Universidade de Aveiro Departamento de Matemática
2016

Maria José Breda Santiago

MÉTODOS DE ESTIMAÇÃO DE FIABILIDADE E CONCORDÂNCIA ENTRE AVALIADORES

DOCUMENTO PROVISÓRIO

Universidade de Aveiro Departamento de Matemática
2016

Maria José Breda Santiago

MÉTODOS DE ESTIMAÇÃO DE FIABILIDADE E CONCORDÂNCIA ENTRE AVALIADORES

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, área de especialização em Estatística e Otimização, realizada sob a orientação científica do Doutor Pedro Miguel Ferreira de Sá Couto, Professor Auxiliar do Departamento de Matemática da Universidade de Aveiro e coorientação científica da Doutora Andreia Oliveira Hall, Professora Associada do Departamento de Matemática da Universidade de Aveiro.

Ao meu pai .

.

O júri / The Jury

Presidente/President	Professor Doutor Pedro Filipe Pessoa Macedo
Vogais/Examiners committee	Professor Doutora Carla Maria Teixeira de Oliveira
	Professor Doutor Pedro Miguel Ferreira de Sá Couto

Agradecimentos / Acknowledgments

Ao professor Pedro e à professora Andreia, por toda a disponibilidade, por todas as revisões, todos os comentários e orientações, enfim, pelo trabalho em equipa.

Palavras-chave Validade, fiabilidade, concordância, intra-avaliador, inter-avaliador, métodos baseados na correção de chance, métodos basedos em rankings; métodos baseados no rácio de variâncias; inferência estatística; cálculo amostral

Resumo Nesta dissertação são apresentados conceitos como validade, fiabilidade e concordância. Validade é definida como a capacidade de um determinado instrumento ser bem fundamentado do ponto de vista teórico e corresponder à realidade que está a ser observada com um elevado grau de exatidão. Fiabilidade refere-se á capacidade de descriminar sujeitos ou objectos enquanto que concordância refere-se ao grau em que os scores ou pontuações medidas são idênticos. A falta de fiabilidade coloca problemas sobre a validade de um instrumento de medição e portanto um instrumento que não seja fiável não pode ser válido. Por outro lado, a existência de fiabilidade não implica a validade. Nesta dissertação são apresentados os métodos para a concordância e fiabilidade mais usuais para as variáveis nominais (métodos baseados na correcção de chance como por exemplo o Kappa de Cohen), variáveis ordinais classificadas em categorias (versões ponderadas dos métodos anteriores) ou classificadas em posições (métodos baseados em rankings, como por exemplo o Kendall tau), e variáveis quantitativas (métodos baseados no rácio das variâncias como por exemplo modelos baseados no coeficiente de correlação intraclasse). O seu cálculo e a sua respetiva inferência estatística (e cálculo amostral) são ilustrados através de casos práticos com recurso ao software R. Foram apresentados e discutidos fluxogramas que auxiliam a escolha apropriada do método de fiabilidade dependendo de algumas condições como o tipo de medida utilizado, do número de avaliadores, ou se o desenho do estudo é inter-avaliador ou intra-avaliador, entre outras condições. Desta forma, esta dissertação vem agrupar e completar muita da informação disponível na literatura, sendo, na nossa opinião, um contributo para uma mais correta aplicação destes métodos de fiabilidade e concordância na construção ou adaptação de instrumentos de medida.

Keywords Validity, reliability, agreement, intra-rater; inter-rater, methods based on chance correction, methods based on rankings, methods based on ratio of variances, inference statistics, sample size calculation

Abstract

This dissertation presents concepts such as validity, reliability and agreement. Validity is defined as the ability of a given instrument is well-founded theoretical point of view and responds to the reality that is being observed with a high degree of accuracy. Reliability refers to the ability to discriminate subjects or objects while agreement refers to the degree to which scores or scores measures are identical.

The unreliability poses problems about the validity of a measuring instrument and therefore an instrument that is not reliable cannot be valid. On the other hand, the existence of reliability does not imply the validity. This dissertation presented methods for the more usual reliability and agreement for nominal variables (methods based on the correction chance such as Kappa Cohen), ordinal variables classified into categories (weighted versions of previous methods), or classified positions (methods based on rankings, such as Kendall tau), and the quantitative variables (methods based on the ratio of the variances such as models based on the intraclass correlation coefficient). The calculation and their respective statistical inference (and sample size calculation) are illustrated through case studies using the software R. Were presented and discussed flowcharts to help select the appropriate reliability method depending on certain conditions such as the type of measure used, the number of raters, or if the study design was inter-rater or intra-rater, among others. Thus, this dissertation regroups and completes the information available in the literature, and in our opinion, contributes to a more correct application of these methods of reliability and agreement for the construction or adaptation of measurement instruments.

ÍNDICE/ INDEX

Índice de tabelas

Índice de figuras

Introdução

Validade e fiabilidade

O conceito de validade geralmente aceite é definido como a capacidade de um determinado construto[1] ser bem fundamentado do ponto de vista teórico e corresponder exatamente à realidade que está a ser observada (Brown, 1970) (Kerlinger, 1986)(McDowell & Newell, 1996). A validade é então um conceito fundamental porque garante que os investigadores estejam a usar métodos que não sejam apenas corretos do ponto de vista ético, clínico ou educacional, mas que sejam válidos e que realmente meçam as variáveis que estão subjacentes a uma determinada questão de investigação. (Brown, 1970)(McDowell & Newell, 1996)

Um trabalho científico geralmente envolve a mensuração de uma ou mais variáveis de interesse, designadas variáveis dependentes (porque normalmente dependem de outras variáveis incluídas no estudo), através de um instrumento de medição. A validade de um instrumento de medição, seja ele um questionário ou um aparelho de medição ou uma escala construída para a recolha de dados, representa o grau de realismo que este instrumento é capaz de medir (Till, 1989).

Um conceito que surge associado ao da validade é o da fiabilidade. No entanto são conceitos bastante diferentes. Uma medida diz-se que tem uma fiabilidade elevada se essa medida for capaz de diferenciar sujeitos ou objetos (Carmines & Zeller, 1979)(Vet & De Vet, 1998). Fiabilidade é também geralmente definida como a coerência das medidas ou a ausência de erro de medição (Carmines & Zeller, 1979)(Vet & De Vet, 1998). Uma medida pode ser

[1] Constructo é definido como um conceito teórico não observável que representa traços, aptidões ou características supostamente existentes e abstratas de uma variedade de comportamentos que tenham significado educacional ou psicológico ou outros como por exemplo a personalidade ou a inteligência.

fiável (precisa), mas pode estar errada e portanto não ser válida, mas não pode ser válida sem que seja fiável. Portanto fiabilidade não implica validade mas é um requisito para avaliar a validade. Ou seja, uma medida para ser válida tem de ser fiável. Deste modo, a fiabilidade é uma condição necessária mas não suficiente para a validade.(Murphy & Davidshofer, 2005)

Na figura A estão ilustradas diferentes combinações possíveis entre fiabilidade e validade que podem ocorrer. Quando uma medida é válida e fiável (gráfico à esquerda), a validade pode ser observada através pelo conjunto de pontos certeiros no alvo e a fiabilidade pode ser analisada através da pequena dispersão dos resultados no centro do alvo. Dos casos apresentados na figura A, o da esquerda é o único que tem interesse científico. No caso de um instrumento não ser válido e também não ser fiável (gráfico à direita), então significa que as medições falham o centro do alvo acertando apenas numa parte dele, ou seja, estão a medir uma realidade diferente ou parcial da pretendida e apresentam uma grande dispersão de resultados mostrando um padrão de respostas aleatório. Por último, uma medida pode ser fiável mas não ser válida (gráfico ao centro), ou seja, os resultados apresentam dispersão reduzida, mas a realidade medida não é a pretendida. No entanto, nestes dois últimos casos, obrigam a uma nova análise do processo de validação e posteriormente um novo estudo de fiabilidade.

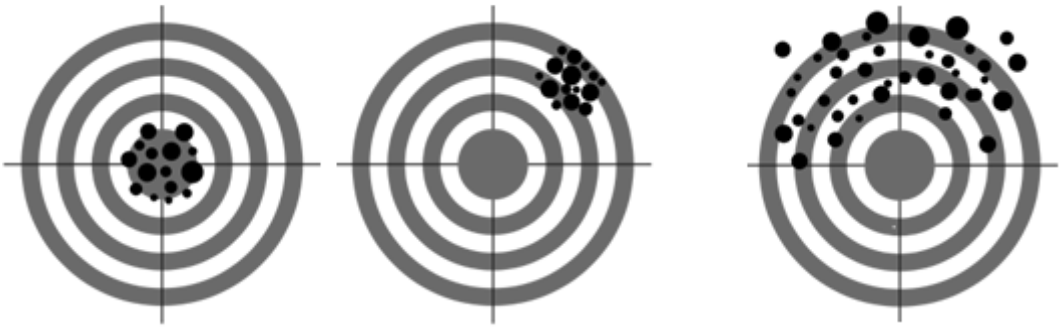

Figura A. Ilustração dos conceitos de validade e fiabilidade. Adaptado a partir de http://en.wikipedia.org/wiki/Validity_(statistics)

Exatidão e precisão

Outros termos frequentemente utilizados neste contexto são a exatidão (*accuracy*) e a precisão (*precison*) de um instrumento. Fiabilidade é uma forma

útil de descrever precisão enquanto validade é usada para descrever exatidão.(Taylor, 1999). A exatidão está associada aos erros sistemáticos de um instrumento enquanto a precisão está associada aos erros aleatórios desse instrumento, respetivamente (Figura B). A precisão de uma medição é o grau com que medidas repetidas sob as mesmas condições mostram os mesmos resultados. Conceitos sobre reprodutibilidade e repetibilidade aparecem então como sinónimos de precisão. Por exemplo, se uma experiência contém um erro sistemático então o aumento da dimensão da amostra geralmente aumenta a precisão mas não melhora a exatidão. O resultado será consistente mas incorreto resultando numa experiência falhada. Eliminando o erro sistemático melhora a exatidão mas não altera a precisão.

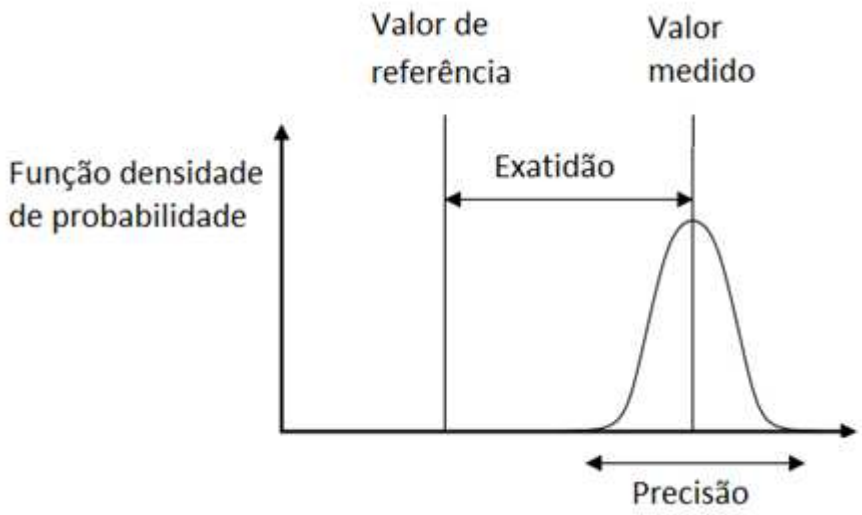

Figura B. Ilustração dos conceitos de exatidão e precisão. Adaptado a partir de http://en.wikipedia.org/wiki/Accuracy_and_precision

Concordância, fiabilidade, repetibilidade e reprodutibilidade.

A concordância (agreement) e a fiabilidade (reliability) são questões muito importantes no desenvolvimento e utilização de um instrumento ou escala de medição. O controlo destes aspetos assegura a qualidade da mediação efetuada (Kottner et al., 2011). Os resultados sobre concordância e fiabilidade indicam informação sobre a quantidade de erro associada a qualquer diagnóstico, resultados ou medição obtidos e por conseguinte determina a validade dos resultados obtidos num determinado estudo.

Introdução

Na literatura estatística (e não só) para descrever estudos baseados em medidas com erro, os termos mais utilizados são concordância, fiabilidade, repetibilidade (repetiability) e reprodutibilidade (reproducibility). Estes termos são muitas vezes utilizados erroneamente como sinónimos (Bartlett & Frost, 2008).

A concordância é definida como o grau em que scores ou pontuações medidas no mesmo sujeito são idênticas (Kottner et al., 2011). Duas medidas no mesmo sujeito podem ser diferentes por várias razões, dependendo das condições sobre as quais as medidas foram feitas (por exemplo, diferentes avaliadores ou diferente momentos de avaliação). A concordância entre as medições é uma característica do instrumento de medida envolvido e não depende da população onde as medidas foram obtidas, a não ser que exista um enviesamento do próprio instrumento (por exemplo, o instrumento esteja mal calibrado). Este grau de concordância vai ser estimado pela medida do erro existente numa situação de medidas repetidas (Bartlett & Frost, 2008)(Kottner et al., 2011).

A fiabilidade é definida como a capacidade de um instrumento de medição diferenciar sujeitos (por exemplo pacientes) ou objetos (por exemplo, imagens de raio X) e é definida matematicamente como:

$$\text{Fiabilidade} = \frac{\text{Variabilidade entre sujeitos (sem erro)}}{\text{Variabilidade entre sujeitos (sem erro)} + \text{variabilidade do erro de medição}}$$

O valor da fiabilidade será elevado se os erros de medição forem pequenos quando comparados com a verdadeira diferença entre sujeitos. Desta forma, significa que os sujeitos podem ser relativamente bem distinguidos entre si. Se os erros de medição tenderem a ser elevados quando comparados com a verdadeira diferença entre os sujeitos, a fiabilidade irá ter um valor baixo porque diferenças entre as medições de dois sujeitos seriam puramente devido a erro em vez de verdadeiras diferenças dos seus valores.

Na figura C estão ilustradas diferentes combinações entre fiabilidade e concordância. Para cada caso, iremos considerar o peso de duas pessoas, medido em cinco dias diferentes. As cinco medições por pessoa mostram

alguma variação. O desvio padrão (SD) dos valores das medições repetidas de uma pessoa representa a concordância, ou seja, diz respeito ao erro de medição avaliando o quão perto estão as medidas repetidas. Para a fiabilidade o erro dessas medidas está relacionado com a variabilidade entre as pessoas, e diz-nos como podem ser distinguidas uma da outra.

Quando uma medida é fiável, os pesos das duas pessoas estão distantes, logo o erro de medição não afetará a discriminação das pessoas, deste modo os resultados medidos aparecerão como na parte superior da figura 2. No entanto no canto superior esquerdo, o erro de medição é baixo no sentido em que não existe um grande desvio (SD) dos valores dos pesos de cada uma das pessoas (consequentemente a concordância é elevada), enquanto que na parte superior direita, os valores dos pesos de cada pessoa apresentam uma maior variabilidade, deste modo o erro de medição será mais elevado (e a concordância mais reduzida). Quando os valores dos pesos das pessoas P1 e P2 estão muito próximos, o erro de medição afetará a capacidade de as discriminar, como acontece na parte inferior da figura 2, deste modo a fiabilidade será bastante inferior. Na parte inferior esquerda, o erro de medição nas medidas repetidas é pequeno, enquanto na parte inferior direita, esse erro é bastante mais elevado.

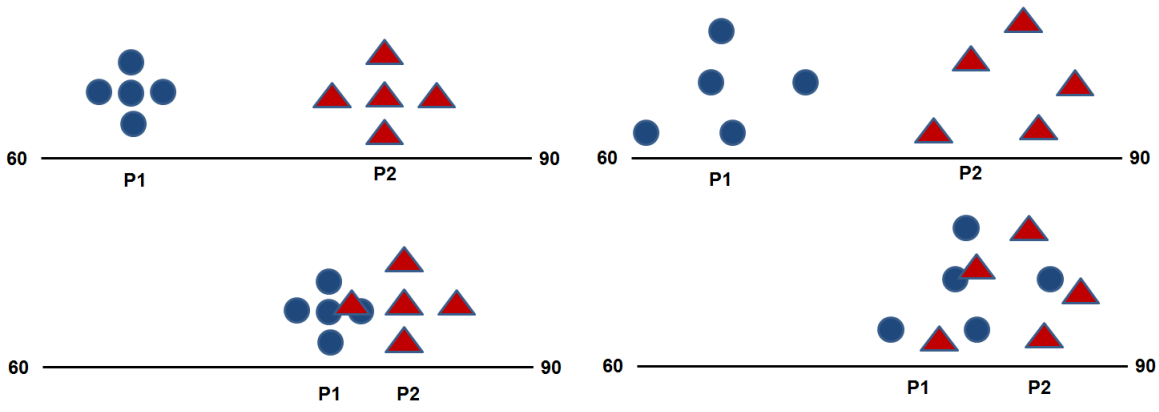

Figura C. Ilustração das diferentes combinações possíveis dos conceitos de fiabilidade e concordância.

A concordância e a fiabilidade têm como objetivo a resposta a duas diferentes questões (Vet, Terwee, Knol, & Bouter, 2006). A concordância está

relacionada com a questão “qual é a concordância entre as medidas repetidas? (*how good is the agreement between repeated measurements?*) enquanto a fiabilidade está relacionada com a questão “quanto fiável é a medição?” (*How reliable is the measurement?*).

O conceito de reprodutibilidade (reproducibility) diz respeito à variação das medidas no mesmo sujeito, mas havendo uma variação nas condições da experiência (Bartlett & Frost, 2008). Estas variações podem ser devidas à utilização de diferentes observadores ou avaliadores, ou a utilização de diferentes instrumentos de medição sobre o mesmo conjunto de sujeitos. Outras formas de variação podem ter origem no próprio instrumento de medição, ou nas circunstâncias em que as medições estão a ser realizadas, por exemplo alguns instrumentos podem ser dependentes da temperatura, ou humor de um entrevistado podendo deste modo influenciar as respostas a um questionário (Vet et al., 2006).

O conceito de repetibilidade (repeatibility) diz respeito à variação das medidas repetidas realizadas sobre o mesmo conjunto de sujeitos e sobre as mesmas condições (Bartlett & Frost, 2008). Isto significa que as medidas são todas realizadas com o mesmo instrumento, o mesmo observador ou avaliador e onde as medidas são realizadas com uma janela temporal pequena. A variabilidade na medição do erro deve estar apenas relacionada com o próprio instrumento e não com os sujeitos que participam no estudo, dado que a pequena janela temporal não permite que haja grandes alterações no comportamento do sujeito. O desenho deste tipo de estudos é muitas vezes referido como teste-reteste (ver próxima secção).

Tipos de desenho de estudos

Como foi atrás descrito, conceitos de fiabilidade e concordância são importantes porque ambos fornecem informação sobre a qualidade das medidas obtidas. Para além disso, o desenho dos estudos para avaliar os dois conceitos são os mesmos e podem ser divididos em duas grandes categorias (Kottner et al., 2011):

- Estudos de fiabilidade/concordância inter-observador (*Inter-rater reliability/agreement).* Neste tipo de estudos, pretende-se avaliar se diferentes avaliadores, usando a mesma escala, classificação, instrumento ou teste avaliam os mesmos sujeitos ou objetos da mesma forma.
- Estudo de fiabilidade/concordância intra-avaliadores (Intra-rater reliability/agreement), também designada como teste-reteste. Neste caso pretende-se avaliar se o mesmo avaliador utilizando a mesma escala, classificação, instrumento ou teste, avalia da mesma forma os mesmos sujeitos ou objetos em momentos diferentes.

Os desenhos de estudos relacionados com a medição da consistência interna (por exemplo, o alfa de Cronbach), muito aplicados na qualidade de informação obtida em questionários, ou na área da teoria da resposta ao item, estão fora do âmbito desta dissertação.

Objetivos da dissertação

Nesta dissertação, os objetivos principais são a apresentação e descrição dos métodos mais usuais utilizados na estimação de concordância (como por exemplo o Kappa de Cohen) e de fiabilidade (como por exemplo o coeficiente de correlação intraclasse) entre dois ou mais avaliadores. Especial ênfase é dado na distinção entre os diferentes métodos, permitindo desta forma uma mais fácil identificação sobre quais serão os métodos mais adequados a um determinado problema de investigação baseados apenas nos tipos de medidas da variável em estudo, distinguindo variáveis nominais, ordinais ou quantitativas.

Para cada método são apresentadas as suas formulações matemáticas principais que permitam a obtenção de estimativas, indicando as vantagens e desvantagens associadas, como objetivos secundários. As suas propriedades estatísticas e respetiva inferência estatística (Testes de hipóteses, intervalos de confiança e calculo da dimensão da amostra) também são abordados. A apresentação de exemplos práticos e do código respetivo realizado com o software R, são uma mais-valia para a compreensão dos métodos apresentados.

Disposição dos capítulos da dissertação

Esta dissertação é dividida pelos seguintes capítulos:

- No capítulo 1 são abordadas questões relativas á validade de um estudo.
- No capítulo 2 são apresentados os principais métodos de concordância para variáveis nominais para dois ou mais avaliadores.
- No capítulo 3 são apresentados os principais métodos de concordância entre dois ou mais avaliadores para variáveis ordinais categorizadas e é também apresentado o estudo para quando existem valores em falta.
- No capítulo 4 são apresentados os principais métodos de fiabilidade para variáveis quantitativas considerando as situações de inter-avaliador e intra-avaliador (dois ou mais avaliadores) sem réplicas.
- No capítulo 5 são apresentados os principais métodos de fiabilidade para variáveis quantitativas considerando as situações de inter-avaliador e intra-avaliador (dois ou mais avaliadores) com réplicas.
- No capítulo 6 são apresentados os principais métodos de fiabilidade entre dois ou mais avaliadores, para variáveis que sejam classificadas através de rankings, sejam ordinais ou quantitativas.
- No capítulo 7 são apresentados resultados de inferência estatística associados aos métodos apresentados nos capítulos 2 a 6.
- No capítulo 8 é apresentada a discussão e as principais conclusões obtidas nesta dissertação.

Referências

Bartlett, J. W., & Frost, C. (2008). Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. *Ultrasound in Obstetrics & Gynecology : The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*, *31*(4), 466–475. http://doi.org/10.1002/uog.5256

Brown, frederick gramm. (1970). *principles of educational and psychogical testing*. (D. Press, Ed.).

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and Validity Assessment*. *PsycCRITIQUES* (Vol. 17). http://doi.org/10.1037/018269

Kerlinger, F. N. (1986). *Foundations of behavioural research*. *Holt Rinehart Winston London* (Vol. Hoit, Rine).

Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., … Streiner, D. L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *International Journal of Nursing Studies*, *48*(6), 661–671. http://doi.org/10.1016/j.ijnurstu.2011.01.016

McDowell, I., & Newell, C. (1996). *Measuing health: A guide to rating scales and questionnaires. Measuring health A guide to rating scales and questionnaires 2nd ed* (Vol. 2nd). Retrieved from http://search.proquest.com/docview/619019361?accountid=11233

Murphy, K. R., & Davidshofer. (2005). *Psychological Testing: Principles and Applications* (6th Editio). Pearson; 6 edition (October 1, 2004).

Taylor, J. R. (1999). *Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. University Science Books.

Till, A. (1989). *Measuring Health — a Guide To Rating Scales and Questionnaires*. *The Journal of the Canadian Chiropractic Association* (2nd ed, Vol. 33). New York: Oxford University Press Inc. http://doi.org/10.1179/108331900786166731

Vet, H., Terwee, C., Knol, D., & Bouter, L. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology*, *59*(10), 1033–9. http://doi.org/10.1016/j.jclinepi.2005.10.015

Vet, H. De, & De Vet, H. C. W. (1998). Observer reliability and agreement. *Encyclopedia of Biostatistics*, 3123–3128. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Observer+Reliability+and+agreement#1

Capítulo 1: Conceitos sobre validade

Este capítulo apresenta o conceito de validade, bem como os tipos de validade sobre um instrumento de medição mais comuns na literatura. Procura-se também fazer uma descrição sucinta das técnicas estatísticas mais relevantes associadas à validade.

1.1 Validade de um instrumento

O conceito de validade é definido como a capacidade de um determinado instrumento ser bem fundamentado do ponto de vista teórico e corresponder à realidade que está a ser observada com um elevado grau de exatidão (Till, 1989). A validade de um instrumento garante que os investigadores estão a usar ferramentas que não são apenas corretas do ponto de vista ético, clínico ou educacional, mas que são válidas para a experiência que pretendem realizar e que medem as variáveis que estão subjacentes a uma determinada questão de investigação. Ou seja, a validade de um instrumento representa o grau de realismo que o instrumento é capaz de medir (Till, 1989). Por exemplo, um eletrocardiograma tem validade para medir a frequência cardíaca e o trajeto do impulso eléctrico dentro do coração, mas não é valido para detetar inflamações ou úlceras no estômago.

Os instrumentos de avaliação utilizados podem ser desde aparelhos de medição utilizados em laboratórios até questionários utilizados nas Ciências Sociais para estudar determinado construto[1]. Para cada instrumento de

[1] Construto é definido como um conceito teórico não observável que representa traços, aptidões ou características supostamente existentes e abstratas de uma variedade de comportamentos que tenham significado educacional ou psicológico ou outros como por exemplo a personalidade ou a inteligência.

avaliação podem ser utilizadas diferentes metodologias para avaliar a sua validade. As formas de validade mais utilizadas são: validade de conteúdo, validade de construto e validade de critério (Carmines & Zeller, 1979; Kirk & Miller, 1986).

Nas próximas subsecções irão ser apresentados alguns métodos de validade associados a um determinado instrumento de avaliação (*test validity*) que têm uma componente matemática/estatística relevante. Na última subsecção, serão brevemente apresentados métodos de validade mais relacionados com a metodologia de investigação (*experimental validity*) como por exemplo, métodos de validade relacionados com o desenho da experiência ou a generalização dos resultados obtidos para a população utilizada num determinado estudo. Uma discussão em profundidade dos conceitos de validade e técnicas associadas estão fora dos objetivos desta dissertação.

1.2 Validade de conteúdo

A validade de conteúdo (*content validity*) analisa o grau de exatidão associado ao conteúdo dum determinado instrumento. Este tipo de validade é geralmente utilizado na fase de construção de instrumentos com base em questionários (ou escalas). O investigador após uma profunda análise da literatura científica conhecida para descrever um determinado construto (ou fator latente), constrói um instrumento para avaliar esse construto (Carmines & Zeller, 1979; Cronbach & Meehl, 1955). Seguidamente, um painel de especialistas avalia a validade de conteúdo do instrumento antes de ser aplicado no campo.

Exemplos de escalas deste tipo são inúmeras e um exemplo é a escala de coma de Glasgow (Teasdale & Jennett, 1974). Esta escala neurológica tem por objetivo pontuar o estado de consciência de uma pessoa após uma lesão cerebral. Após avaliar a resposta dos itens: movimentos oculares, verbais e motores, cada paciente irá ser classificado através de um score final que indicará o grau da lesão (construto).

Para demonstrar validade de conteúdo, o investigador terá de demonstrar:

- Validade de face (*face validity*), onde avalia se o instrumento de avaliação "parece" ser uma boa medida ou não. Esta avaliação poderá ser feita por um não especialista da matéria e será sempre considerada um ponto de partida no processo de validação (Litwin, 1995);
- Validade de representação (*representation validity*), onde avalia em que medida uma construção teórica abstrata está bem representada num teste prático. Geralmente esta validação é feita utilizando um painel de especialistas (*experts*) sobre determinada área (Litwin, 1995) e o método estatístico mais utilizado é o índice de validade de conteúdo.

Índice de validade de conteúdo

O índice de validade de conteúdo (IVC) mede o grau de satisfação do painel de especialistas quando avalia um determinado instrumento de avaliação (Alexandre & Coluci, 2011; Polit & Beck, 2006). É solicitado a cada especialista que pontue, de forma independente, os itens apresentados num determinado instrumento segundo uma escala *Likert*, construída da seguinte forma: 1- não concorda com o tópico e sugere a eliminação do mesmo; 2- não concorda e propõe alterações substanciais de forma a constar no instrumento; 3- concorda, na generalidade, mas propõe alterações; 4- concorda totalmente. Nas pontuações 1, 2 e 3 é usualmente pedido aos especialistas a justificação da classificação e sugestão de mudanças (caso existam). O IVC para cada item (IVC_{item}) pode ser calculado:

$$\mathrm{IVC_{Item}} = \frac{\text{nº de avaliadores com pontuação 3 ou 4}}{\text{nº total de avaliadores com pontuação } 1, 2, 3 \text{ ou } 4} \qquad 1.1$$

Após o cálculo do IVC_{item}, calcula-se a validade de conteúdo para o instrumento de avaliação global (SIVC). Polit e Beck (2006) apresentam duas formas diferentes para realizar este cálculo. A primeira é a proporção de itens que foram pontuados com cotação 3 ou 4 por todos os especialistas, designado

por índice de validade de conteúdo por concordância universal ($SIVC_{UA}$) enquanto a segunda forma representa a média dos valores IVC_{item} obtidos para cada item ($SIVC_{AVE}$):

$$\mathrm{SIVC}_{UA} = \frac{\text{nº de itens } \textbf{onde todos os especialistas} \text{ pontuam com 3 ou 4}}{\text{nº total de itens avaliados}} \quad 1.2$$

$$\mathrm{SIVC}_{\mathrm{AVE}} = \frac{\sum_{i=1}^{N} IVC_{item_i}}{\mathrm{N}} \quad 1.3$$

onde N representa o número de itens avaliados. O autor Lynn (1986) aconselha o uso de pelo menos 3 especialistas até um máximo de 10. O mesmo autor indica que se forem 5 ou menos avaliadores, o valor do IVC_{item} deve ser igual a 1 enquanto se forem mais do que 5 avaliadores o valor mínimo deve ser não inferior a 0.83 (Lynn, 1986). Outra referência refere um valor para o IVC_{item} de 0.78 como mínimo aceitável (Polit & Beck, 2006). O valor mínimo aceitável do SIVC deve ser de 0.80 (Berk, 1990) embora o valor de 0.90 também é citado como recomendado (Davis, 1992). Como se pode observar existe alguma discrepância nos valores de referência encontrados na literatura, o que reflete a dificuldade de estabelecer critérios globalmente aceites.

Na tabela 1.1 é apresentado o exemplo que vem descrito no artigo de Polit e Beck (2006). Neste exemplo é utilizado um painel de 6 especialistas e o valor 1 representa uma pontuação de 3 ou 4 para o item medido. Como se pode observar, todos os valores do IVC_{item} indicam relevância. Nas medidas para a avaliação global do instrumento, o $SIVC_{AVE}$ é bastante elevado, indicando validade do conteúdo. Esta medida podia ser calcula de outra forma com valores sempre idênticos, indicando o nº de casos onde se obteve uma classificação 3 ou 4, sobre o total de classificações (54/60=0.9). No caso da medida $SIVC_{UA}$ o valor obtido é fraco e revela que quanto maior for o número de especialistas utilizados maior será a probabilidade de se obter valores baixos para esta medida, independentemente de todos os valores dos IVC_{item} serem elevados. Estes autores sugerem que se publiquem estes três resultados.

Tabela 1.1- Exemplo ilustrativo do cálculo de IVC_{item}, $SIVC_{UA}$ e $SIVC_{AVE}$ para um conjunto de dados apresentados por Polit e Beck (2006)

	Especialistas						IVC_{item}
Item	E1	E2	E3	E4	E5	E6	
1	0	1	1	1	1	1	5/6=0.83
2	1	0	1	1	1	1	5/6=0.83
3	1	1	0	1	1	1	5/6=0.83
4	1	1	1	0	1	1	5/6=0.83
5	1	1	1	1	0	1	5/6=0.83
6	1	1	1	1	1	0	5/6=0.83
7	1	1	1	1	1	1	6/6=1
8	1	1	1	1	1	1	6/6=1
9	1	1	1	1	1	1	6/6=1
10	1	1	1	1	1	1	6/6=1
Proporção de relevância	9/10= 0.9	9/10= 0.9	9/10= 0.9	9/10= 0.9	9/10= 0.9	9/10= 0.9	

$SIVC_{UA}$=4/10=0.4; $SIVC_{AVE}$=(6*0.83+4*1)/10=0.90

1.3 Validade de construto

Após ter sido feita a validade do conteúdo, a próxima etapa é a validade do construto (*construct validity*). Esta envolve utilizar suporte empírico e teórico para a interpretação do construto ou variável latente medido através da análise dos resultados obtidos após a aplicação de um instrumento de avaliação (Carmines & Zeller, 1979; Cronbach & Meehl, 1955). Exemplos deste tipo de validade estão relacionados com medições sobre o cérebro humano como a inteligência ou o nível de emoção, representando situações onde existe uma grande subjetividade ou variabilidade. Este tipo de validade é geralmente dividido em:

- Validade convergente (*convergent validity*) refere-se ao grau em que duas ou mais medidas do instrumento, que se esperam estar relacionadas entre si, estão realmente relacionadas (Campbell & Fiske, 1959; Carmines & Zeller, 1979).
- Validade discriminante (*discriminant validity*), refere-se ao grau ou à capacidade de duas ou mais medidas que supostamente são independentes entre si, serem realmente independentes (Campbell & Fiske, 1959).

Os métodos estatísticos mais usuais são os métodos de correlação, construindo matrizes de correlação entre os diferentes itens utilizados ou métodos multivariados que estudam o padrão de respostas aos itens como a análise factorial exploratória.

Por exemplo, num estudo sobre a autoestima (construto principal do estudo) são medidos também os seguintes construtos: autoestima e locus de controlo[2]. Na tabela 1.2 estão apresentados os valores de correlação de Pearson para estes dois constructos através de uma matriz de correlação entre os 6 itens considerados.

Tabela 1.2- Matriz de correlações entre os 3 itens (AE1, AE2 e AE3) do constructo autoestima e os 3 itens (LC1, LC2 e LC3) do constructo locus de controlo.

	AE1	AE2	AE3	LC1	LC2	LC3
AE1	1					
AE2	0.83	1				
AE3	0.89	0.85	1			
LC1	0.02	0.05	0.04	1		
LC2	0.12	0.11	0.01	0.84	1	
LC3	0.09	0.03	0.06	0.93	0.91	1

Para demostrar validade convergente, para o construto autoestima, os 3 itens (AE1, AE2 e AE3) têm de apresentar correlações elevadas entre si e para o construto locus de controlo os 3 itens (LC1, LC2 e LC3) também devem apresentar correlações elevadas entre si. Para demonstrar validade discriminante, os 3 itens da autoestima devem apresentar correlações baixas com os 3 itens do locus de controlo. Este exemplo pode ser facilmente estendido para os outros construtos. Os itens dentro do construto devem apresentar correlações elevadas entre si, mostrando validade convergente e os

[2] Locus de controlo é a expectativa do indivíduo sobre a medida em que os seus estímulos a um comportamento se encontram sob controle interno (esforço pessoal, competência, etc.), ou externo (as outras pessoas, sorte, acaso, etc.).

itens entre os diferentes constructos devem apresentar correlações baixas entre si, mostrando validade discriminante.

Medidas de correlação

As medidas de correlação ou associação quantificam a intensidade e a direção da associação entre duas variáveis, isto é permitem observar o grau de dependência entre duas variáveis. As correlações podem ser bivariadas (se envolvem apenas duas variáveis) ou multivariadas (se envolvem mais de duas variáveis). Existem vários coeficientes de correlação bivariados que são definidos em função da escala de medida das variáveis consideradas. Alguns, dos coeficientes de correlação usados mais frequentemente são os coeficientes de correlação de *Pearson* e de *Spearman*.

O coeficiente de correlação de *Pearson* mede a intensidade e a direção da associação de tipo linear entre duas variáveis contínuas e para a realização estatística inferencial, pressupõe que a distribuição conjunta seja normal bivariada. Esta associação é calculada a partir da covariância ($Cov(X_1, X_2)$) entre duas variáveis X_1 e X_2 dada pela equação 1.4 e o coeficiente de correlação de Pearson ($R_{X_1X_2}$) pode ser obtido estandardizando a covariância pelos desvios- padrão das duas variáveis (equação 1.5):

$$Cov(X_1, X_2) = \frac{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{n-1} \qquad 1.4$$

$$R_{X_1X_2} = \frac{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sqrt{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2}\sqrt{\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2}} = \frac{Cov(X_1, X_2)}{\sqrt{Var(X_1)}\sqrt{Var(X_2)}} \qquad 1.5$$

O coeficiente de correlação de *Spearman* é uma medida de associação não-paramétrica entre duas variáveis pelo menos ordinais ou quantitativas. O coeficiente pode ser calculado usando a fórmula do coeficiente de correlação

de Pearson, substituindo os valores das observações de X_1 e X_2 pelas respectivas ordens r_1 e r_2 (Zar, 1999):

$$R_S = \frac{\sum_{i=1}^{n}(r_{1i} - \bar{r}_1)(r_{2i} - \bar{r}_2)}{\sqrt{\sum_{i=1}^{n}(r_{1i} - \bar{r}_1)^2}\sqrt{\sum_{i=1}^{n}(r_{2i} - \bar{r}_2)^2}} = 1 - \frac{6\sum_{i=1}^{n}(r_{1i} - r_{2i})^2}{n^3 - n} \quad 1.6$$

Os valores de correlação de Pearson e Spearman estão sempre no intervalo [-1.0;+1.0] e quanto mais próximo dos extremos, maior será a relação de dependência entre as duas variáveis. Quanto mais próximo do valor 0, menor será a dependência entre as duas variáveis. Correlações positivas significam que o aumento de uma das variáveis se traduz num aumento da outra. Correlações negativas indicam que o aumento de uma das variáveis se traduz na diminuição da outra. Na tabela 1.2 estão apresentados valores de correlação entre duas variáveis.

Análise fatorial exploratória

A análise fatorial exploratória (AFE) é uma técnica de análise exploratória de dados que tem por objectivo descobrir e analisar a estrutura de um conjunto de variáveis interrelacionadas de modo a construir uma escala de medida para factores latentes e intrínsecos que de alguma forma controlam as variáveis originais. Em princípio, se duas ou mais variáveis estão correlacionadas (e essa correlação não é espúria), essa associação resulta da partilha de uma característica comum não diretamente observável (fator comum latente).

Uma AFE usa correlações observadas entre variáveis originais para estimar o(s) fator(es) comum(ns) e as relações estruturais que ligam os fatores (latentes) às p variáveis, podendo ser modeladas por:

$$\begin{cases} x_1 = \mu_1 + \lambda_{11} f_1 + \lambda_{12} f_2 + \ \ldots + \lambda_{1m} f_m + \eta_1 \\ x_2 = \mu_2 + \lambda_{21} f_1 + \lambda_{22} f_2 + \ \ldots + \lambda_{2m} f_m + \eta_2 \\ \ldots \\ x_p = \mu_p + \lambda_{p1} f_1 + \lambda_{p2} f_2 + \ \ldots + \lambda_{pm} f_m + \eta_p \end{cases} \quad 1.7$$

onde f_j representa os fatores comuns (ou latentes), sendo desejável que o nº de fatores comuns seja bastante inferior ao nº de variáveis (m<p), η_i representa os p factores específicos e λ_{ij} representa o peso da variável i no fator j (designados por *factor loadings*), ou seja cada λ_{ij} mede a contribuição do fator comum j na variável i (Hair, Black, Babin, & Anderson, 2009).

As seguintes propriedades devem ser respeitadas:

- Os fatores comuns (f_j) devem ser independentes (e ortogonais) e identicamente distribuídos com média 0 e variância 1, com j=1,…,m;
- Os fatores específicos (η_i) devem ser independentes e igualmente distribuídos com média 0 e variância constante, com i=1,…,p;

Os fatores f_j e η_i devem ser independentes. Se esta condição se verificar, o modelo fatorial diz-se ortogonal, mas se f_j e η_i se apresentarem correlacionadas então o modelo factorial diz-se oblíquo.

O objectivo final de uma AFE é agrupar os itens medidos sob um determinado fator latente ou construto. A tabela 1.3 representa um resultado hipotético sobre uma AFE com 10 itens e 3 fatores latentes.

Da tabela 1.3 observa-se que os itens 1, 4 e 8 estão agrupados ao fator F1, os itens 2, 5, 6 e 10 estão agrupados ao fator F2 e os itens 3, 7 e 9 estão agrupados ao fator F3. Se esta distribuição dos itens estiver de acordo com o comportamento teórico esperado então o resultado da AFE apresentado é bastante satisfatório. No entanto se o item 1 que está associado ao F1, teoricamente estivesse associado a um outro fator, o resultado da tabela já não seria satisfatório. Esta técnica indica simultaneamente resultados para a validade convergente, através da associação dos itens a cada fator, e para a validade discriminante, um item só pertence a um único fator. Se os

agrupamentos/divisões dos itens propostos tiverem um suporte lógico e coerente que justifiquem estes resultados, então a validade de constructo fica verificada.

Tabela 1.3. Sumário de uma análise fatorial exploratória hipotética com 10 itens distribuídos por 3 fatores latentes (F1, F2, e F3).

	Fatores		
Item	F1	F2	F3
1	0.80		
2		0.83	
3			0.79
4	0.75		
5		0.86	
6		0.75	
7			0.78
8	0.90		
9			0.83
10		0.86	

Os valores apresentados na tabela 1.3 são só os valores relevantes para o exemplo hipotético. Os valores inferiores a 0.5 em módulo não são apresentados porque são considerados irrelevantes.

1.4 Validade de critério

A validade de critério (criterion-oriented validity) avalia os resultados obtidos no novo instrumento de avaliação contra um instrumento de referência, designado por *gold standard*. Uma limitação importante à avaliação deste tipo de validade reside no facto que para a maioria das medidas usadas nas ciências sociais, não existirem variáveis de critério que sejam de referência (gold standard) e, quando existem, torna-se difícil provar de forma imparcial a validade da medição de critério. Este tipo de validade é normalmente dividido em:

- Validade simultânea ou concorrente (*concurrent validity*) que relaciona o grau de correlação da nova medida (ou instrumento) com outras medidas (ou

instrumentos) previamente validadas. Todas as medidas são aplicadas no mesmo instante de tempo e aos mesmos sujeitos;

- Validade preditiva (*predictive validity*) que relaciona os valores obtidos para a nova medida (ou instrumento) com a sua capacidade de previsão de um conjunto de medidas num tempo futuro (utilizando os mesmos sujeitos).

Um exemplo de validade simultânea é a utilização de escalas que sejam consideradas como tendo uma validade elevada como escala do coeficiente de inteligência, coeficiente emocional, ou outras escalas previamente validadas e reconhecidas por outros investigadores como *gold standards*. Então a nova escala (ou medida) será correlacionada contra estas escalas definidas como gold standard.

Um exemplo para validade preditiva é o processo de seleção de estudantes para uma Universidade. A questão colocada é se uma determinada escala, como por exemplo, a nota de entrada na Universidade, pode ser preditiva da nota final do curso (num tempo futuro). Se assim for então a nota de entrada da Universidade tem a propriedade de validade preditiva. Numa situação de validade preditiva, a variável de interesse é medida numa primeira fase e as medidas de critério numa segunda fase, num tempo futuro.

Os métodos mais usuais utilizados para a convergência concorrente são métodos de correlação, enquanto para a validade preditiva, os métodos mais usuais são os modelos de previsão, como por exemplo métodos de regressão (linear ou não) entre duas variáveis.

Modelos de previsão

O termo “Análise de Regressão” define um conjunto vasto de técnicas estatísticas usadas para modelar relações entre variáveis e predizer o valor de uma ou mais variáveis dependentes a partir de um conjunto de variáveis independentes (ou preditoras). O termo variável dependente implica

geralmente uma relação do tipo causa-e-efeito. Porém, a análise de regressão pode ser usada para modelar a relação funcional entre duas variáveis, ou seja, através de uma função matemática, independentemente de existir ou não uma relação de causa-e-efeito (que nem sempre é fácil de demonstrar).

O modelo que irá ser representado nesta dissertação é o modelo de regressão linear univariado. Existem outros modelos de regressão como por exemplo: modelos de regressão não linear, modelos de regressão logística binária ou multinomial, regressão ordinal e análise sobrevivência**.** No modelo de regressão linear univariado, a relação funcional entre uma variável dependente (Y, modelo univariado) e uma ou mais variáveis independentes (X_i, i=1,…,p) é do tipo:

$$\mathrm{Y} = b_0 + b_1 X_1 + b_2 X_2 \ \ldots + b_p X_p + \varepsilon \qquad 1.8$$

onde os b_i são os chamados coeficientes de regressão e ε representa os erros do modelo. O coeficiente b_0 é a ordenada na origem e os coeficientes b_i representam os declives parciais (ou seja, representa a variação de Y por unidade de variação de X_i). O termo ε reflete os erros de medição e a variação natural em Y. Este modelo exige que os erros sejam aleatórios, independentes e com distribuição normal de média zero e variância constante (Hair et al., 2009). Caso exista apenas uma variável independente o modelo simplifica-se e designa-se por modelo de regressão linear simples:

$$\mathrm{Y} = b_0 + b_1 X_1 + \varepsilon \qquad 1.9$$

Uma condição necessária nestes modelos é a não existência de multicolinearidade entre as variáveis independentes, ou seja, é necessário que as variáveis independentes sejam ortogonais, que não estejam correlacionadas ou quanto muito apenas apresentem correlações fracas entre si.

1.5 Outros tipos de validade

A validade experimental (*experimental validity*) está relacionada com a validade do desenho da experiência e com questões éticas sobre a mesma. Sem um desenho experimental válido, não é possível obter conclusões científicas válidas. A validade experimental pode ser dividida: em validade interna, validade externa, validade ecológica e validade das conclusões estatísticas:

- Validade interna (*internal validity*) avalia as relações entre as variáveis independentes e as variáveis dependentes. Esta validade exige o controlo das variáveis estranhas ou covariantes com o objetivo de eliminar qualquer contaminação que essas variáveis possam ter nos resultados das variáveis medidas. A qualidade de uma validade interna pode ser assegurada pelo controlo do desenho experimental através da utilização de uma amostragem aleatória na seleção dos sujeitos a incluir no estudo, a repartição aleatória dos sujeitos pelos grupos de controlo ou experimentais, utilização de instrumentos de medida fiáveis, manipulação rigorosa dos processos utilizados (utilização de procedimentos duplamente ocultos[3]) e utilização de técnicas para identificação de variáveis de confundimento[4] (através da identificação de relações espúrias através de correlações totais ou parciais, análises de covariância ou de regressão).
- Validade externa (*external validity*) diz respeito à generalização dos resultados, tendo por base os resultados obtidos (Messick, 1995). Esta validade permite dizer se os resultados obtidos podem ser reproduzidos com uma outra amostra de indivíduos, noutro local e tempo, permitindo avaliar o grau de generalização dos resultados obtidos.
- Validade ecológica (*ecologic validity*) ou ambiente natural relaciona como os resultados estatísticos obtidos poderão ser aplicados a situações reais fora do ambiente de investigação. Este conceito está próximo da validade

[3] Duplamente ocultado (*double blind*) significa que enquanto a experiência científica durar, ambos participantes e os intervenientes da experiência não sabem se pertencem ao grupo de controlo ou grupo experimental.
[4] Confundimento (*confounding*) significa que uma variável externa se correlaciona simultaneamente com a variável dependente e independente.

externa, mas o seu objetivo é demonstrar como um estudo experimental pode ser reproduzido numa situação real ou num ambiente natural.

Os fatores que mais influenciam a validade interna são a existência de fatores históricos/experimentais ligados ao desenvolvimento do estudo e que afetam o dia-a-dia dos participantes desse estudo (exemplo: situações de crise), a maturação dos sujeitos do estudo, a existência de variáveis de confundimento, o viés na seleção dos participantes para o estudo ou para os grupos de investigação em análise, a repetição do teste (os participantes podem se lembrar das respostas dadas ou serem condicionados pelo facto de saberem que vão ser novamente testados), a mudança do instrumento de medição, a utilização de recompensas ou castigos, a perda de sujeitos por desistência do estudo e, finalmente, o viés do investigador.

Os fatores que mais influenciam a validade externa são: a seleção da amostra utilizada no estudo (amostragem aleatória versus não aleatória) e a sua representatividade na população em estudo (tamanho da amostra utilizado). Outros fatores que influenciam a generalização dos resultados são o efeito de contágio entre os grupos experimentais e o grupo de controlo, o efeito da reatividade (reação dos participantes ao fato de serem estudados), a interação entre a intervenção e as condições experimentais, a existência da interferência de tratamentos múltiplos, e finalmente uma relação causal ambígua.

A validade das conclusões estatísticas (validade externa) indica o grau em que as conclusões obtidas baseadas nos dados obtidos são corretas ou razoáveis. Geralmente pode ser aferida através do uso adequado de técnicas de amostragem corretas, escolha acertada dos testes estatísticos, a indicação das medidas de fiabilidade sobre os instrumentos utilizados e o controlo dos erros tipo I e tipo II.

Segundo Pasquali (2007), o número de diferentes tipos de validade que se encontram reportados da literatura é vasto, apresentando 31 tipos diferentes e convidando outros investigadores a aumentarem o número apresentado (Pasquali, 2007). Como é referido pelo próprio autor, este número indica a complexidade e a importância que o conceito de validade de um instrumento/experiência assume.

Referências

Alexandre, N. M. C., & Coluci, M. Z. O. (2011). Validade de conteúdo nos processos de construção e adaptação de instrumentos de medidas. *Ciência & Saúde Coletiva*, *16*, 3061–3068. doi:10.1590/S1413-81232011000800006

Berk, R. a. (1990). Importance of expert judgment in content-related validity evidence. *Western Journal of Nursing Research*, *12*(5), 659–671. doi:10.1177/019394599001200507

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105. doi:10.1037/h0046016

Carmines, E., & Zeller, R. (1979). *Reliability and validity assessment*. Sage Publications. doi:http://dx.doi.org/10.4135/9781412985642

Cicchetti, D. V, & Feinstien, a R. (1990). High Agreement but Low Kappa. *Journal of Clinical Epidemiology*, *43*, 551 – 585.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302. doi:10.1037/h0040957

Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research*, *5*(4), 194–197. doi:10.1016/S0897-1897(05)80008-4

Fleiss, J. (2004). *Statistical Methods for Rates and Proportions*. *Technometrics* (Vol. 46, pp. 263–264). New York: John Wiley. doi:10.1198/tech.2004.s812

Hair, J., Black, W., Babin, B., & Anderson, R. (2009). *Multivariate data analysis*.

Kirk, J., & Miller, M. L. (1986). *Reliability and Validity in Qualitative Research* (p. 87). Sage Publications. doi:10.4135/9781412985659

Litwin, M. S. (1995). *How to Measure Survey Reliability and Validity* (pp. 5 – 8). Sage Publications. doi:10.4135/9781483348957

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, *35*, 382–385. doi:10.1097/00006199-198611000-00017

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. doi:10.1037/0003-066X.50.9.741

Pasquali, L. (2007). Validade dos Testes Psicológicos : Será Possível Reencontrar o Caminho ? The Validity of the Psychological Tests : Is It Possible to Find the Way Again ? A Confusão do Conceito Validade, *23*, 99–107.

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing and Health*, *29*, 489–497. doi:10.1002/nur.20147

Teasdale, G., & Jennett, B. (1974). Assessment of coma and impaired consciousness. A practical scale. *Lancet*, *2*, 81–84. doi:10.1016/S0140-6736(74)91639-0

Till, A. (1989). *Measuring Health — a Guide To Rating Scales and Questionnaires*. *The Journal of the Canadian Chiropractic Association* (2nd ed., Vol. 33, p. 208). New York: Oxford University Press Inc. doi:10.1179/108331900786166731

Capítulo 2: Métodos para a estimação de concordância e de fiabilidade em análises com variáveis nominais

Neste capítulo iremos descrever com detalhe os métodos mais utilizados para medir a concordância e a fiabilidade entre avaliadores, quando estamos na presença de variáveis medidas numa escala nominal e em função do número de avaliadores (Shoukri, 2010).

Para medir a concordância os métodos utilizados serão as percentagens de concordância e para a fiabilidade utilizaremos as estatísticas Kappa.

2.1 Percentagem de concordância

Os dados nominais obtidos de num estudo com dois avaliadores (*inter-rater study*) são geralmente apresentados numa tabela de contingência *qxq*, onde *q* designa o número de categorias não sobrepostas em que um determinado sujeito pode ser classificado.

A situação mais simples é quando temos uma avaliação dicotómica (sim/não, doente/não doente, geralmente codificado como 1/0) originando uma tabela 2x2 (Shoukri, 2010). Na tabela 2.1 estão resumidas as pontuações (*ratings*) de dois avaliadores, onde n_{11}, n_{10}, n_{01} e n_{00} denotam as frequências absolutas observadas para cada possível combinação de classificações dos avaliadores A e B. Os totais apresentados podem ser descritos como frequências marginais, bastando para isso, dividi-las pela dimensão da amostra.

A proporção de concordância P_a é a proporção de casos em que os avaliadores A e B concordam e é dada por:

$$P_a = \frac{n_{11} + n_{00}}{n} \qquad 2.1$$

ou seja, P_a é o quociente entre o número de respostas concordantes e o número total de respostas.

Tabela 2.1. Tabela básica 2×2 para dois avaliadores

		Avaliador A		
		Doente(1)	Não Doente(0)	Total
Avaliador B	Doente(1)	n_{11}	n_{10}	$n_{1+}= n_{11}+ n_{10}$
	Não Doente(0)	n_{01}	n_{00}	$n_{0+}= n_{01}+ n_{00}$
	Total	$n_{+1}= n_{11}+ n_{01}$	$n_{+0}= n_{10}+ n_{00}$	n

Esta proporção é informativa e útil, mas usada por si só tem limitações. Por exemplo, numa aplicação epidemiológica onde uma classificação positiva corresponde a um diagnóstico positivo para uma doença muito rara, com uma prevalência de 1 em 1000000. Nesta situação iremos obter um valor de P_a muito elevado, acima de 0.99, e este resultado deve-se unicamente a um acordo sobre a ausência da doença, não nos informando diretamente sobre a capacidade de diagnosticar corretamente a doença.

Para ilustrar o cálculo da proporção de concordância (P_a) em situações dicotómicas, vamos considerar o exemplo hipotético extraído de (Gwet, 2010) (exemplo2.1)

Exemplo 2.1. Suponhamos que é efetuado um estudo em que dois médicos (avaliadores) pretendem determinar a utilidade de um instrumento para diagnosticar uma doença em 100 pacientes (tabela 2.2). Os dados foram retirados do estudo apresentado por Gwet (Gwet, 2010)

Tabela 2.2. Resultados de dois avaliadores sobre a utilidade de um instrumento

		Avaliador A		
		Sim	Não	Total
Avaliador B	Sim	35	20	55
	Não	5	40	45
Total		40	60	100

A tabela 2.2 indica que os avaliadores A e B classificam 35 dos 100 indivíduos na categoria 1, e 40 dos 100 indivíduos na categoria 2 ou seja o avaliador A e o Avaliador B concordam que o instrumento é útil em 35% do tempo e não útil em 40% das vezes. No entanto eles estão em desacordo na classificação de 25 indivíduos. Neste caso a percentagem de concordância P_a, é dada por

$$P_a = \frac{35 + 40}{100} = 0.75$$

Duma forma geral podemos ter dois avaliadores e um número de categorias superior a 2 (tabela 2.3). Seja n_{ij} o número de casos atribuído pelo avaliador A à categoria *i* e à categoria *j* pelo avaliador B, onde *i,j*=1,2,…,*q* e *n* indica o número total de observações.

Tabela 2.3 Tabela básica *q*×*q* para dois avaliadores

	Avaliador B					
Av. A	1	2	3	…	q	total
1	n_{11}	n_{12}	n_{13}	…	n_{1q}	$n_{1+}=\sum_{j=1}^{q} n_{1,j}$
2	n_{21}	n_{22}	n_{23}	…	n_{2q}	$n_{2+}=\sum_{j=1}^{q} n_{2,j}$
3	n_{31}	n_{32}	n_{33}		n_{3q}	$n_{3+}=\sum_{j=1}^{q} n_{3,j}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
q	n_{q1}	n_{q2}	n_{q3}	…	n_{qq}	$n_{q+}=\sum_{j=1}^{q} n_{q,j}$
Total	$n_{+1}=\sum_{i=1}^{q} n_{i,1}$	$n_{+2}=\sum_{i=1}^{q} n_{i,2}$	$n_{+3}=\sum_{i=1}^{q} n_{i,3}$	…	$n_{+q}=\sum_{i=1}^{q} n_{i,q}$	n

Neste caso a proporção de concordância é dada pelo quociente entre a soma das frequências da diagonal da tabela pelo total de casos, ou seja:

$$P_a = \frac{1}{n}\sum_{i=1}^{q} n_{ii} = \sum_{i=1}^{q} p_{ii} \qquad 2.2$$

onde p_{ii} representa a frequência relativa n_{ii}/n.

Exemplo 2.2. Na tabela 2.4 estão apresentados os diagnósticos efetuados por 2 médicos do tipo de síndrome em função das dores de coluna que os pacientes apresentam. Os dados foram adaptados do estudo apresentado por Gwet (Gwet, 2010) de modo a se obter um valor de concordância mais elevado.

Tabela 2.4. Avaliações efetuadas pelos 2 médicos no diagnóstico de um determinado tipo de síndrome

	Médico B			
Médico A	Síndrome degenerativo	Síndrome disfuncional	Síndrome Postural	Total
Síndrome degenerativo	31	1	2	34
Síndrome Disfuncional	3	37	4	44
Síndrome Postural	2	1	21	24
Total	36	39	27	102

A percentagem de concordância é dada por:

$$P_a = \frac{1}{n}\sum_{i=1}^{3} n_{ii} = \frac{31 + 37 + 21}{102} = \frac{89}{102} \approx 0.873$$

Tal como na situação anterior, também temos uma elevada proporção de concordância, uma vez que dos 102 pacientes avaliados, ambos os médicos concordam no diagnóstico da mesma síndrome em 89 dos pacientes (para ambos, 31 pacientes têm o Síndrome degenerativo, 37 pacientes, o Síndrome disfuncional e 21 pacientes o Síndrome postural). No entanto, um facto conhecido deste tipo de problemas é que algumas destas 89 concordâncias podem ocorrer puramente devido ao acaso. O resultado disto é o inflacionar ou sobrestimar o verdadeiro valor da concordância entre os dois avaliadores.
Como não é possível identificar as concordâncias que ocorreram devido ao acaso e eliminá-las do cálculo da concordância, então tenta-se calcular uma correção de concordância devido ao acaso. Nas seções seguintes, serão apresentadas medidas de concordância que fazem esta correção da concordância devido ao acaso (Shoukri, 2010).

2.2 Dois avaliadores: Kappa de Cohen e o Kappa de Scott (ou π de Scott)

Atendendo a que as proporções de concordância não são sensíveis ao facto de existir uma certa quantidade de concordância baseada apenas no acaso, Cohen (Cohen, 1960) propôs a estatística Kappa como uma medida de concordância[1] que inclui uma correção. Esta consiste em estimar o valor esperado de concordância devido ao acaso utilizando a teoria das probabilidades. Este valor é obtido utilizando o produto das frequências marginais da tabela de contingência (representadas pelos totais de cada linha e de cada coluna apresentadas nas tabelas 2.1 e 2.3, respetivamente).

Atualmente, o coeficiente de Kappa de Cohen (K_{Cohen}) continua a ser amplamente utilizado, sendo os pressupostos básicos apresentados pelo autor para o seu cálculo: “(1) as unidades em análise são independentes; (2) as categorias da escala nominal são independentes, mutuamente exclusivas e exaustivas; (3) os avaliadores atuam independentemente (Cohen, 1960). Cada avaliador pode distribuir as unidades de análise pelas diferentes categorias

[1] Cohen propôs a estatística Kappa como uma medida de concordância, no entanto esta é uma medida de fiabilidade.(Kottner et al., 2011)

livremente, partindo-se do princípio que ambos os avaliadores são considerados igualmente aptos para a realização da tarefa.

A notação usualmente utilizada é apresentada em termos de frequência relativa e não tanto em frequência absoluta. Estes valores são calculados através dos respetivos quocientes entre as classificações atribuídas pelos avaliadores A e B e o número total de classificações.

O coeficiente Kappa de Cohen pode então ser definido como a proporção de acordo entre os avaliadores após ser retirada a proporção esperada de acordo devido ao acaso (*expected chance agreement*), exprimindo-se pela seguinte expressão:

$$K_{Cohen} = \frac{P_a - P_e}{1 - P_e} \qquad 2.3$$

onde P_a representa a proporção de concordância observado dado pela expressão 2.2 (no caso de uma matriz de contingência *qxq*). O valor de P_e que representa a proporção esperada de acordo devido ao acaso, ou seja, a proporção de unidades classificadas pelos avaliadores nas mesmas categorias por mera coincidência e é obtido pela soma dos produtos entre o total linha pelo total coluna, isto é:

$$P_e = \sum_{k=1}^{q} p_{k+}\, p_{+k} = \sum_{k=1}^{q} \frac{n_{k+}}{n} * \frac{n_{+k}}{n} \qquad 2.4$$

onde p_{k+} e p_{+k} representam as respetivas frequências relativas marginais.

O denominador do Kappa de Cohen representa a percentagem de indivíduos onde não seria de esperar qualquer acordo devido ao acaso, enquanto o numerador, segundo Cohen (Cohen, 1960) representa "…the percent of units in which beyond-chance agreement occurred…" . Cohen (1960) considera Kappa como "…the proportion of agreement after chance agreement is removed from consideration…" (Cohen, 1960)

Este coeficiente de fiabilidade só é aplicado para dados nominais ou ordinais, em matrizes de contingência quadradas (de ordem *qxq* não muito

elevada) e quando estamos na presença de apenas dois avaliadores ou de dois momentos (Cohen, 1960). A estatística Kappa de Cohen é uma medida que assume valores entre -1 e 1, onde 1 significa concordância perfeita e -1 uma discordância perfeita. O valor 0 verifica-se quando a concordância observada é exatamente a mesma da concordância esperada devido ao acaso (Cohen, 1960). Os valores positivos indicam que a concordância observada é maior que a concordância esperada devido ao acaso, enquanto os valores negativos representam a situação contrária (Cohen, 1960). Existe uma grande variação na interpretação dos valores de Kappa de Cohen. No ponto de vista de vários autores o mais abrangente é a interpretação proposta por Landis e Koch (Landis JR, 1977) apresentado na tabela 2.5.

Tabela 2.5. Interpretações dos valores de Kappa de Cohen sugerido por Landis e Koch (Landis JR, 1977).

Kappa	Interpretação
<0	Pobre
0.00-0.20	Fraco
0.21-0.40	Considerável
0.41-0.60	Moderado
0.61-0.80	Substancial
0.81-1.00	Quase Perfeito

Para ilustrar o cálculo do coeficiente de Kappa de Cohen, vamos voltar ao exemplo 2.1 e à tabela 2.2. O valor obtido para P_a (0.75) é considerado como um “bom” grau de concordância entre os avaliadores A e B. Cohen (1960) mostrou como ajustar P_a através do acordo devido ao acaso (P_e) para obter o coeficiente de Kappa:

$$P_e = p_{1+}p_{+1} + p_{2+}p_{+2} = \frac{55}{100} * \frac{40}{100} + \frac{45}{100} * \frac{60}{100} = \frac{49}{100} = 0.49$$

$$\widehat{K}_{Cohen} = \frac{0.75 - 0.49}{1 - 0.49} \approx 0.51$$

Atendendo à interpretação dada por Landis & Koch, representada na tabela 2.5, podemos indicar que os avaliadores estão em concordância moderada em relação à utilidade do instrumento, no diagnóstico da referida doença.

De seguida, com base no exemplo 2.2, em que os dados relativos à avaliação estão representados na tabela 2.4, onde P_a é igual a 0.87 e o P_e é dada por:

$$P_e = \sum_{k=1}^{3} p_{k+}\, p_{+k} = \frac{34}{102} * \frac{36}{102} + \frac{44}{102} * \frac{39}{102} + \frac{24}{102} * \frac{27}{102} \approx 0.34$$

$$\widehat{K}_{cohen} = \frac{0.87 - 0.34}{1 - 0.34} \approx 0.80$$

Segundo a tabela 2.5, pode-se observar que os 2 avaliadores (médicos), estão em concordância substancial no que diz respeito ao diagnóstico do síndrome em função das dores de coluna dos pacientes.

O Kappa de Scott (em homenagem a William A. Scott, K_{Scott}) é uma estatística semelhante ao Kappa de Cohen, para variáveis nominais ou ordinais, apresentada antes do Kappa de Cohen, em 1955 por William A. Scott. Ambos são calculados para uma situação onde só existem dois avaliadores e para uma matriz de contingência quadrada. A diferença entre os dois Kappas é no cálculo da frequência esperada devido ao acaso:

$$P_e = \sum_{k=1}^{q} \pi_k^2 \qquad 2.5$$

$$\pi_k = \frac{(p_{k+} + p_{+k})}{2} \qquad 2.6$$

Baseando-nos no exemplo 2.1, os valores de P_e e do Kappa de Scott são calculados, respetivamente:

$$P_e = \left(\frac{\frac{40}{100} + \frac{55}{100}}{2}\right)^2 + \left(\frac{\frac{60}{100} + \frac{45}{100}}{2}\right)^2 \approx 0.50$$

$$\widehat{\mathrm{K}}_{\mathrm{Scott}} = \frac{0{,}75 - 0{,}50}{1 - 0{,}50} \approx 0.50$$

No exemplo 2.2, os valores de P_e e do Kappa de Scott são calculados, respetivamente:

$$P_e = \left(\frac{\frac{36}{102} + \frac{34}{102}}{2}\right)^2 + \left(\frac{\frac{39}{102} + \frac{44}{102}}{2}\right)^2 + \left(\frac{\frac{27}{102} + \frac{24}{102}}{2}\right)^2 \approx 0.35$$

$$\widehat{\mathrm{K}}_{\mathrm{Scott}} = \frac{0.87 - 0.35}{1 - 0.35} \approx 0.80$$

Convém referir que para o mesmo exemplo, o coeficiente Kappa de Cohen e o Kappa de Scott tendem a produzir resultados semelhantes.

2.3 Mais de dois avaliadores: Kappa de Fleiss e o kappa de Conger

Uma extensão do Kappa de Cohen foi desenvolvida por Fleiss (Fleiss, 1971), para o caso de mais do que dois avaliadores em simultâneo. No entanto, prova-se que o Kappa de Fleiss (K_{Fleiss}) é a generalização do π de Scott para mais do que 2 avaliadores e não do Kappa de Cohen. A generalização do Kappa de Cohen foi desenvolvida por Conger (Conger, 1980), designado por Kappa de Conger (K_{Conger}). A tabela 2.6 mostra a distribuição dos r avaliadores pelos n indivíduos nas q categorias disponíveis. O cálculo da concordância observada (P_a) destes dois Kappas entre os avaliadores é dada por:

$$P_a = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{q}\frac{r_{ik}(r_{ik}-1)}{r_i(r_i-1)} \qquad 2.7$$

onde r_{ik} é o número de avaliadores que atribuíram um determinado valor x_k ao sujeito i e r_i o número de avaliadores que avaliaram o sujeito i com q categorias possíveis. Note-se que poderão existir valores em falta, daí o valor de r_i não ser

sempre o mesmo (constante). Se não existirem valores em falta, ri poderá ser simplesmente substituído por r.

Tabela 2.6: Distribuição dos *r* avaliadores por *n* Sujeitos e *q* categorias de resposta

	Categoria de resposta					Total
Sujeitos	1	…	*k*	…	*q*	
1	r_{11}	…	r_{1k}	…	r_{1q}	r_1
…	…		…		…	…
i	r_{i1}	…	r_{ik}	…	r_{iq}	r_i
…	…		…		…	…
n	r_{n1}	…	r_{nk}	…	r_{nq}	r_n
Média	$\bar{r}_{.1}$	…	$\bar{r}_{.k}$	…	$\bar{r}_{.q}$	$\bar{r}$

A concordância esperada devido ao acaso é calculada de forma diferente para o Kappa de Fleiss ou para o Kappa de Conger. No caso do Kappa de Fleiss, o valor de P_e é dado pela expressão:

$$P_e = \sum_{k=1}^{q} \pi_k^2 \qquad 2.8$$

$$\pi_k = \frac{1}{n} \sum_{i=1}^{n} \frac{r_{ik}}{r_i} \qquad 2.9$$

A concordância esperada pelo Kappa de Conger necessita de uma segunda tabela (Tabela 2.7), que indica a distribuição dos n indivíduos por r avaliadores nas q categorias disponíveis.

Tabela 2.7: Distribuição dos *n* sujeitos por *r* avaliadores e *q* categorias de resposta

	Categoria de resposta					Total
Avaliador	1	…	*k*	…	*q*	
1	n_{11}	…	n_{1k}	…	n_{1q}	n_1
…	…		…		…	…
g	n_{g1}	…	n_{gk}	…	n_{gq}	n_g
…	…		…		…	…
r	n_{r1}	…	n_{rk}	…	n_{rq}	n_r
Média	$\bar{n}_{.1}$	…	$\bar{n}_{.k}$	…	$\bar{n}_{.q}$	$\bar{n}$

Considerando $p_{gk}=n_{gk}/n_g$, a proporção de indivíduos que o avaliador *g* classifica na categoria *k*, e S^2_k a variância da amostra das *r* proporções $p_{1k},\ldots,p_{rk}$, então a concordância esperada devido ao acaso para o Kappa de Conger é dada por:

$$P_e = \sum_{k=1}^{q} \bar{p}_{+k}^2 - \sum_{k=1}^{q} s_k^2/r \qquad 2.10$$

$$\bar{p}_{+k} = \frac{1}{r}\sum_{g=1}^{r} p_{gk} \qquad 2.11$$

$$s_k^2 = \frac{1}{r-1}\sum_{g=1}^{r}\left(p_{gk} - \bar{p}_{+k}\right)^2 \qquad 2.12$$

O coeficiente Kappa para múltiplos avaliadores é dado pela expressão 2.3, para ambos os métodos. O exemplo 2.3 foi adaptado (Gwet, 2010),com a finalidade de ilustrar as várias etapas do cálculo dos coeficientes Kappa de Fleiss e Kappa de Conger.

Exemplo 2.3. Supondo que temos 4 médicos (r) que pretendem diagnosticar uma determinada doença (com 5 categorias: a, b, c, d, e), em 12 indivíduos (n) selecionados aleatoriamente. Os dados estão representados na tabela 2.8 e os cálculos nas tabelas 2.9 e 2.10.

Tabela 2.8. Diagnóstico atribuído pelos 4 médicos aos 12 pacientes.

Pacientes	Avaliador 1	Avaliador 2	Avaliador 3	Avaliador 4
1	a	a	b	a
2	b	b	c	b
3	c	c	c	c
4	c	c	c	c
5	b	b	b	b
6	a	b	c	d
7	d	d	d	d
8	a	a	b	a
9	b	b	b	b
10	e	e	e	e
11	e	e	a	a
12	b	b	c	b

Tabela 2.9. Distribuição das classificações dos 4 médicos por individuo e categoria (doença)

	Doença					
Pacientes	a	b	c	d	e	P_i
1	3	1	0	0	0	0.50
2	0	3	1	0	0	0.50
3	0	0	4	0	0	1.00
4	0	0	4	0	0	1.00
5	0	4	0	0	0	1.00
6	1	1	1	1	0	0.00
7	0	0	0	4	0	1.00
8	3	1	0	0	0	0.50
9	0	4	0	0	0	1.00
10	0	0	0	0	4	1.00
11	2	0	0	0	2	0.33
12	0	3	1	0	0	0.50
p_k	0.19	0.35	0.23	0.10	0.13	

Tabela 2.10. Distribuição das classificações dos 12 sujeitos por médico e categoria (doença)

	Doença					
Avaliadores	a	b	c	d	e	Total
1	3	4	2	1	2	12
2	2	5	2	1	2	12
3	1	4	5	1	1	12
4	3	4	2	2	1	12
Média	2.25	4.25	2.75	1.25	1.50	

Para o exemplo dado, temos, *n*=12, *r*=4 e *k*=5, deste modo para cada *k*=1,2, …5 obtém-se para as colunas:

$$p_1 = \frac{3+0+0+0+0+1+0+3+0+0+2+0}{12*4} = 0.1875 \approx 0.19$$

e assim sucessivamente, obtemos p_2, p_3, p_4 e p_5 encontrando-se os seus resultados na ultima linha da tabela 2.9. Para cada *i*=1,2, … ,12,obtém-se para as linhas:

$$P_1 = \frac{1}{4*(4-1)}\left((3^2+1^2+0^2+0^2+0^2)-4\right) = 0.50$$

Note que P_3=P_4=P_5=P_7=P_9=P_{10}=1, o que significa que todos os 4 médicos diagnosticam a mesma doença aos pacientes 3,4,5,7,9 e 10,respectivamente, estando em concordância perfeita. No entanto em relação ao paciente 6 estão em total discordância, daí P_6 ser nulo.

Deste modo a proporção de concordância observada para ambos os kappas é dada por:

$$P_a = \frac{1}{12}(0.5+0.5+1+1+1+0+1+0.5+1+1+0.3+0.5) = 0.69$$

Quanto à concordância esperada devido ao acaso, para o Kappa de Fleiss é dada pela expressão 2.8:

$$P_{eF} = 0.19^2 + 0.35^2 + 0.23^2 + 0.10^2 + 0.13^2 \approx 0.24$$

e o respectivo valor do coeficiente Kappa de Fleiss é dado por:

$$\widehat{K}_{Fleiss} = \frac{0.69 - 0.24}{1 - 0.24} \approx 0.59$$

Para o Kappa de Conger, o valor da estimativa da concordância esperada devido ao acaso exige bastantes mais cálculos (tabela 2.11). Para obter o valor da variância, primeiro vamos obter os valores da proporção de pacientes que cada médico classifica em cada uma das categorias, considerando $p_{gk}=n_{gk}/n_{g.}$

A proporção de pacientes que o medico 1 classificou na categoria a, é $p_{11} = n_{11}/n_1 = 3/12 = 0.25$, e assim sucessivamente. Em seguida calculamos os valores de $\bar{p}_{.k}$, que são obtidos calculando a média para cada k com *k*=1, 2, 3, 4, 5. Assim, para $k = 1$, $\bar{P}_{.1} = (p_{11} + p_{21} + p_{31} + p_{41})/4 = 0.1875,$ e assim sucessivamente.

Tabela 2.11. Distribuição das classificações dos 12 sujeitos por médico e categoria (doença)

	Doença					
Avaliadores	a	b	c	d	e	Soma
1	0.25	0.33	0.17	0.08	0.17	
2	0.17	0.42	0.17	0.08	0.17	
3	0.08	0.33	0.42	0.08	0.08	
4	0.25	0.33	0.17	0.17	0.08	
$\bar{p}_{\cdot k}$	0.19	0.35	0.23	0.10	0.13	1
S_k^2	0.01	0.00	0.02	0.00	0.00	0.03
$\bar{P}_k^2$	0.04	0.13	0.05	0.01	0.02	0.24

A concordância esperada devido ao acaso será dada pela expressão 2.10:

$$P_e = 0.24 - \frac{0.03}{4} = 0.23$$

e o coeficiente Kappa de Conger é dado por:

$$\widehat{K}_{Conger} = \frac{0.69 - 0.23}{1 - 0.23} = 0.60$$

Como podemos observar pelo exemplo o Kappa de Conger exige muito mais cálculos, sendo esta a sua grande desvantagem em relação ao Kappa de Fleiss. A única vantagem deste coeficiente é ser uma extensão mais natural do Kappa de Cohen para o caso de 3 ou mais avaliadores.(Gwet, 2010)

Ambos os coeficientes tendem a ser idênticos á medida que o número de avaliadores aumenta (Gwet, 2010). Como acontece numa situação de 2 avaliadores, os valores de Kappa tende a tornar-se maior à medida que o número de categorias diminui, dado que a possibilidade de discordar é menor.

2.4 Kappa de Brennan-Prediger

O Kappa de Brennan-Prediger ($K_{B\text{-}P}$) foi apresentado por vários autores diferentes e com vários nomes diferentes, como por exemplo G-Index (Holley and Guilford(1964)). Nesta abordagem o cálculo do valor esperado de concordância devido ao acaso é igual ao inverso do número de categorias disponíveis para os avaliadores (q), simplificando as abordagens apresentadas anteriormente:

$$P_e = \frac{1}{q} \qquad 2.13$$

No caso de dois avaliadores, o valor de P_a será dado pela expressão 2.2 enquanto no caso de haver mais do que 2 avaliadores, o valor de P_a será dado pela expressão 2.7. O resultado do $K_{B\text{-}P}$ é dado pela expressão 2.3.

Os resultados deste Kappa podem ser muito diferentes dos Kappa anteriormente apresentados. No exemplo 2.1, o valor de $K_{B\text{-}P}$=0.50, no exemplo 2.2, $K_{B\text{-}P}$=0.81 e no exemplo 2.3, $K_{B\text{-}P}$=0.62. (Resultados mais detalhados estão apresentados no capitulo 7). A grande vantagem deste coeficiente está na simplificação do cálculo da concordância esperada devido ao acaso (inverso do número de categorias). Outra grande vantagem é que resolve um dos paradoxos da estatística Kappa, quando existe um acordo praticamente perfeito numa das categorias, como no caso do exemplo 2.4 da próxima seção.

2.5 Paradoxos do coeficiente Kappa

A estatística Kappa produz frequentemente valores que são inesperadamente baixos comparando com a percentagem de acordo global (P_a). Estas discrepâncias têm sido referidas na literatura como *os paradoxos da estatística Kappa*. Feinstein e Cicchetti (1990) fornecem uma discussão detalhada sobre dois desses paradoxos (Feinstein & Cicchetti, 1990). O uso das distribuições marginais, com o objetivo de quantificar o valor esperado de concordância devido ao acaso (P_e), está na origem dos paradoxos:

- Se o valor do P_e é elevado, o processo de cálculo do Kappa pode converter um elevado valor de concordância numa estatística Kappa reduzida (exemplo 2.4);
- Se a tabela de contingência produzida pelos avaliadores for assimétrica (ou não balanceada) então os valores da estatística kappa serão mais elevados do que se a tabela de contingência for “mais” simétrica (ou balanceada) (exemplo 2.5).

Exemplo 2.4. Supondo que dois avaliadores avaliam a utilidade (sim/não) de um instrumento na determinação de uma determinada doença em 100

pacientes. Neste caso, uma das categorias de concordância é muito superior em relação a uma segunda categoria e consequentemente os totais marginais são desequilibrados. Os resultados encontram-se na tabela 2.12.

Tabela 2.12. Resultados de dois avaliadores sobre a utilidade de um instrumento

		Avaliador B		
		Sim	Não	Total
Avaliador A	Sim	**95**	5	100
	Não	0	**0**	0
	Total	95	5	100

A tabela 2.12 sugere-nos que existe um acordo praticamente perfeito em relação á utilidade do instrumento proposto para diagnosticar a referida doença, pois só em 5 indivíduos é que os dois avaliadores discordam. No entanto, a proporção observada de concordância (P_a) é de 0.95 e a proporção esperada de concordância devida ao acaso (P_e) pela expressão 2.4 é também de 0.95, desta forma o coeficiente Kappa de Cohen associado a estes dados é nulo.

Este é um exemplo onde um investigador poderia esperar uma concordância quase perfeita entre os avaliadores, independentemente da forma em que são medidos, no entanto o coeficiente Kappa é nulo, o que sugere uma total ausência de acordo entre os avaliadores. Neste caso estamos perante um paradoxo, atendendo que o Kappa não os quantificou corretamente. Por outro lado, o valor da percentagem de concordância global observada é de 0.95, como seria de esperar, mas o valor da percentagem da concordância esperada devida ao acaso é também de 0.95, o que é totalmente inesperado. O uso das distribuições marginais para quantificar a proporção de concordância devida ao acaso pode não ser razoável no caso em que estas são muito desequilibradas para uma dada categoria.

Como se observou, o coeficiente Kappa é fortemente influenciado pela prevalência de um determinado atributo. Para uma situação em que os

avaliadores têm de escolher entre classificar casos como positivo ou negativo em relação a determinado atributo, um efeito de prevalência existe quando a proporção de concordância sobre a classificação positiva difere da classificação negativa. Esta situação pode ser expressa pelo índice de prevalência (Banerjee & Fielding, 1997):

$$\text{Indice de prevalência} = \frac{|n_{11} - n_{22}|}{n} \qquad 2.14$$

onde $|n_{11} - n_{22}|$ é o valor absoluto da diferença das células onde ambos os avaliadores concordam. Se o índice de prevalência for elevado, a proporção de acordo esperado devido ao acaso também será muito alta e o coeficiente Kappa será reduzido (Brennan & Silman, 1992). No caso apresentado na tabela 2.12, o índice de prevalência é alto:

$$\text{Indice de prevalência} = \frac{|95 - 0|}{100} = 0.95$$

Exemplo 2.5. Supondo, novamente que dois avaliadores avaliam a utilidade (sim/não) de um instrumento na determinação de uma determinada doença em 100 pacientes. Neste caso, o total das categorias que representam a concordância é semelhante, os totais marginais são também semelhantes, mas a tabela de contingência no caso A apresenta uma homogeneidade dos resultados nas categorias consideradas, o que não acontece no caso B. Os resultados encontram-se na tabela 2.13.

Tabela 2.13. Tabela de contingência que mostra divergências “mais” simétricas (esquerda) ou “menos” simétricas (direita).

		Av B (esquerda)			Av C (direita)		
		Sim	Não	Total	Sim	Não	Total
Av A	Sim	45	15	60	25	35	60
	Não	25	15	40	5	35	40
	Total	70	30	100	30	70	100

O valor de kappa para a tabela de contingência da esquerda é igual a 0.13 enquanto para a tabela da direita é de 0.26. A razão destes resultados está novamente relacionado com o cálculo do P_e (0.54 na tabela esquerda versus 0.46 na tabela direita). Tabelas de contingência mais assimétricas permitem a obtenção de valores de kappa superiores. Gwet (Gwet, 2010) sugere a utilização de kappa com pesos ponderados para minimizar o impacto das discordâncias.

Um índice sobre a discordância também pode ser calculado, designado por índice de viés. O viés é a medida em que os avaliadores discordam sobre a proporção de casos positivos ou negativos e é dado pela diferença das células n_{12} e n_{21}:

$$\text{Indice do Viés} = \frac{|n_{12} - n_{21}|}{n} \qquad 2.15$$

No caso apresentado na tabela 2.12, o índice de viés é baixo (0.05) enquanto na tabela 2.13 são considerados moderados (0.40). Quando o valor do índice de viés é alto, o coeficiente Kappa também aumenta, estando em contraste com o índice de prevalência (Byrt, Bishop, & Carlin, 1993).

2.6 Outros coeficientes Kappas

Na literatura são apresentados outros Kappas alternativos aos apresentados. O Kappa generalizado de Light (Light, 1971) é uma extensão do Kappa de Cohen para múltiplos avaliadores, envolvendo a média de todos os pares de avaliadores dois a dois, utilizando o Kappa de Cohen para esse efeito. Os coeficientes de BAK (Bias-Adjusted Kappa) e PABAK (Prevelance-Adjusted and Bias-Adjusted Kappa) (Byrt et al (1993)) são utilizados como uma tentativa de resolver os paradoxos do coeficiente Kappa. Outros coeficientes que tentam resolver estes paradoxos são propostos por Aickin (1990) e Gwet (2008) designados por α de Aickin e por AC1, respetivamente. Estes coeficientes são computacionalmente exigentes e estão claramente apresentados no capítulo 4 em Gwet (Gwet, 2010), não sendo apresentados nesta dissertação.

Referências

Banerjee, M., & Fielding, J. (1997). Interpreting kappa values for two-observer nursing diagnosis data. *Research in Nursing and Health*, *20*, 465–470.

Brennan, P., & Silman, a. (1992). Statistical methods for assessing observer variability in clinical measures. *BMJ (Clinical Research Ed.)*, *304*(6840), 1491–1494. http://doi.org/10.1136/bmj.304.6840.1491

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, *46*(5), 423–429. http://doi.org/10.1016/0895-4356(93)90018-V

Cohen, J. (1960). *A coefficient of agreement for nominal scales*. Educational and Psychological measurement,. http://doi.org/10.1177/001316446002000104

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543–549. http://doi.org/10.1016/0895-4356(90)90158-L

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*. http://doi.org/10.1037/h0031619

Gwet, K. L. (2010). *Handbook of Inter-Rater Reliability: the definitive guide to measuring the extent of agreement among raters*. *Gaithersburg, MD: STATAXIS Publishing Company*. Advanced Analytics, LLC.

Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., … Streiner, D. L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *International Journal of Nursing Studies*, *48*(6), 661–671. http://doi.org/10.1016/j.ijnurstu.2011.01.016

Landis JR, K. G. (1977). *The measurement of observer agreement for categorical data.* Biometrics.

Shoukri, M. M. (2010). *Measures of Interobserver Agreement and Reliability, Second Edition*.

Capitulo 3: Métodos para a estimação da fiabilidade para variáveis ordinais

O coeficiente de Cohen discutido no capítulo 2 é adequado somente para avaliações medidas numa escala nominal. Em escalas nominais, a classificação dos indivíduos nas várias categorias não tem uma estrutura de ordem, ou seja duas categorias nominais consecutivas são tão diferentes como a primeira e a ultima categoria. Mas se por exemplo, as categorias estão ordenadas de “muito baixo” até “muito alto”, então os coeficientes Kappa apresentados poderão subestimar drasticamente o grau de concordância entre os avaliadores (Gwet, 2010).

Neste capítulo iremos então descrever com algum detalhe os métodos mais utilizados para medir a fiabilidade quando estamos na presença de variáveis ordinais e em função do numero de avaliadores.

3.1 Ponderação (weights) para os coeficientes kappa

Em resposta à necessidade sentida por alguns investigadores em diferenciar o grau de discordância entre as diferentes categorias, surgiu a ideia de atribuir pesos diferentes a essas categorias discordantes. Quanto mais afastada estiver a categoria discordante da categoria concordante, menor será o peso atribuído a essa categoria.

Por exemplo, supondo que tínhamos quatro categorias para medir a utilidade de um instrumento na avaliação de uma doença: “muito útil ”, “ útil ”, ”neutro”, e ” nada útil” e dois avaliadores. Neste caso, a discordância entre um

avaliador classificar como "muito útil" enquanto outro classificar como “útil" não será muito relevante, mas se um classificar como "muito útil” enquanto o outro categorizar como “nada útil” (ou seja, nos opostos da escala), esta discordância será mais relevante.

Nestas situações o Kappa proposto por Cohen (1968) é ineficiente para analisar avaliações medidas numa escala ordinal. O próprio autor propôs a versão ponderada do kappa para corrigir esse problema. (Cohen, 1968). A passagem de um coeficiente kappa (K) para um coeficiente Kappa ponderado (K_W) permite atribuir diferentes pesos às discordâncias, tornando-se assim uma estatística preferível para dados com categorias ordenadas.(Cohen, 1968). Como referido anteriormente, o kappa ponderado atribui menos peso para o acordo quando as categorias estão mais afastadas. Uma discordância de “muito útil” versus “neutro” ainda seria considerado um acordo parcial, mas um desacordo de “muito útil” versus "nada inútil" seria contado como um total desacordo, sendo atribuído um peso muito baixo.

O Kappa ponderado é então um índice estatístico utilizado para determinar a fiabilidade quando as variáveis são ordinais e os resultados podem ser expressos por mais de duas categorias sendo considerado uma extensão do kappa de Cohen (dado que este pode ser utilizado em situações com variáveis nominais/categóricas ou ordinais. Enquanto o Kappa de Cohen não ponderado considera somente concordância ou discordância, o Kappa ponderado permite a atribuição de pesos às diferentes categorias, de tal forma que categorias semelhantes podem estar em acordo parcial.(Cohen, 1968)

O kappa ponderado tem as mesmas limitações dos Kappas não ponderados. Esta estatística é adequada quando temos entre 3 e 10 categorias ordinais e o tamanho mínimo da amostra necessário para se poder aproximar a uma distribuição normal é de $2*q^2$ onde q é o número de categorias (Domenic V. Cicchetti & Feinstein, 1990). Os pesos estão compreendidos num intervalo $0 \leq w_{kl} \leq 1$, onde k=1,2,…,q e l=1,2,…,q. O peso máximo será atribuído quando o acordo entre os dois avaliadores é exato, isto é, $W_{kk} = 1$, e a todos os desacordos será atribuído um peso com um valor inferior ao peso máximo (D. V. Cicchetti, 1981).

Os pesos mais utilizados obtêm-se utilizando uma ponderação quadrática (Streiner, 1995) e para uma matriz *qxq* são definidos por:

$$w_{kl} = \begin{cases} 1 - \dfrac{(x_k - x_l)^2}{(q-1)^2}, se\ k \neq l \\ 1, \qquad\qquad se\ k = l \end{cases} \tag{3.1}$$

onde x_k e x_l representam o valor numérico da linha *k* e da coluna *l*. Independentemente de Cohen (1960), Cicchetti e Allison (1971), (D. Cicchetti & Allison, 1971) propuseram uma formulação para pesos lineares:

$$w_{kl} = \begin{cases} 1 - \dfrac{|x_k - x_l|}{|q-1|}, se\ k \neq l \\ 1, \qquad\qquad se\ k = l \end{cases} \tag{3.2}$$

Quando as categorias são ordinais, (Gwet, 2010), sugere a utilização de pesos ordinais definidos através da relação:

$$w_{kl} = \begin{cases} 1 - \dfrac{M_{kl}}{M_{max}}, se\ k \neq l \\ 1, \qquad\qquad se\ k = l \end{cases} \tag{3.3}$$

onde $M_{kl} = C_2^{\max(k,l)-\min(k,l)+1}$ e $M_{max} = C_2^q$. Se os dados medidos forem quantitativos numa escala de razões, o mesmo autor sugere a utilização da seguinte relação para o cálculo dos pesos:

$$w_{kl} = \begin{cases} 1 - \dfrac{[(x_k - x_l)/(x_k + x_l)]}{[(q-1)/(q+1)]}, se\ k \neq l \\ 1, \qquad\qquad\qquad se\ k = l \end{cases} \tag{3.4}$$

Outras formulações para o cálculo dos pesos podem ser encontradas em (Gwet, 2010) como pesos com base na raiz quadrada (como alternativa aos pesos quadráticos ou lineares), pesos circulares (se a variável medida for angular em graus ou radianos) e pesos bipolares (que tem comportamento idêntico aos pesos numa escala de razões no centro da escala e um

comportamento idêntico aos pesos quadráticos quando se afasta do centro da escala). Os pesos quadráticos, lineares, numa escala ordinal e numa escala de razões para 3 categorias são apresentados na tabela 3.1.

Tabela 3.1: Pesos quadráticos (topo esquerdo), lineares (topo direito), numa escala ordinal (inferior esquerdo) e em escala de razão (inferior direito) para uma escala com 3 categorias pelo menos ordinais

	Categorias (pesos quadráticos)				Categorias (pesos lineares)		
Categorias	A	B	C		A	B	C
A	1	0.75	0		1	0.50	0
B	0.75	1	0.75		0.50	1	0.50
C	0	0.75	1		0	0.50	1
	Categorias (pesos numa escala ordinal)				Categorias (pesos numa escala de razões)		
Categorias	A	B	C		A	B	C
A	1	0.67	0		1	0.56	0
B	0.67	1	0.67		0.56	1	0.84
C	0	0.67	1		0	0.84	1

Como acontece para as estatísticas não ponderadas apresentadas no capitulo 2, os Kappa ponderados são calculados de forma similar, bastando para isso corrigir a proporção de concordância e a proporção esperada devido ao acaso através uma matriz de pesos (Fleiss, Levin, & Cho Paik, 2003) e (Cohen, 1968).

3.2 Kappa ponderado para 2 avaliadores

Para o cálculo do Kappa de Cohen ponderado (weighted Kappa, K_{CW}) é necessário calcular a proporção de concordância observada ponderada (P_{aw}) e

a proporção de acordo devido ao acaso ponderado (P_{aw}). O cálculo (P_{aw}) é dado pela seguinte relação:

$$P_{aw} = \sum_{k=1}^{q} \sum_{l=1}^{q} w_{kl} p_{kl} \qquad 3.5$$

onde as proporções p_{kl} representam as avaliações (concordantes e discordantes) dadas entre os dois avaliadores. O valor obtido desta relação é interpretado como a percentagem de concordância ponderada (ver secção 2.1). A proporção ponderada de acordo devida ao acaso (P_{ew}) é dada por:

$$P_{ew} = \sum_{k=1}^{q} \sum_{l=1}^{q} w_{kl} p_{k+} p_{+l} \qquad 3.6$$

onde as proporções p_{k+} e p_{+l} representam as respetivas frequências marginais (ver expressão 2.4). Consequentemente, o Kappa de Cohen ponderado é dado por:

$$K_{CW} = \frac{P_{aw} - P_{ew}}{1 - P_{ew}} \qquad 3.7$$

Convém salientar que quando todos os desacordos são considerados igualmente graves, ou seja w_{kl}=0 para todo o $k \neq l$ e w_{kl}=1 para todo o k=l, então o kappa ponderado é idêntico ao kappa não ponderado dado pela expressão 2.3 do capitulo 2.

O kappa de Scott (K_{SW}) e o kappa de Brennan-Prediger (K_{BP}) nas suas versões ponderadas irão ser apresentados pela sua relativa importância como alternativas ao kappa de Cohen. O kappa de Scott ponderado é dado pelas seguintes expressões:

$$K_{SW} = \frac{P_{aw} - P_{ew}}{1 - P_{ew}} \qquad 3.8$$

$$P_{aw} = \sum_{k=1}^{q}\sum_{l=1}^{q} w_{kl}p_{kl} \quad 3.9$$

$$P_{ew} = \sum_{k=1}^{q}\sum_{l=1}^{q} w_{kl}\pi_k\pi_l \quad 3.10$$

onde $\pi_k = (p_{k+} + p_{+k})/2$. O kappa de Brennan-Prediger ponderado é dado por:

$$K_{BPW} = \frac{P_{aw} - P_{ew}}{1 - P_{ew}} \quad 3.11$$

$$P_{aw} = \sum_{k=1}^{q}\sum_{l=1}^{q} w_{kl}p_{kl} \quad 3.12$$

$$P_{ew} = \frac{1}{q^2}\sum_{k=1}^{q}\sum_{l=1}^{q} w_{kl} \quad 3.13$$

A interpretação da magnitude dos valores do kappa ponderado é idêntica à do kappa não ponderado e os valores esperados do Kappa ponderado tendem a ser maiores do que os valores do Kappa não ponderado, independentemente do estimador kappa utilizado (Soeken & Prescott, 1986). No exemplo que se segue, vamos ilustrar o cálculo do coeficiente Kappa ponderado e não ponderado

Exemplo 3.1: Consideremos o conjunto de dados apresentados na tabela 3.2, onde 2 Avaliadores denominados por Avaliador 1 e Avaliador 2, têm de pontuar 11 indivíduos em cada uma das 3 possíveis categorias, denotadas por A, B e C do tipo ordinais.

Tabela 3.2. Avaliação dos 11 indivíduos pelos 2 avaliadores

Indivíduos	Avaliador 1	Avaliador 2
1	A	B
2	B	C
3	C	C
4	C	C
5	B	B
6	B	A
7	A	A
8	A	B
9	B	B
10	B	B
11	A	A

A Tabela 3.3 mostra a distribuição dos indivíduos por avaliador e inclui a proporção dos totais marginais, este tipo de tabela é muito útil quando a experiência envolve um grande número de indivíduos a serem avaliados.

Tabela 3.3: Distribuição dos indivíduos por avaliador

		Avaliador 2			
Aval 1	A	B	C	Total	p_{i+}
A	2	2	0	4	0.36
B	1	3	1	5	0.46
C	0	0	2	2	0.18
Total	3	5	3	11	
p_{+j}	0.27	0.46	0.27		1

A tabela 3.1 apresenta os pesos quadráticos associados às categorias A, B, e C. Decorre desta tabela que todos os pesos da diagonal são iguais a 1 o que representa a concordância perfeita, enquanto os elementos fora da diagonal têm um peso 0 ou 0.75, representando uma concordância parcial. De

seguida apresentaremos a tabela 3.4, das proporções conjuntas das classificações dadas pelos 2 avaliadores nas 3 categorias.

Tabela 3.4: proporções conjuntas das classificações dos avaliadores 1 e 2 nas 3 categorias

	Avaliador 2				
Aval 1	A	B	C	Total	p_{i+}
A	0.18	0.18	0	4	0,36
B	0.09	0.27	0.09	5	0.46
C	0	0	0.18	2	0.18
Total	3	5	3	11	
p_{+j}	0.27	0.46	0.27	1	

Deste modo, e atendendo às expressões 3.5 e 3.6

$$P_{aw} = \sum_{i=1}^{3}\sum_{j=1}^{3} W_{ij}p_{ij} = 0.91$$

$$P_{ew} = \sum_{i=1}^{3}\sum_{j=1}^{3} W_{ij}p_{i+}p_{+j} = 0.73$$

Consequentemente o valor do coeficiente do Kappa ponderado é dado por:

$$\widehat{K}_{CW} = \frac{P_{aw} - P_{ew}}{1 - P_{ew}} = \frac{0.91 - 0.73}{1 - 0.73} = 0.67$$

O Kappa não ponderado (*unweighted Kappa*), dado pela expressão 2.3 é dado por:

$$\widehat{K}_C = \frac{P_a - P_e}{1 - P_e} = \frac{0.64 - 0.36}{1 - 0.36} = 0.44$$

De forma análoga seria possível produzir os resultados para o kappa de Scott e para o B-P, nos casos ponderado e não ponderado, que será efetuado no capitulo 7.

3.3 Kappa ponderado para mais do que 2 avaliadores e q categorias

Nesta secção irão ser apresentados os kappas ponderados para situações com mais do que 2 avaliadores. Os kappas apresentados nesta secção serão os kappa de Conger, de Fleiss e de Brennan-Prediger (B-P). As suas versões não ponderadas foram apresentadas no capítulo anterior, e por isso só se irá apresentar as modificações necessários para o seu cálculo ponderado.

Como referido no capítulo anterior, o cálculo da percentagem de concordância é idêntico para todos os métodos e a sua versão ponderada irá ser dada pela seguinte relação:

$$P_a = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{q}\frac{r_{ik}^{*}(r_{ik}-1)}{r_i(r_i-1)} \qquad 3.14$$

$$r_{ik}^{*} = \sum_{l=1}^{q} w_{kl} r_{il} \qquad 3.15$$

onde n é o número de sujeitos avaliados por dois ou mais avaliadores.

O que difere nestes métodos de concordância é a forma como a proporção esperada de acordo devido ao acaso é calculada. No caso do Kappa de Fleiss, o valor de P_e é dado pela expressão:

$$P_e = \sum_{k=1}^{q}\sum_{l=1}^{q} w_{kl}\pi_k\pi_l \quad 3.16$$

$$\pi_k = \frac{1}{n}\sum_{i=1}^{n}\frac{r_{ik}}{r_i} \quad 3.17$$

No caso do Kappa de Conger, o valor de P_e é dado pelas expressões:

$$P_e = \sum_{k=1}^{q}\sum_{l=1}^{q}\left(\bar{p}_{+k}\bar{p}_{+l} - s_k^2/r\right) \quad 3.18$$

$$\bar{p}_{+k} = \frac{1}{r}\sum_{g=1}^{r} p_{gk} \quad 3.19$$

$$p_{gk} = \frac{n_{gk}}{n_g}$$

$$s_{kl}^2 = \frac{1}{r-1}\left(\sum_{g=1}^{r} p_{gk}p_{gl} - r\bar{p}_{+k}\bar{p}_{+l}\right) \quad 3.20$$

Para um valor especifico de x_k, a concordância é determinada não apenas através do número de avaliadores associado com o sujeito *i* e score x_k, mas também incorporando os valores dos outros scores x_l que estão ligado a x_k através do peso w_{kl}.

No caso do Kappa de Brennan-Prediger, o valor de P_e é dado pela expressão:

$$P_e = \frac{1}{q^2}\sum_{k=1}^{q}\sum_{l=1}^{q} w_{kl} \quad 3.21$$

O seguinte exemplo ilustra o cálculo destes kappas. Note-se, como foi referido no capítulo anterior, que estes coeficientes de fiabilidade podem ser utilizados quando existem valores em falta (*missing values*). Para os Kappa de Cohen e de Scott é necessário fazer algumas modificações como iremos ver na próxima secção.

Exemplo 3.2. Num estudo, quatro avaliadores podem pontuar sujeitos usando 5 scores definidos da seguinte forma: 0.5, 1, 1.5, 2 e 2.5. Como os dados estão em intervalos de valores a estratégia mais correta é utilizar kappa ponderados. Os dados estão na seguinte tabela e a ponderação será a quadrática.

Tabela 3.5. Distribuição das classificações dos 4 avaliadores por individuo

	Avaliadores			
Pacientes	L	K	W	B
1	1	1.5	1	
2	2	2	2	2
3	0.5	1	1.5	1.5
4	1	1	1	1
5	1	1	1	1.5
6		1	2.5	
7	2.5	2.5	2.5	2.5
8	1	1		1
9		1	2	1
10	1	1	0.5	1
11	1.5	1.5	1.5	1.5
12	1	1.5	1	
13	1	1	1.5	
14	1	2	2.5	2
15		1	1.5	1
16	0.5	0.5	0.5	0.5
p_k	0.19	0.35	0.23	0.10

Tabela 3.6. Ponderação quadrática para quatro avaliadores.

	Categorias				
Categorias	**0.5**	**1**	**1.5**	**2**	**2.5**
0.5	1	0.9375	0.75	0.4375	0
1	0.9375	1	0.9375	0.75	0.4375
1.5	0.75	0.9375	1	0.9375	0.75
2	0.4375	0.75	0.9375	1	0.9375
2.5	0	0.4375	0.75	0.9375	1

O valor obtido para a percentagem de concordância é de Pa=0.9206, e para os valores dos kappa obtém-se:

- Kappa de Conger: K_{CW}=(0.9206-0.8314)/(1-0.8314)=0.5290
- Kappa de Fleiss: K_{FW}=(0.9206-0.8377)/(1-0.8377)=0.5107
- Kappa de B-P: K_{BPW}=(0.9206-0.75)/(1-0.75)=0.6823

Como acontece na prática, o coeficiente B-P irá apresentar resultados mais elevados que os coeficientes de Conger e Fleiss. Segundo o autor (Gwet, 2010), este facto deve-se a que estes últimos geralmente "exageram" na proporção esperada de acordo devido ao acaso.

3.4. Cálculo da fiabilidade com valores em falta para dois avaliadores

Até este momento, só lidamos com situações de concordância sem dados em falta *(missing data*), ou seja, exemplos em que os avaliadores classificam todos os indivíduos. Porém, na prática podem existir situações em que os avaliadores não tenham a oportunidade de pontuar uma parte dos indivíduos que participam no estudo.

De modo a lidar com estes valores em falta, devemos organizar os dados de classificação numa tabela de contingência, onde cada avaliador classifica os indivíduos nas várias categorias (na tabela 3.7 designadas por 1 e 2), e para todos os indivíduos que não são classificados por ambos os avaliadores cria-se uma categoria fictícia, denominada por X, como se mostra na tabela 3.7.

Na tabela n_{iX} representa o número de indivíduos que o avaliador A classifica na categoria *i* e que o avaliador B não pontua e n_{Xj} representa o número de indivíduos que o avaliador B classifica na categoria j e que o avaliador A não pontua. Obrigatoriamente a célula (X,X) deve ter o valor zero, o que significa que nem o avaliador A nem o avaliador B pontuam, sendo deste modo eliminados da análise.(Gwet, 2010).

Tabela 3.7. Distribuição de *n* indivíduos, por avaliador e com uma categoria com valores em falta.

		Avaliador B			
		1	2	X	Total
Avaliador A	1	n_{11}	n_{12}	n_{1X}	n_{1+}
	2	n_{21}	n_{22}	n_{2X}	n_{0+}
	X	n_{X1}	n_{X2}	*0*	n_{X+}
	Total	n_{+1}	n_{+2}	n_{+X}	*n*

As únicas alterações necessárias para a determinação dos coeficientes Kappa estão relacionadas com o cálculo das probabilidades associadas ás células da tabela 3.7 e respetivas frequências marginais. Assim que a tabela de contingência tiver toda escrita em termos probabilísticos, as expressões que foram apresentadas no capítulo 2 e neste capítulo são de aplicação direta.

As frequências associadas a cada célula terão de ser corrigidas pelos valores em falta, ou seja, ao número total de sujeitos terão de ser retirados, os sujeitos não avaliados pelo avaliador A e pelo avaliador B:

$$p_{ij} = \frac{n_{ij}}{n - (n_{+X} + n_{X+})} \qquad 3.22$$

As frequências marginais continuam a ser calculadas através dos totais coluna e dos totais linha, apenas para os sujeitos avaliados por ambos os avaliadores, dividindo depois pela dimensão da amostra, que inclui os sujeitos que têm valores em falta:

$$P_e = \sum_{k=1}^{q} p_{k+}\, p_{+k} = \sum_{k=1}^{q} \frac{n_{k+}}{n} * \frac{n_{+k}}{n} \qquad 3.23$$

onde q representa o número de categorias que os sujeitos foram avaliados por ambos os avaliadores e n representa a dimensão da amostra, incluindo os sujeitos que foram avaliados por apenas um dos avaliadores. Note-se que se houver sujeitos que não foram avaliados nem pelo avaliador A e nem pelo avaliador B são excluídos desta análise.

Exemplo 3.3. Voltando ao exemplo 2.1, vamos considerar o caso de alguns valores em falta (tabela 3.8).

Tabela 3.8. Resultados de dois avaliadores sobre a utilidade de um instrumento

		Avaliador B			
		Sim	Não	X	Total
Avaliador A	Sim	30	15	5	50
	Não	5	32	5	42
	X	3	5	0	8
Total		38	52	10	100

O primeiro passo seria transformar a tabela dos valores absolutos em frequências relativas, seguindo as equações 3.22 e 3.23:

Tabela 3.9. Frequências relativas dos resultados de dois avaliadores sobre a utilidade de um instrumento

		Avaliador B			
		Sim	Não	X	Total
Avaliador A	Sim	30/82	15/82	5/82	50/100
	Não	5/82	32/82	5/82	42/100
	X	3/82	5/82	0	8/100
Total		38/100	52/100	10/100	100

Atendendo à expressão 2.2, o valor do P_a é dado por:

$$P_a = \sum_{i=1}^{2} p_{ii} = \frac{30}{82} + \frac{32}{82} = 0.76$$

O valor de P_e dado pela expressão 2.4, não considerando as frequências marginais dos valores em falta, será dado por:

$$P_e = p_{1+}p_{+1} + p_{2+}p_{+2} = \frac{50}{100} \times \frac{38}{100} + \frac{42}{100} \times \frac{52}{100} = 0.41$$

Deste modo, o coeficiente Kappa de Cohen é dado por:

$$\hat{k}_{Cohen} = \frac{0.756 - 0.408}{1 - 0.408} \approx 0.59$$

Se não incorporássemos a correção devido aos valores em falta, o valor do P_a seria mais pequeno (0.62=62/100) em vez de 0.756. Esta seria a consequência de se tomar X como uma categoria e a classificação dos 18 indivíduos classificados por um único avaliador ser considerado como um desacordo.

No caso das frequências marginais, por exemplo, para p_{+2} o valor seria aproximadamente 0.51 ((15+32)/(50+42)), em oposição a 0.52(52/100). No entanto nós sabemos que quanto maior for o número de indivíduos, a frequência marginal é mais precisa, do que quando temos poucos indivíduos. No entanto neste caso proposto trata-se apenas de uma situação simples de fiabilidade com apenas duas categorias.

O exemplo anterior foi tratado com apenas duas categorias, podendo este ser estendido a mais que duas categorias, ou com a utilização de pesos para as discordâncias, em que a última coluna e a última linha representarão os valores em falta.

Tabela 3.10. Distribuição dos n indivíduos, por avaliador e categoria com valores em falta.

		Avaliador B					
		1	2	…	q	X	Total
Avaliador A	1	n_{11}	n_{12}	…	n_{1q}	n_{1X}	n_{1+}
	2	n_{21}	n_{22}	…	n_{2q}	n_{2X}	n_{2+}
	…					…	
	q	n_{q1}	n_{q2}	…	n_{qq}	n_{qX}	n_{q+}
	X	n_{X1}	n_{X2}	…	n_{Xq}	0	n_{X+}
	Total	n_{+1}	n_{+2}	…	n_{+q}	n_{+X}	n

No exemplo 3.4 iremos considerar os dados referentes á avaliação de um determinado síndrome em função das dores de coluna, com 18 indivíduos que foram avaliados unicamente por um dos avaliadores.

Exemplo 3.4. Avaliações efetuadas pelos 2 médicos no diagnóstico de um determinado síndrome, onde existem indivíduos que não são avaliados pelo avaliador A ou pelo avaliador B.

Tabela 3.11. Avaliações efetuadas pelos 2 médicos no diagnóstico de um determinado síndrome, onde existem indivíduos que não são avaliados pelo avaliador A ou pelo avaliador B

Médico. A	Médico B				
	Síndrome degenerativo	Síndrome disfuncional	Síndrome Postural	X	Total
Síndrome degenerativo	31	1	2	3	37
Síndrome disfuncional	3	37	4	2	46
Síndrome Postural	2	1	21	3	27
X	3	1	6	0	10
Total	39	40	33	8	120

Desta forma, utilizando as expressões 3.22 e 3.23, a percentagem de concordância e a percentagem de concordância devida ao acaso são dadas por:

$$P_a = \frac{31 + 37 + 21}{120 - (8 + 10)} = \frac{89}{102} \approx 0.87$$

$$P_e = \frac{37}{120} * \frac{39}{120} + \frac{46}{120} * \frac{40}{120} + \frac{27}{120} * \frac{33}{120} \approx 0.29$$

Consequentemente o Kappa de Cohen é:

$$\hat{k}_{Cohen} = \frac{0.87 - 0.29}{1 - 0.29} \approx 0.82$$

Se tivéssemos a utilizar pesos quadráticos, os valores seriam:

$$P_a = \frac{31 * 1 + 37 * 1 + 21 * 1 + 1 * 0.75 + 3 * 0.75 + 4 * 0.75 + 1 * 0.75}{120 - (8 + 10)}$$

$$= \frac{95.75}{102} \approx 0.94$$

$$P_e = \frac{37}{120} * \frac{39}{120} * 1 + \frac{37}{120} * \frac{40}{120} * 0.75 +$$

$$\frac{46}{120} * \frac{40}{120} * 1 + \frac{46}{120} * \frac{39}{120} * 0.75 + \frac{46}{120} * \frac{33}{120} * 0.75 +$$

$$\frac{27}{120} * \frac{33}{120} * 1 + \frac{27}{120} * \frac{40}{120} * 0.75 \approx 0.60$$

$$\hat{k}_{Cohen} = \frac{0.94 - 0.60}{1 - 0.60} = 0.85$$

Como se pode verificar, as estimativas obtida para o Kappa de Cohen não ponderado e ponderado são muito próximas, no entanto o ponderado fornece uma melhor estimativa, uma vez que que estes dão menores pesos às categorias mais afastadas.

Referências

Cicchetti, D., & Allison, T. (1971). A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal EEG Technology*, *11*, 101–109.

Cicchetti, D. V. (1981). Testing the Normal Approximation and Minimal Sample Size Requirements of Weighted Kappa When the Number of Categories is Large. *Applied Psychological Measurement*, *5*, 101–104. http://doi.org/10.1177/014662168100500114

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 551–558. http://doi.org/10.1016/0895-4356(90)90159-M

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. http://doi.org/10.1037/h0026256

Fleiss, J., Levin, B., & Cho Paik, M. (2003). *Statistical Methods for Rates and Proportions*. *John Wiley & Sons*. http://doi.org/10.1198/tech.2004.s812

Gwet, K. L. (2010). *Handbook of Inter-Rater Reliability: the definitive guide to measuring the extent of agreement among raters*. *Gaithersburg, MD: STATAXIS Publishing Company*. Advanced Analytics, LLC.

Soeken, K., & Prescott, P. (1986). Issues in the Use of Kappa to Estimate Reliability on JSTOR. Retrieved from

http://www.jstor.org/stable/3765100?seq=1#page_scan_tab_contents

Streiner, D. L. (1995). Learning how to differ: Agreement and reliability statistics in psychiatry. *Canadian Journal of Psychiatry*, *40*(2), 60–66.

Capitulo 4: Métodos paramétricos de estimação da fiabilidade para variáveis quantitativas: estudos de fiabilidade inter-avaliador e intra-avaliador baseados no ICC

Para medir a relação entre duas variáveis que representam diferentes classes de medição (variáveis que são medidas em escalas diferentes) deve-se usar um coeficiente de correlação, sendo o mais comum o coeficiente de correlação de Pearson. Este coeficiente é utilizado para relacionar/correlacionar medidas quantitativas que devem possuir uma relação linear e que devem ter uma distribuição Gaussiana. Por exemplo, medir a correlação entre os valores do quociente de inteligência e as classificações obtidas na disciplina de Matemática. No entanto, a utilização deste coeficiente para medir fiabilidade deve ser realizado com muito cuidado. É possível ter um valor de correlação elevado sem no entanto ter um valor de fiabilidade elevado (McGraw & Wong, 1996). Só no caso, da existência de um modelo linear da forma: y=x+b é que o valor de correlação será idêntico ao valor da fiabilidade.

Quando estamos interessados no relacionamento entre variáveis da mesma classe, ou seja, variáveis avaliadas com a mesma escala, o coeficiente mais utilizado é o coeficiente de correlação intraclasse (ICC, *intraclass correlation coefficient*). Uma das vantagens deste coeficiente é permitir medir a homogeneidade das avaliações, não apenas para pares de medições mas, para um grande número de medições (McGraw & Wong, 1996).

Neste capítulo iremos descrever com algum detalhe os diferentes tipos de ICC'S para medir a fiabilidade quando estamos na presença de variáveis quantitativas e em função do número de avaliadores. Será também apresentado para cada modelo o seu respetivo erro de medição (SEM-*standard error mean*), que tem grande utilidade na interpretação dos

resultados, dado que indica a precisão do instrumento utilizado (Weir, 2005), sendo considerado uma medida de concordância definida pelas linhas orientadoras apresentadas por (Kottner et al., 2011) .

.

4.1 Definição do coeficiente de correlação intraclasse

O ICC surge como uma tentativa de superar algumas limitações da correlação clássica (não detetam quaisquer erros sistemáticos), sendo uma das ferramentas estatísticas mais utilizadas para determinar a fiabilidade de medidas, especialmente quando o número de avaliadores é elevado e as variáveis são quantitativas.

Grande parte das medições nas ciências comportamentais (e não só) envolvem erros de medição associados aos avaliadores (seres humanos), podendo estes erros afetar seriamente a análise estatística e a sua interpretação. Desta forma torna-se muito importante quantificar este erro através de um índice de fiabilidade (Shrout & Fleiss, 1979a).

No cálculo de qualquer modelo do ICC serão necessárias múltiplas avaliações sobre o mesmo conjunto de sujeitos. Uma vez que os sujeitos são selecionados aleatoriamente, estes irão representar um fator aleatório do ponto de vista do desenho do estudo. Uma forma útil de apresentar as classificações é através da tabela 4.1, em que i (i=1,…,n) é o índice utilizado para os sujeitos ou objetos a serem medidos, j (j=1,…,k) é o índice que se refere aos avaliadores ou múltiplas observações (McGraw & Wong, 1996).

O ICC é apresentado na literatura e é calculado como a proporção da variabilidade atribuída aos objetos em estudo (razão entre esta variabilidade e a variabilidade total, incluindo erro de medição associado). A expressão 4.1 apresenta a definição geral para o ICC (Shrout & Fleiss, 1979a) (McGraw & Wong, 1996):

$$\mathrm{ICC} = \frac{\text{Variabilidade entre objetos de estudo}}{\text{Variabilidade entre objetos de estudo} + \text{var. do erro de medição}} \qquad 4.1$$

Tabela 4.1. Estrutura dos dados usados no cálculo do ICC para uma situação de inter-avalidores.(McGraw & Wong, 1996)

Objeto de medição	**Avaliadores**					
	1	2	…	*j*	…	*k*
1	X_{11}	X_{12}	…	X_{1j}	…	X_{1k}
2	X_{21}	X_{22}	…	X_{2j}	…	X_{2k}
...	…	…		…		…
i	X_{i1}	X_{i2}	…	X_{ij}	…	X_{ik}
…	…	…		…		…
n	X_{n1}	X_{n2}	…	X_{nj}	…	X_{nk}

A fiabilidade inter-avaliadores (designado por ρ) tem por base a definição do coeficiente de correlação entre duas pontuações quantitativas X_{ij} e $X_{ij'}$ associadas com dois avaliadores j e j', sobre o mesmo sujeito i:

$$\rho = corr\left(X_{ij}, X_{ij'}\right) = Cov(X_{ij}, X_{ij'}) / \left(\sqrt{Var(X_{ij})} \times \sqrt{Var\left(X_{ij'}\right)}\right) \quad 4.2$$

A expressão 4.2 pode ser generalizada para k avaliadores. Em geral, para qualquer combinação linear, as seguintes propriedades da variância e da covariância são úteis para determinar os modelos do ICC correspondentes:

$$var\left(\sum_{i=1}^{N} a_i X_i\right) = \sum_{i=1}^{N} a_i^2 var(X_i) + \sum_{i \neq j} a_i a_j cov\left(X_i, X_j\right) \quad 4.3$$

e:

$$var(aX + b) = \ a^2 var(X) \quad 4.4$$

$$var(aX + \ CY) = var(aX) + var\ (cY) + 2 * cov(aX, CY) \quad 4.5$$

$$cov(a, X) = 0 \quad 4.6$$

$$cov(X, X) = var(X) \quad 4.7$$

$$cov\ (aX + b, cY + d) = cov(aX, cY) + cov(aX, d) + cov(b, cY) + cov(b, d) \quad 4.8$$
$$= a * c\ cov(X, Y)$$

O ICC é baseado na análise dos modelos de análise variância (ANOVA) de medidas repetidas e é calculado a partir de estimativas das diferentes componentes de variância, através da decomposição da variância total nas variâncias entre sujeitos (*between subjects*) e dentro dos sujeitos (within subjects).

O primeiro passo para a utilização de qualquer um dos procedimentos que iremos descrever consiste em escolher o modelo de variância mais adequado aos dados da amostra. Os autores Shrout e Fleiss (Shrout & Fleiss, 1979a) apresentam 3 modelos principais para o cálculo do ICC. Os modelos são:

1. Um fator de efeitos aleatórios (*model 1: one-way random effects*). Neste modelo apenas os sujeitos participantes do estudo serão considerados como fator aleatório (dado que são retirados de uma amostra aleatória). Situações como a utilização de diferentes avaliadores ou a utilização de várias medidas pelo mesmo avaliador (medidas repetidas ou réplicas) não são consideradas neste modelo. Este modelo será apresentado na seção 4.2.

2. Dois fatores de efeitos aleatórios (*model 2*: *two-way random effects*). Neste modelo os dois fatores são os sujeitos (fator 1) e os avaliadores (fator 2), não havendo réplicas nas avaliações feitas por estes. Os avaliadores são retirados de uma amostra aleatória de avaliadores, sendo considerado um fator aleatório. Este modelo será apresentado na seção 4.3.

3. Dois fatores de efeitos mistos (*model 3*: *one-way mixed effects*). Este modelo é similar ao anterior, mas os avaliadores não são considerados aleatórios por terem sido escolhidos pelo investigador, sendo este fator considerado fixo. Novamente, a situação de haver réplicas pelo mesmo avaliador não é contemplada neste modelo. Este modelo será apresentado na seção 4.4.

Os autores McGraw e Wong (McGraw & Wong, 1996) desenvolveram o trabalho anterior, apresentando cinco modelos para o cálculo do ICC. No entanto, os dois modelos “extra” são meramente casos particulares dos modelos gerais apresentados. Como indicado, existem vários modelos para o cálculo do ICC, cada um deles sendo adequado para uma situação específica,

e podendo dar resultados muito diferentes quando aplicados ao mesmo conjunto de dados.

Para os modelos 2 e 3 faremos a distinção formal entre concordância absoluta *(agreement or absolute agreement)* e consistência (consistency), uma vez que em certos estudos será de grande importância considerar a variabilidade entre os avaliadores no cálculo do ICC. Em concordância absoluta a variabilidade entre os avaliadores é incluída enquanto a opção consistência, esta variabilidade é considerada irrelevante.

Em relação à unidade em análise, esta poderá ser em forma individual (quando o erro de medição não é corrigido pelo número de avaliadores) ou em média (quando o erro de medição é dividido pelo número de avaliadores). Os valores de fiabilidade apresentados num formato de média, tem tendência a serem superiores, dado que o valor do erro de medição é menor nestas situações.

Em resumo, as diretrizes para a escolha apropriada do tipo de ICC, pedem á partida três decisões (Shrout & Fleiss, 1979a):

- Para a análise do estudo de fiabilidade, os dados serão tratados por um modelo da ANOVA de medidas repetidas de um fator ou de dois fatores;
- Se as diferenças entre as classificações dos avaliadores são ou não importantes para o estudo de fiabilidade (concordância vs consistência);
- Se a unidade em análise representa uma classificação individual ou uma média de várias classificações (individual vs média).

Ao longo deste capítulo, iremos apresentar os modelos estatísticos para determinar a fiabilidade para os modelos 1, 2 e 3 apresentados por Shrout e Fleiss (Shrout & Fleiss, 1979a) e McGraw e Wong (McGraw & Wong, 1996).

4.2 Modelo de um fator (*one way factor*)

Neste tipo de modelo, os sujeitos são considerados como a única fonte de variabilidade, não sendo incorporado no modelo qualquer outro tipo de

informação. Deste modo, apenas se pode estudar o efeito do fator “sujeito”. Por este motivo este modelo é conhecido como um modelo de um fator. No entanto, para este modelo, não faz qualquer sentido estudar o efeito das “colunas”. As colunas podem representar diferentes avaliadores, não sendo conhecido quais os sujeitos que foram avaliados por um determinado avaliador, ou mesmo se um avaliador avaliou todos os sujeitos participantes no estudo. No entanto, as colunas também podem representar diferentes medições (ou réplicas) realizadas pelo mesmo avaliador. Ou seja, este modelo 1 pode representar uma situação inter-avaliador (vários avaliadores e uma única medição) ou uma situação intra-avaliador (um só avaliador e várias medições), dependo da informação que esteja nas colunas.

Estes modelos são designados por ICC(1,1) no caso da unidade em análise ser uma classificação individual, ou ICC(1,k) se unidade em análise for a classificação média dos avaliadores. Na tabela 4.1, a variável linha (que representa os sujeitos) é assumida como aleatória e a forma como são recolhidos na sua ordenação j (nas coluna) é irrelevante, ou seja, cada sujeito poderá ser avaliado por um grupo diferente de avaliadores. Embora os k avaliadores estejam rotulados como “avaliadores” (j=1,2,…,k), eles poderão de facto, representar diferentes avaliadores ou medidas repetidas de um só avaliador, como foi referido.

Seja x_{ij} a representação da pontuação quantitativa atribuída ao sujeito i (i=1,…,n) e pelo avaliador j (j=1,…,k), onde os dados estão representados pela tabela 4.1. O modelo matemático para este caso é dado:

$$x_{ij} = \mu + s_i + w_{ij} \qquad 4.9$$

onde μ representa a média global de todas as pontuações, ou seja, o efeito global comum nas varias pontuações, s_i o efeito do individuo i e w_{ij} o erro aleatório. As suposições do modelo definido na expressão 4.9 são: $s_i \underset{\sim}{iid} N(0, \sigma_s^2)$, $w_{ij} \underset{\sim}{iid} N(0, \sigma_w^2)$, ou seja, o efeito do sujeito e do erro associado tem uma distribuição gaussiana de média 0 e de variância constante,

independente e identicamente distribuída. Atendendo à expressão 4.2, o valor do ICC(1,1) vem:

$$\rho = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_w^2} \quad 4.10$$

dado que:

$$Cov(x_{ij}.x_{ij'}) = Cov(\mu + s_i + w_{ij}, \mu + s_i + w_{ij'}) = \sigma_s^2 \quad 4.11$$

$$Var(x_{ij}) = Var(\mu + s_i + w_{ij}) = \sigma_s^2 + \sigma_w^2 \quad 4.12$$

$$Var(x_{ij'}) = Var(\mu + s_i + w_{ij'}) = \sigma_s^2 + \sigma_w^2 \quad 4.13$$

Partindo da definição apresentada 4.1, σ_w^2 irá representar a variância do erro de medição associado ao instrumento de medição e σ_s^2 irá representar a variabilidade entre os sujeitos. Para a estimação das variâncias apresentadas na expressão 4.10, é necessário obter os resultados da ANOVA que estão apresentados na tabela 4.2, onde MS_S é a média dos quadrados dos sujeitos e MS_W, a média dos quadrados da variância dos erros.

Tabela 4.2. Quadrados médios esperados para a análise da variância no modelo1.

Modelo e fonte da variância	df	MS	EMS
Entre linhas	n-1	MS_S	$k\sigma_s^2 + \sigma_w^2$
Dentro das linhas	n(k-1)	MS_W	σ_w^2

MS: média dos quadrados ; EMS: média esperada dos quadrados

Com base na tabela 4.2, a estimativa ICC(1,1) e do erro de medição será dado por (McGraw & Wong, 1996):

$$\widehat{ICC}(1,1) = \frac{MS_S - MS_W}{MS_S + (k-1)MS_W} \quad 4.14$$

$$SEM = \sqrt{\sigma_w^2} = \sqrt{MS_W} \quad 4.15$$

Usando a tabela 4.2, substituindo $MS_w = \hat{\sigma}_w^2$ e $MS_s = \mathrm{k}\hat{\sigma}_\mathrm{r}^2 + \hat{\sigma}_\mathrm{w}^2$ na expressão 4.14, obtêm-se o modelo para o ICC apresentado na expressão 4.10.

As avaliações feitas por observações individuais são muitas vezes consideradas pouco fiáveis, este problema pode ser resolvido através do cálculo da média das k medições realizadas. Desta forma, o ICC(1,k) é definido por:

$$\rho_k = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_w^2/k} \quad 4.16$$

A estimativa do ICC(1,k) é dada por (McGraw & Wong, 1996):

$$\widehat{ICC}(1,k) = \frac{MS_s - MS_w}{MS_R} \quad 4.17$$

De forma análoga, utilizando os estimadores apresentados na tabela 4.2 na expressão 4.17, obtêm-se o modelo para o ICC apresentado na expressão 4.16.

Exemplo 4.1.Para ilustrar o cálculo do ICC sob o modelo inter-avaliador de um fator, vamos apresentar a situação proposta por Shrout e Fleiss (1979) com 4 avaliadores e 6 sujeitos.

Tabela 4.3 Pontuações atribuídas a 6 sujeitos por 4 avaliadores.

	Avaliadores			
Sujeitos	1	2	3	4
1	9	2	5	8
2	6	1	3	2
3	8	4	6	8
4	7	1	2	6
5	10	5	6	9
6	6	2	4	7

Recorrendo ao software R, o valor do ICC(1,1)=0.166 e do ICC(1,4)=0.443. Neste modelo, os sujeitos são a única fonte sistemática de variância, o valor baixo do ICC(1,1) significa que não é fácil fazer a discriminação dos sujeitos, no entanto, se considerarmos a média das pontuações este valor já aumenta consideravelmente, dado que neste ultimo o erro de medição é corrigido pelo número de avaliadores.

4.3 Modelo de dois fatores de efeitos aleatórios (*Two- way random effects model*)

Para o caso em que as *k* observações por individuo classificado diferem de alguma forma sistemática, o modelo de dois fatores deve ser utilizado para representar os dados. A razão pela escolha é que existe uma fonte sistemática da variância associada com as colunas (avaliadores) e com as linhas (sujeitos) da tabela 4.1. Por exemplo, se as colunas representarem os itens de um teste de matemática, estes itens podem diferir em grau de dificuldade, criando desta forma uma fonte separável da variância. O mesmo poderia acontecer se as colunas representassem diferentes avaliadores, que podem diferir na sua pontuação. Estas situações são específicas de um modelo de dois fatores.

Neste tipo de modelo, para além dos sujeitos serem considerados como uma fonte de variabilidade, os avaliadores também serão considerados como uma segunda fonte de variabilidade. No entanto, neste modelo, não há possibilidade do mesmo avaliador realizar várias medições, e por isso este modelo só avalia uma situação inter-avaliador (vários avaliadores e uma única medição). Como referido anteriormente, se os avaliadores forem escolhidos de uma forma aleatória então o modelo estatístico é designado por modelo de dois fatores de efeitos aleatórios, enquanto que, se os avaliadores forem definidos á priori pelos investigadores, o modelo estatístico será designado por modelo de dois fatores de efeitos mistos (próxima secção).

No desenho fatorial aleatório os avaliadores participantes na experiência são selecionados aleatoriamente de uma população maior de avaliadores, e os sujeitos são também selecionados aleatoriamente de um universo maior de sujeitos. Os resultados obtidos poderão então ser generalizados para a população. No desenho fatorial misto, apenas os sujeitos são selecionados aleatoriamente de uma população maior de sujeitos, enquanto os avaliadores que participam no estudo são os únicos avaliadores de interesse. Os resultados obtidos só fazem sentido para esses avaliadores, não se podendo desta forma generalizar para outros possíveis avaliadores.

Por exemplo, vamos imaginar um estudo de fiabilidade, onde a finalidade consiste em avaliar o nível de coerência entre dois dispositivos de medição utilizados em exames clínicos de reumatologia. Nesta situação o avaliador está interessado em investigar estes dois dispositivos específicos e não pretende efetuar uma generalização a outros dispositivos similares, nestas situações estamos perante um desenho fatorial misto que geralmente produz valores mais altos para o ICC, do que aqueles que se baseiam num desenho fatorial aleatório, na medida em que nenhuma variação é gerada pelo efeito do avaliador. (Gwet, 2010)

Os modelos 2 e 3 diferem do modelo 1 na medida em que as componentes de W_{ij} são especificadas. Seja X_{ij}, a representação da pontuação quantitativa atribuída ao sujeito i (i=1,…,n) pelo avaliador j (j=1,…,k) onde os k

avaliadores avaliam todos os n sujeitos. A equação para este modelo no caso de dois fatores com interação é dada por:

$$x_{ij} = \mu + s_i + r_j + (sr)_{ij} + e_{ij} \qquad 4.18$$

onde, μ representa a média global de todas as pontuações para todos os sujeitos e avaliadores, ou seja o efeito global comum nas varias pontuações, sendo desta forma constante, s_i efeito no sujeito i (efeito nas linhas) aleatórios e independentes e r_j,o efeito das colunas, $(sr)_{ij}$ efeito da interação, sujeito i x avaliador j, e e_{ij}, os erros aleatórios.

As suposições do modelo definido em 4.18 são: $s_i \overset{iid}{\sim} N(0, \sigma_s^2)$, $r_j \overset{iid}{\sim} N(0, \sigma_r^2)$, $(sr)_{ij} \overset{iid}{\sim} N(0, \sigma_{sr}^2)$, $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$. Além disso, os fatores sujeito e avaliador e a interação entre os efeitos são considerados mutuamente independentes (a magnitude de um deles não afeta a magnitude do outro efeito). A expressão 4.18 estipula que os diferentes efeitos são aditivos, independentes e seguem uma distribuição normal. Um caso particular é a situação onde existe ausência de interação entre os sujeitos e os avaliadores. Assim, a expressão 4.18 pode ser então simplificada:

$$x_{ij} = \mu + s_i + r_j + e_{ij} \qquad 4.19$$

Para os modelos de dois fatores (aleatórios ou mistos), podemos ainda considerar dois tipos diferentes: Consistência (Consistency) e concordância absoluta (absolute agreement).

Para as medidas do ICC enquanto medida de consistência, a variância das colunas (ou seja dos avaliadores) é excluída, atendendo que esta é considerada uma fonte de variância irrelevante, enquanto que em concordância absoluta, esta fonte de variabilidade é considerada relevante. Os modelos para o ICC em concordância absoluta e em consistência são dados, respetivamente, por:

$$\rho = \frac{\sigma_s^2}{\sigma_s^2 + (\sigma_r^2 + \sigma_{sr}^2 + \sigma_e^2)} \qquad para\ o\ ICC\ em\ concordância\ absoluta \tag{4.20}$$

$$\rho = \frac{\sigma_s^2}{\sigma_s^2 + (\sigma_{sr}^2 + \sigma_e^2)} \qquad para\ o\ ICC\ em\ consistência \tag{4.21}$$

No cálculo do ICC, o denominador representa a variância total das pontuações, enquanto o numerador representa a variabilidade associada aos sujeitos. Como se pode observar nas expressões anteriores, a diferença entre concordância absoluta e consistência está relacionada com a incorporação ou não da componente de variância σ^2_r no denominador do ICC. Assim, iremos denotar os modelos de dois fatores, como, $ICC_A(2,1)$, ICC_A (2,k), $ICC_C(2,1)$ e $ICC_C(2,k)$.

Utilizando a definição de ICC apresentado na expressão 4.2 e para a situação de concordância absoluta (para a consistência os resultados são análogos), temos:

$$\begin{aligned} Cov(x_{ij}.x_{ij'}) &= Cov(\mu + s_i + r_j + (sr)_{ij} + e_{ij}, \mu + s_i + r_{j'} + (sr)_{ij'} + e_{ij'}) \\ &= Cov(s_i, s_i) + Cov(r_j, r_{j'}) + Cov((sr)_{ij}, (sr)_{ij'}) + Cov(\mathrm{e}_{ij}, \mathrm{e}_{ij'}) \end{aligned} \tag{4.22}$$

considerando que a covariância dos produtos cruzados são nulos. Utilizando as propriedades indicadas no início do capítulo, a expressão 4.22 fica:

$$Cov(x_{ij}.x_{ij'}) = \sigma_s^2 \tag{4.23}$$

dado que os avaliadores, a sua respetiva interação com os sujeitos e os erros aleatórios são independentes e identicamente distribuídos, a sua covariância é nula. A variância será dada por:

$$Var(x_{ij}) = Var(\mu + s_i + r_j + (sr)_{ij} + e_{ij}) = \sigma_s^2 + \sigma_r^2 + \sigma_{sr}^2 + \sigma_e^2 \tag{4.24}$$

De forma análoga se obtém a $Var(x_{ij'})$. Os quadrados médios esperados apropriados para este tipo de estudo aparecem na tabela 4.4.

Tabela 4.4. Quadrados médios esperados para a análise da variância para o modelo 2.

Fonte de variância	df	MS	EMS
Entre linhas	n-1	MS_S	$k\sigma_s^2 + \sigma_{sr}^2 + \sigma_e^2$
Dentro das linhas	n(k-1)	MS_W	$\sigma_r^2 + \sigma_{sr}^2 + \sigma_e^2$
Entre colunas	k-1	MS_R	$n\sigma_r^2 + \sigma_{sr}^2 + \sigma_e^2$
Erro	(n-1)(k-1)	MS_E	$\sigma_{sr}^2 + \sigma_e^2$

Com base na tabela anterior, as estimativas de $ICC_A(2,1)$ e $ICC_C(2,1)$ e os respetivos erros de medição são definidos pelas expressões 4.25 a 4.28 (McGraw & Wong, 1996), respectivamente:

$$\widehat{ICC}_A(2,1) = \frac{MS_S - MS_E}{MS_S + (k-1)MS_E + \frac{k}{n}(MS_R - MS_E)} \quad 4.25$$

$$SEM = \sqrt{\sigma_r^2 + \sigma_{sr}^2 + \sigma_e^2} = \sqrt{MS_w} \quad 4.26$$

$$\widehat{ICC}_C(2,1) = \frac{MS_S - MS_E}{MS_S + (k-1)MS_E} \quad 4.27$$

$$SEM = \sqrt{\sigma_{sr}^2 + \sigma_e^2} = \sqrt{MS_E} \quad 4.28$$

Com base na tabela 4.4, substituindo os valores MS_S, MS_E e MS_R pelos respetivos valores esperados nas expressões 4.25 e 4.27, obtêm-se as expressões 4.20 e 4.21.

Voltando ao exemplo 4.1, iremos agora considerar que os sujeitos e os avaliadores são escolhidos de uma população maior (ou seja, são ambos fatores aleatórios), recorrendo ao software R (package IRR), obtivemos os valores $ICC_A(2,1)=0.29$ e o $ICC_C(2,1)=0.72$. Como se pode constatar, os valores encontrados são muito diferentes, o que significa que a variabilidade das pontuações dos avaliadores é bastante elevada.

Tal como foi referido para o modelo 1, por vezes as avaliações individuais são pouco fiáveis, atendendo a que existem imensas possibilidades de erro. A forma de ultrapassar esta problemática consiste em considerar a média das classificações dadas pelos avaliadores. Para a concordância absoluta e para a consistência, as expressões 4.20 e 4.21 vão ser corrigidas pelo número de avaliadores k, diminuído desta forma as componentes de variabilidade que lhes estão associadas:

$$\rho_k = \frac{\sigma_s^2}{\sigma_s^2 + (\sigma_r^2 + \sigma_{sr}^2 + \sigma_e^2)/k} \quad para\ o\ ICC\ em\ concordância\ absoluta \tag{4.29}$$

$$\rho_k = \frac{\sigma_s^2}{\sigma_s^2 + (\sigma_{sr}^2 + \sigma_e^2)/k} \quad para\ o\ ICC\ em\ consistencia \tag{4.30}$$

Os quadrados médios esperados estão representados na tabela 4.4, desta forma as estimativas dos $ICC_A(2,k)$ e $ICC_C(2,k)$ são dados por (McGraw & Wong, 1996):

$$\widehat{ICC}_A(2,k) = \frac{MS_S - MS_E}{MS_S + \frac{MS_R - MS_E}{n}} \tag{4.31}$$

$$\widehat{ICC}_C(2,k) = \frac{MS_S - MS_E}{MS_E} \qquad 4.32$$

Baseando-nos no exemplo 4.1, iremos determinar os ICC's atrás referidos, recorrendo ao software R. Deste modo, considerando a média das classificações dos quatro avaliadores obtém-se $ICC_A(2,4)=0.62$ e $ICC_C(2,4)=0.909$, que são bastante mais elevados que os respetivos valores obtidos anteriormente considerando a unidade de medida individual.

4.4 Modelo de dois fatores de efeitos mistos

Como referido anteriormente, neste modelo o fator avaliador é fixo, ou seja os avaliadores participantes no estudo são os únicos avaliadores de interesse. Em certos estudos o efeito avaliador não pode ser considerado aleatório, por exemplo se realizarmos uma experiência com um único instrumento de avaliação para classificar os mesmos sujeitos ou objetos em 10 situações diferentes, o efeito avaliador deve ser considerado fixo.

Este tipo de estudo combina o efeito do avaliador fixo com o efeito do sujeito ser aleatório, que levará a um desenho experimental denominado por desenho fatorial misto.

As expressões e as premissas são as mesmas do modelo 2 de efeitos aleatórios (expressões 4.20 e 4.21, respetivamente), no entanto, as componentes do avaliador e da sua respetiva interação são fixas, de modo que é necessário adicionar duas restrições ao modelo:

$$\sum_{j=1}^{k} r_j = 0 \qquad 4.33$$

$$\sum_{j=1}^{k}(sr)_{ij} = 0 \qquad 4.34$$

ou seja, a soma dos efeitos dos avaliadores e da respetiva interação com os sujeitos tem de ser obrigatoriamente nulo.

No modelo de efeitos aleatórios, os avaliadores representam uma variável aleatória de média 0 e variância σ_r^2. Neste modelo, como os avaliadores são fixos (expressão 4.33), então uma estimativa não enviesada para a variância dos avaliadores é dado por:

$$\theta_r^2 = \frac{\sum_{j=1}^{k} r_j^2}{k-1} \qquad 4.35$$

A segunda restrição (expressão 4.34) implica que, para o mesmo sujeito, o pressuposto que as observações são independentes não é verificado (como acontece no modelo de efeitos aleatórios) mas sim através de uma correlação negativa (Shrout & Fleiss, 1979a). Calculando a variância da expressão 4.34:

$$var\left(\sum_{j=1}^{k}(sr)_{ij}\right) = 0 \Leftrightarrow \mathrm{k}\, var\big((sr)_{ij}\big) + \sum_{i\neq \mathrm{j}}^{k} cov\big((sr)_{ij},(sr)_{ij'}\big) = 0 \Leftrightarrow \qquad 4.36$$

$$k\sigma_{sr}^2 + \sum_{i\neq \mathrm{j}}^{k} cov\big((sr)_{ij},(sr)_{ij'}\big) = 0 \Leftrightarrow$$

$$\sum_{i\neq \mathrm{j}}^{k} cov\big((sr)_{ij},(sr)_{ij'}\big) = -\frac{\sigma_{sr}^2}{k} = c$$

ou seja, c representa a covariância comum entre os efeitos da interação no mesmo sujeito.

As suposições deste modelo são diferentes do modelo anterior devido ás restrições impostas. Enquanto os sujeitos e os resíduos são definidos da

mesma forma: $s_i \underset{\sim}{iid} N(0, \sigma_s^2)$, $e_{ij} \underset{\sim}{iid} N(0, \sigma_e^2)$, os avaliadores e interação avaliadores*sujeitos são definidas como: $r_j \underset{\sim}{iid} N\left(0, \frac{k-1}{k}\sigma_r^2\right)$, $(sr)_{ij} \underset{\sim}{iid} N\left(0, \frac{k-1}{k}\sigma_{sr}^2\right)$. O fator $\frac{k-1}{k}$ está relacionado com a relação utilizada para o cálculo da estimativa amostral versus estimativa da população[1]. Esta relação terá um impacto nos modelos apresentados por Shrout e Fleiss (Shrout & Fleiss, 1979a) e McGraw e Wong (McGraw & Wong, 1996), como iremos constatar.

Utilizando novamente a definição de ICC apresentado na expressão 4.2 e para a situação de concordância absoluta (para a consistência os resultados são análogos), temos:

$$Cov(x_{ij}.x_{ij'}) = Cov(s_i, s_i) + Cov\left(r_j, r_{j'}\right) + Cov\left((sr)_{ij}, (sr)_{i'j}\right) + Cov(\mathrm{e}_{ij}, \mathrm{e}_{ij'}) \qquad 4.37$$
$$= \sigma_s^2 - \frac{\sigma_{sr}^2}{k}$$

Dado que os avaliadores, a sua respetiva interação com os sujeitos e os erros aleatórios são independentes e identicamente distribuídos, logo a sua covariância é nula. A variância vai ser dada por:

$$Var(x_{ij}) = Var(\mu + s_i + r_j + (sr)_{ij} + e_{ij}) = \sigma_s^2 + \frac{k-1}{k}\sigma_r^2 + \frac{k-1}{k}\sigma_{sr}^2 + \sigma_e^2 \qquad 4.38$$

De forma análoga se obtém a $Var(x_{ij'})$. Os valores do ICC para a concordância absoluta e para a consistência são então dados por Eliasziw et. al (1994) (Eliasziw, Young, Woodbury, & Fryday-Field, 1994):

$$\rho = \frac{\sigma_s^2 - \sigma_{sr}^2/k}{\sigma_s^2 + (\frac{k-1}{k}\sigma_r^2 + \frac{k-1}{k}\sigma_{sr}^2 + \sigma_e^2)} \qquad 4.39$$

[1] O estimador da população seria dado por $\hat{\sigma}^2 = \frac{\sum_{j=1}^{k} r_j^2}{k}$, enquanto o estimador da amostra seria dado por $\hat{s}^2 = \frac{\sum_{j=1}^{k} r_j^2}{k-1}$ e portanto $\hat{\sigma}^2 = \frac{k-1}{k}\hat{s}^2$

$$\rho = \frac{\sigma_s^2 - \sigma_{sr}^2/k}{\sigma_s^2 + \left(\frac{k-1}{k}\sigma_{sr}^2 + \sigma_e^2\right)} \quad 4.40$$

No trabalho desenvolvido por Shrout e Fleiss (Shrout & Fleiss, 1979a), este autores aplicaram um fator de correção (f) às componentes relacionadas com as variâncias dos avaliadores e respetiva interação entre avaliadores e sujeitos, com o objetivo de obterem linearidade na soma das variâncias no denominador dos ICCs:

$$f = \frac{k}{(k-1)} \quad 4.41$$

e após incorporação desse fator de correção nas expressões 4.39 e 4.40, obtém-se os resultados apresentados nos trabalhos por Shrout e Fleiss (Shrout & Fleiss, 1979a) e McGraw e Wong (McGraw & Wong, 1996):

$$\rho = \frac{\sigma_s^2 - \sigma_{sr}^2/(k-1)}{\sigma_s^2 + (\theta_r^2 + \sigma_{sr}^2 + \sigma_e^2)} \quad 4.42$$

$$\rho = \frac{\sigma_s^2 - \sigma_{sr}^2/(k-1)}{\sigma_s^2 + (\sigma_{sr}^2 + \sigma_e^2)} \quad 4.43$$

onde θ_r^2 é dado pela expressão 4.35. A tabela 4.5 apresenta os resultados da ANOVA para este estudo. Com base nesta tabela, as estimativas de $ICC_A(3,1)$ e $ICC_C(3,1)$ são definidos pelas expressões 4.44 e 4.45 (McGraw & Wong, 1996), idênticas às do modelo de efeitos aleatórios:

$$\widehat{ICC_A}(3,1) = \frac{MS_S - MS_E}{MS_S + (k-1)MS_E + \frac{k}{n}(MS_R - MS_E)} \quad 4.44$$

$$\widehat{ICC}_C(3,1) = \frac{MS_S - MS_E}{MS_S + (k-1)MS_E} \quad 4.45$$

Tabela 4.5. Quadrados médios esperados para a análise da variância no modelo 3 apresentado por Shrout e Fleiss (Shrout & Fleiss, 1979a) com a incorporação do fator de correção f=k/(k-1).

Fonte de variância	**df**	**MS**	**EMS**
Entre linhas	n-1	MS_S	$k\sigma_s^2 + \sigma_e^2$
Dentro das linhas	n(k-1)	MS_W	$\theta_r^2 + \frac{k}{k-1}\sigma_{sr}^2 + \sigma_e^2$
Entre colunas	k-1	MS_R	$n\theta_r^2 + \frac{k}{k-1}\sigma_{sr}^2 + \sigma_e^2$
Erro	(n-1)(k-1)	MS_E	$\frac{k}{k-1}\sigma_{sr}^2 + \sigma_e^2$

As fórmulas para os erros de medição não podem ser obtidas partindo da dos resultados da ANOVA proposta por Shrout e Fleiss (1979) (Shrout & Fleiss, 1979a), mas com a notação proposta por Eliasziw et. al (1994), estas relações serão, respetivamente:

$$SEM = \sqrt{\theta_r^2 + \sigma_{sr}^2 + \sigma_e^2} \quad para\ concordância\ absoluta \quad 4.46$$

$$SEM = \sqrt{\sigma_{sr}^2 + \sigma_e^2} \quad para\ consistência \quad 4.47$$

Com base na tabela 4.5, e de forma análoga à secção anterior, substituindo os valores MS_S, MS_E e MS_R pelos respetivos valores esperados nas expressões 4.44 e 4.45, obtém-se as expressões 4.42 e 4.43.

No caso da unidade de medida ser a média, os valores do ICC são dados por:

$$\rho_k = \frac{\sigma_s^2 - \sigma_{sr}^2/(k-1)}{\sigma_s^2 + (\theta_r^2 + \sigma_{sr}^2 + \sigma_e^2)/k} \qquad para\ o\ ICC\ em\ concordância\ absoluta \tag{4.48}$$

$$\rho_k = \frac{\sigma_s^2 - \sigma_{sr}^2/(k-1)}{\sigma_s^2 + (\sigma_{sr}^2 + \sigma_e^2)/k} \qquad para\ o\ ICC\ em\ consistencia \tag{4.49}$$

Como base na ANOVA (tabela 4.5) não se consegue obter as respetivas estimativas, como é indicado por (McGraw & Wong, 1996) e (Shrout & Fleiss, 1979a). No entanto, se a interação estiver ausente ($\sigma_{sr}^2 = 0$), as suas estimativas são definidas pelas expressões 4.50 e 4.51:

$$\widehat{ICC}_A(3,k) = \frac{MS_S - MS_E}{MS_S + \frac{MS_R - MS_E}{n}} \tag{4.50}$$

$$\widehat{ICC}_C(3,k) = \frac{MS_S - MS_E}{MS_E} \tag{4.51}$$

Como indicado pelos autores (Shrout & Fleiss, 1979a) e (McGraw & Wong, 1996), as expressões para as estimativas dos modelos do ICC de dois fatores aleatórios ou de dois fatores mistos baseados na ANOVA são idênticas, mas os seus respetivos modelos teóricos são diferentes, bem como as tabelas das ANOVAS associadas.

Baseando-nos no exemplo 4.1, iremos determinar os ICC's atrás referidos, recorrendo ao software R. Os valore obtidos foram: $ICC_A(3,1)$=0.29 e $ICC_C(3,1)$= 0.72. Os resultados obtidos são os mesmos que no caso dos efeitos serem aleatórios. A distinção entre eles está na forma como os resultados são interpretados e não no cálculo do ICC. Embora os modelos sejam diferentes,

as suas estimativas são iguais. Quando os avaliadores são selecionados aleatoriamente (Modelo 2), os resultados podem ser generalizados e quando são fixos (Modelo 3) não.

Resumindo, na prática não existe uma subdivisão dos dois fatores em efeitos aleatórios ou efeitos mistos, para além da sua interpretação. Existe sim, uma subdivisão entre consistência e concordância absoluta, como é indicado por Shrout & Fleiss (1979a) e (McGraw & Wong, 1996).

Referências:

Eliasziw, M., Young, S. L., Woodbury, M. G., & Fryday-Field, K. (1994). Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. *Physical Therapy*, *74*(8), 777–88. http://doi.org/10.1186/1471-2474-7-60

Gwet, K. L. (2010). *Handbook of Inter-Rater Reliability: the definitive guide to measuring the extent of agreement among raters*. *Gaithersburg, MD: STATAXIS Publishing Company*. Advanced Analytics, LLC.

Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., … Streiner, D. L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *International Journal of Nursing Studies*, *48*(6), 661–671. http://doi.org/10.1016/j.ijnurstu.2011.01.016

McGraw, K. O., & Wong, S. P. (1996). “Forming inferences about some intraclass correlations coefficients”: Correction. *Psychological Methods*, *1*(4), 390–390. http://doi.org/10.1037/1082-989X.1.4.390

Shrout, P. E., & Fleiss, J. L. (1979a). Intraclass Correlation: Uses in Assessing Rater Reliabitlity. *Psychological Bulletin*, *86*(2), 420–428.

Shrout, P. E., & Fleiss, J. L. (1979b). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428.

Weir, J. P. (2005). the I Ntraclass C Orrelation C Oefficient and the Sem.

Journal of Strength and Conditioning Research, *19*(1), 231–240. http://doi.org/10.1519/15184.1

Capitulo 5: Métodos paramétricos de estimação da fiabilidade para variáveis quantitativas baseados no ICC com múltiplas medições

Em muitos estudos, nomeadamente na área da saúde, é vulgar realizar mais de uma medição por mais do que um avaliador, no entanto os métodos para o calculo do ICC apresentados por Shrout, Fleiss (1979) , McGraw e Wong (1996), presumem que cada avaliador efetue apenas uma medição. Isto significa que os métodos por eles aplicados não podem ser estendidos a estudos de fiabilidade, quer intra ou inter-avaliador em que os avaliadores fazem mais que uma medição.

Neste capítulo iremos apresentar os métodos de fiabilidade intra e inter-avaliador para situações com mais do que uma medição por avaliador e com vários avaliadores em simultâneo.

5.1 Problemas com múltiplas observações por avaliador

Nos estudos em que são realizadas repetições, os investigadores utilizam os métodos para o cálculo do ICC como se apenas de uma medição se tratasse, fazendo por vezes a média das repetições de cada avaliador. Este procedimento tem o efeito de inflacionar a fiabilidade inter-avaliador. Sabe-se que um ICC calculado através da média de várias medições será mais elevado do que um com base numa única medição (Hayen, Dennis, & Finch, 2007).

Em alguns estudos os investigadores têm utilizado apenas as medidas repetidas para um único avaliador, mas este método é também ineficiente, uma vez que não utiliza toda a informação disponível.

Os autores Eliasziw e Young (Eliasziw, Young, Woodbury, & Fryday-Field, 1994) recomendam que, quando se avalia a fiabilidade intra-avaliador e inter-avaliador, deve-se utilizar um desenho de medidas repetidas, de forma a tirar partido do aumento de precisão obtida ao usar todas as observações.

Na tabela 5.1 é apresentado o desenho de medidas repetidas do qual os coeficientes de ICC de intra-avaliador e inter-avaliador podem ser obtidos. As m medições repetidas (l=1,..,m) são feitas por cada um dos avaliadores (j=1,…,k) de uma amostra aleatória de n sujeitos (i=1,…,n).

Tabela 5.1. Tabela de dados para um estudo de medidas repetidas. (Eliasziw et al., 1994)

	Avaliadores						
	1			**…**	***k***		
	Medidas				**Medidas**		
Sujeitos	**1**	**…**	***m***	**…**	**1**	**…**	***m***
1	x_{111}	…	x_{11m}	…	x_{1k1}	…	x_{1km}
2	X_{211}	…	X_{21m}	…	X_{2k1}	…	X_{2km}
…	…	…	…	…	…	…	…
n	X_{n11}	…	X_{n1m}	…	X_{nk1}	…	X_{nkm}

Partindo da tabela 5.1, se k=1 obtém-se o modelo de um fator para a fiabilidade intra-avaliador, apresentado na secção 4.2. Se m=1 então obtém-se os modelos de dois fatores para a fiabilidade inter-avaliador apresentados nas secções 4.3 e 4.4, respetivamente. Como referido, os modelos apresentados neste capítulo serão relativos a uma situação de múltiplos avaliadores e múltiplas avaliações realizadas por esses avaliadores.

5.2 Modelo de dois fatores de efeitos aleatórios

O modelo geral para medidas repetidas, sem dados em falta e com interação é dada por:

$$x_{ijl} = \mu + s_i + r_j + (sr)_{ij} + e_{ijl} \qquad 5.1$$

onde, μ representa a média global de todas as pontuações para todos os sujeitos e avaliadores, ou seja o efeito global comum nas varias pontuações, sendo desta forma constante, s_i efeito do sujeito *i* (efeito nas linhas) aleatórios e independentes e r_j, o efeito dos avaliadores, $(sr)_{ij}$ efeito da interação sujeito* avaliador e e_{ijl} representa os erros aleatórios .

As suposições do modelo definido em 5.1 são: $s_i \underset{\sim}{iid} N(0, \sigma_s^2)$, $r_j \underset{\sim}{iid} N(0, \sigma_r^2)$. $(sr)_{ij} \underset{\sim}{iid} N(0, \sigma_{sr}^2)$, $e_{ijl} \underset{\sim}{iid} N(0, \sigma_e^2)$. Além disso, os fatores: sujeito, avaliadores e a sua interação são considerados mutuamente independentes (a magnitude de um deles não afeta a magnitude do outro efeito).

Da mesma forma que no capítulo anterior, os modelos de dois fatores (aleatórios ou mistos), podem-se ainda dividir em dois tipos diferentes: consistência (*consistency*) e concordância absoluta (*agreement*). A diferença, como já referido, está na incorporação ou não da variabilidade relativa aos avaliadores.

Os modelos para o ICC inter-avaliador em concordância absoluta e em consistência são dados, respetivamente, por:

$$\rho = Cov(X_{ijl}, X_{ij'l}) \Big/ \left(\sqrt{Var\left(X_{ij'l}\right)} \times \sqrt{Var\left(X_{ij'l}\right)} \right) = \qquad 5.2$$

$$= \frac{\sigma_s^2}{\sigma_s^2 + (\sigma_r^2 + \sigma_{sr}^2 + \sigma_e^2)} \qquad para\ o\ ICC\ em\ concordância\ absoluta$$

$$\rho = \frac{\sigma_s^2}{\sigma_s^2 + (\sigma_{sr}^2 + \sigma_e^2)} \qquad para\ o\ ICC\ em\ consistência \qquad 5.3$$

A sua derivação foi apresentada no capítulo anterior (secção 4.3).

Para os modelos do ICC intra-avaliador, utilizando a definição de ICC apresentado pela expressão 4.2, temos que a covariância para a situação de concordância absoluta vai ser dada:

$$Cov(x_{ijl}.x_{ijl'}) = Cov(\mu + s_i + r_j + (sr)_{ij} + e_{ijl}, \mu + s_i + r_j + (sr)_{ij} + e_{ijl'}) \quad 5.4$$
$$= Cov(s_i, s_i) + Cov(r_j, r_j) + Cov((sr)_{ij}, (sr)_{ij}) + Cov\left(\mathrm{e}_{ijl}, \mathrm{e}_{ijl'}\right)$$
$$= \sigma_s^2 + \sigma_r^2 + \sigma_{sr}^2$$

e a respetiva variância é dada por:

$$Var(x_{ijl}) = Var(\mu + s_i + r_j + (sr)_{ij} + e_{ijl}) = \sigma_s^2 + \sigma_r^2 + \sigma_{sr}^2 + \sigma_e^2 \quad 5.5$$

para a situação de consistência os resultados serão análogos

Desta forma as fórmulas do ICC em concordância absoluta e em consistência para situação intra-avaliador são dadas, respetivamente, por:

$$\rho = \frac{\sigma_s^2 + \sigma_r^2 + \sigma_{sr}^2}{\sigma_s^2 + \sigma_r^2 + \sigma_{sr}^2 + (\sigma_e^2)} \quad em\ concordância\ absoluta \quad 5.6$$

$$\rho = \frac{\sigma_s^2 + \sigma_{sr}^2}{\sigma_s^2 + \sigma_{sr}^2 + (\sigma_e^2)} \quad em\ consistência \quad 5.7$$

Neste modelo de medidas repetidas é possível calcular o ICC intra-avaliador, associado a cada avaliador. Em vez de utilizar o valor total dos erros associados (σ_e^2), utiliza-se os erros associados a cada avaliador j (σ_{ej}^2), respectivamente, e as formulas são dadas por 5.8 e 5.9.

$$\rho_j = \frac{\sigma_s^2 + \sigma_r^2 + \sigma_{sr}^2}{\sigma_s^2 + (\sigma_r^2 + \sigma_{sr}^2 + \sigma_{ej}^2)} \quad em\ concordância\ absoluta \quad 5.8$$

$$\rho_j = \frac{\sigma_s^2 + \sigma_{sr}^2}{\sigma_s^2 + (\sigma_{sr}^2 + \sigma_{ej}^2)} \qquad em\ consistência \tag{5.9}$$

Os quadrados médios esperados apropriados para este tipo de estudo surgem na tabela 5.2. Estas tabelas são diferentes das apresentadas no capítulo 4 porque incorporam a componente das medidas repetidas produzidas pelos avaliadores

Tabela 5.2. Quadrados médios esperados para a análise da variância no caso dos efeitos no avaliador serem aleatórios num desenho de medidas repetidas

Fonte de variância	**df**	**MS**	**EMS**
Sujeitos	n-1	MS_S	$\mathrm{mk}\sigma_\mathrm{s}^2 + \mathrm{m}\sigma_\mathrm{sr}^2 + \sigma_\mathrm{e}^2$
Avaliadores	k-1	MS_R	$\mathrm{mn}\sigma_\mathrm{r}^2 + \mathrm{m}\sigma_\mathrm{sr}^2 + \sigma_\mathrm{e}^2$
Erro (interação)	(n-1)(k-1)	MS_E	$\mathrm{m}\sigma_\mathrm{sr}^2 + \sigma_\mathrm{e}^2$
Erro (Avaliador)	nk(m-1)	MS_{RE}	σ_e^2
Erro(Avaliador 1)	n(m-1)	MS_{R1E}	$\sigma_{\mathrm{e}1}^2$
…	…	…	…
Erro(Avaliador j)	n(m-1)	MS_{RjE}	σ_ej^2
…	…	…	…
Erro(Avaliador k)	n(m-1)	MS_{RkE}	σ_ek^2

Da tabela 5.2, rapidamente se encontra as estimativas das diferentes componentes da variância através de um sistema de equações (Eliasziw et al., 1994):

$$\begin{cases} \sigma_s^2 = \dfrac{MS_S - MS_E}{mk} \\ \sigma_r^2 = \dfrac{MS_R - MS_E}{mn} \\ \sigma_{sr}^2 = \dfrac{MS_E - MS_{RE}}{m} \\ \sigma_e^2 = MS_{RE} \end{cases} \quad 5.10$$

Com base no sistema apresentado na expressão 5.10, as estimativas dos ICCs inter-avaliadores e o seu erro de medição associado serão dados pelas expressões:

$$\widehat{ICC}_A(2,1) = \frac{\dfrac{MS_S - MS_E}{mk}}{\dfrac{MS_S - MS_E}{mk} + \dfrac{MS_R - MS_E}{mn} + \dfrac{MS_E - MS_{RE}}{m} + MS_{RE}} \quad 5.11$$

$$SEM = \sqrt{\sigma_r^2 + \sigma_{sr}^2 + \sigma_e^2} = \sqrt{\frac{MS_R - MS_E}{mn} + \frac{MS_E - MS_{RE}}{m} + MS_{RE}} \quad 5.12$$

$$\widehat{ICC}_C(2,1) = \frac{\dfrac{MS_S - MS_E}{mk}}{\dfrac{MS_S - MS_E}{mk} + \dfrac{MS_E - MS_{RE}}{m} + MS_{RE}} \quad 5.13$$

$$SEM = \sqrt{\sigma_{sr}^2 + \sigma_e^2} = \sqrt{\frac{MS_E - MS_{RE}}{m} + MS_{RE}} \quad 5.14$$

Para a situação intra-avaliador, os valores do ICC são definidos respetivamente:

$$\widehat{ICC}_A(2,1,m) = \frac{\dfrac{MS_S - MS_E}{mk} + \dfrac{MS_R - MS_E}{mn} + \dfrac{MS_E - MS_{RE}}{m}}{\dfrac{MS_S - MS_E}{mk} + \dfrac{MS_R - MS_E}{mn} + \dfrac{MS_E - MS_{RE}}{m} + MS_{RE}} \quad 5.15$$

$$\widehat{ICC}_C(2,1,m) = \frac{\frac{MS_S - MS_E}{mk} + \frac{MS_E - MS_{RE}}{m}}{\frac{MS_S - MS_E}{mk} + \frac{MS_E - MS_{RE}}{m} + MS_{RE}} \quad 5.16$$

$$SEM = \sqrt{\sigma_e^2} = \sqrt{MS_{RE}} \quad 5.17$$

De forma analogia se poderá obter os ICCs para a situação intra-avaliador associados para cada um dos avaliadores (j=1,..,k):

$$\widehat{ICC}_A(2,1,m,av_j) = \frac{\frac{MS_S - MS_E}{mk} + \frac{MS_R - MS_E}{mn} + \frac{MS_E - MS_{RjE}}{m}}{\frac{MS_S - MS_E}{mk} + \frac{MS_R - MS_E}{mn} + \frac{MS_E - MS_{RjE}}{m} + MS_{RjE}} \quad 5.18$$

$$SEM = \sqrt{\sigma_{ej}^2} = \sqrt{MS_{RjE}} \quad 5.19$$

5.3. Modelo de dois fatores de efeitos mistos

No caso do modelo de dois fatores de efeitos mistos, as mesmas restrições aplicadas a uma situação inter-avaliador (expressões 4.33 e 4.34), também são aplicadas numa situação intra-avaliador.

As equações e as premissas são as mesmas do modelo de efeitos aleatórios, no entanto, as componentes do avaliador e da sua respetiva interação são fixas.

Os valores do ICC na situação inter-avaliador para a concordância absoluta e para a consistência foram apresentados no capitulo anterior e são dados por Eliasziw et. al (1994) (Eliasziw, Young, Woodbury, & Fryday-Field, 1994):

$$\rho = \frac{\sigma_s^2 - \sigma_{sr}^2/k}{\sigma_s^2 + (\theta_r^2 + \frac{k-1}{k}\sigma_{sr}^2 + \sigma_e^2)} \quad para\ concordância\ absoluta \tag{5.20}$$

$$\rho = \frac{\sigma_s^2 - \sigma_{sr}^2/k}{\sigma_s^2 + \left(\frac{k-1}{k}\sigma_{sr}^2 + \sigma_e^2\right)} \quad para\ consistência \tag{5.21}$$

As suposições deste modelo são diferentes do modelo de fatores aleatórios devido às restrições impostas. Enquanto os sujeitos e os resíduos são definidos da mesma forma: $s_i \overset{iid}{\sim} N(0, \sigma_s^2)$, $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$, os avaliadores e interação avaliadores*sujeitos são definidas como: $r_j \overset{iid}{\sim} N\left(0, \frac{k-1}{k}\sigma_r^2\right)$, $(sr)_{ij} \overset{iid}{\sim} N\left(0, \frac{k-1}{k}\sigma_{sr}^2\right)$. A covariância irá ser dada:

$$\begin{aligned} Cov(x_{ijl}.x_{ijl'}) &= Cov(\mu + s_i + r_j + (sr)_{ij} + e_{ijl}, \mu + s_i + r_j + (sr)_{ij} + e_{ijl'}) \\ &= Cov(s_i, s_i) + Cov(r_j, r_j) + Cov((sr)_{ij}, (sr)_{ij}) + Cov\left(\mathrm{e}_{ijl}, \mathrm{e}_{ijl'}\right) \\ &= \sigma_s^2 + \frac{k-1}{k}\ \sigma_r^2 + \frac{k-1}{k}\sigma_{sr}^2 \end{aligned} \tag{5.22}$$

e a respetiva variância é dada por:

$$Var(x_{ijl}) = Var(\mu + s_i + r_j + (sr)_{ij} + e_{ijl}) = \sigma_s^2 + \frac{k-1}{k}\sigma_r^2 + \frac{k-1}{k}\sigma_{sr}^2 + \sigma_e^2 \tag{5.23}$$

Os valores do ICC na situação intra-avaliador para a concordância absoluta e para a consistência, são dados por (Eliasziw, Young, Woodbury, & Fryday-Field, 1994):

$$\rho = \frac{\sigma_s^2 + (k-1)(\theta_r^2 + \sigma_{sr}^2)/k}{\sigma_s^2 + ((k-1)(\theta_r^2 + \sigma_{sr}^2)/k + \sigma_e^2)} \quad para\ o\ ICC\ concordância\ absoluta \tag{5.24}$$

$$\rho = \frac{\sigma_s^2 + (k-1)\sigma_{sr}^2/k}{\sigma_s^2 + ((k-1)\sigma_{sr}^2/k + \sigma_e^2)} \qquad para\ o\ ICC\ Consistência \qquad 5.25$$

Os autores Eliasziw et al. (1994) (Eliasziw, Young, Woodbury, & Fryday-Field, 1994) apenas apresentam os quadrados médios esperados para a situação de consistência e os resultados encontram-se na tabela 5.3. Quando os avaliadores são fixos, os autores não apresentaram os quadrados médios esperados na situação de concordância absoluta.

Tabela 5.3. Quadrados médios esperados para a análise da variância no caso dos efeitos no avaliador serem fixos num desenho de medidas repetidas apenas para a situação de consistência (a pontuação entre os avaliadores é irrelevante).

Fonte de variância	*df*	*MS*	*EMS*
Sujeitos	*n-1*	MS_S	$mk\sigma_s^2 + \sigma_e^2$
Avaliadores	*k-1*	MS_R	$mn\theta_r^2 + m\sigma_{sr}^2 + \sigma_e^2$
Erro (interação)	*(n-1)(k-1)*	MS_E	$m\sigma_{sr}^2 + \sigma_e^2$
Erro (Avaliador)	*nk(m-1)*	MS_{RE}	σ_e^2
Erro (Avaliador 1)	*n(m-1)*	MS_{R1E}	σ_{e1}^2
…	…	…	
Erro (Avaliador j)	*n(m-1)*	MS_{RjE}	σ_{ej}^2
…	…	…	
Erro (Avaliador k)	*n(m-1)*	MS_{RkE}	σ_{ek}^2

As estimativas das diferentes componentes da variância para uma situação de consistência são:

$$\begin{cases} \sigma_s^2 = \dfrac{MS_S - MS_{RE}}{mk} \\ \theta_r^2 = \dfrac{MS_R - MS_E}{mn} \\ \sigma_{sr}^2 = \dfrac{MS_E - MS_{RE}}{m} \\ \sigma_e^2 = MS_{RE} \end{cases} \quad 5.26$$

Com base na tabela anterior, as estimativas para os ICCs inter-avaliador e intra-avaliador apenas para a consistência são idênticas ás apresentadas na secção anterior:

$$\rho = \frac{\dfrac{MS_S - MS_{RE}}{mk} + \dfrac{(k-1)\left(\dfrac{MS_E - MS_{RE}}{m}\right)}{k}}{\dfrac{MS_S - MS_{RE}}{mk} + \dfrac{(k-1)\left(\dfrac{MS_E - MS_{RE}}{m}\right)}{k} + MS_{RE})} \quad 5.22$$

$$= \frac{\dfrac{MS_S - MS_{RE}}{mk} + \left(\dfrac{MS_E - MS_{RE}}{m}\right) - (\dfrac{MS_E - MS_{RE}}{mk})}{\dfrac{MS_S - MS_{RE}}{mk} + \left(\dfrac{MS_E - MS_{RE}}{m}\right) - (\dfrac{MS_E - MS_{RE}}{mk}) + MS_{RE})} =$$

$$= \frac{\dfrac{MS_S - MS_E}{mk} + \dfrac{MS_E - MS_{RE}}{m}}{\dfrac{MS_S - MS_E}{mk} + \dfrac{MS_E - MS_{RE}}{m} + MS_{RE}} = \widehat{ICC}_C(2,1,m)$$

5.4. Exemplo para o cálculo da fiabilidade intra-avaliador

O exemplo prático que a seguir apresentamos foi retirado de Ellasziw e Young, e tem como objetivo ilustrar um estudo de fiabilidade baseado num desenho de medidas repetidas.

Exemplo 5.1. Num estudo de teste/re-teste, para avaliar o nível de fiabilidade foram utilizados 2 goniómetros na medição de um ângulo associado a uma articulação (em graus): um de plástico (goniómetro1 denotado por av1) e um de outro tipo (Lamoreux eletrogoniometro, goniómetro 2 denotado por av2). Uma amostra de 29 doentes foram medidos três vezes consecutivas por ambos os goniometros. Os dados referentes ao estudo encontram-se representados na

tabela 5.4 em relação a uma posição comum do joelho, a extensão passiva total. Recorrendo ao software R, obtivemos as várias estimativas do ICC, num desenho de medidas repetidas, apresentadas na tabela 5.5.

Neste exemplo os únicos avaliadores de interesse são o goniómetro 1 e o goniómetro 2, dai só analisarmos a situação de efeitos mistos. Desta forma $ICC_C(2,k)$=0.961, que nos indica que as medições efetuadas nos 29 indivíduos foram consistentes, por ambos os avaliadores (goniómetro 1 como avaliador 1) e goniómetro 2 como avaliador 2).

Para verificar a consistência e reprodutibilade das medições analisamos as estimativas dos coeficientes intra-avaliador e constatamos que a estimativa geral é boa $ICC_C(2,k,m)$=0.984, no entanto como estamos interessados em verificar a eficiência de cada avaliador (goniómetro), ou seja cada goniómetro é o único avaliador de interesse, é conveniente apresentar cada uma das estimativas em separado, $ICC_C(2,1,m,av1)$=0.986 e $ICC_C(2,1,m,av2)$=0.982, ambas as estimativas são consideradas boas, embora neste estudo o goniómetro 1 seja ligeiramente mais fiável que o goniómetro 2.

Tabela 5.5. Resultados para os vários ICC's, considerando as situações inter-avaliador e intra-avaliador no desenho de medidas repetidas.

Modelos	ICC
Inter-avaliador	
$ICC_A(2,1)$	0.945
$ICC_C(2,1)$	0.961
Intra-avaliador	
$ICC_A(2,1,m)$	0.984
$ICC_A(2,1,m,av1)$	0.986
$ICC_A(2,1,m,av2)$	0.982
$ICC_C(2,1,m)$	0.984
$ICC_C(2,1,m,av1)$	0.986
$ICC_C(2,1,m,av2)$	0.982

Tabela 5.4. Dados relativos aos 29 pacientes na avaliação do ângulo em graus da articulação do joelho na posição extensiva passiva total, avaliados por dois goniómetros.

	Goniómetro 1			Goniómetro 2		
Pacientes	R1	R2	R3	R1	R2	R3
1	-2	0	1	2	1	1
2	16	16	15	12	14	13
3	5	6	6	4	4	4
4	11	10	10	9	7	8
5	7	8	6	5	6	6
6	-7	-8	-8	-9	-10	-9
7	18	19	19	17	17	17
8	4	5	5	5	5	5
9	0	-3	-2	-7	-6	-5
10	0	0	-2	1	2	1
11	-3	-2	-2	-4	-3	-3
12	3	-1	1	-1	-2	1
13	7	9	9	4	4	2
14	-6	-7	-6	-8	-10	-9
15	1	1	0	-2	-2	-3
16	-13	-14	-14	-12	-12	-12
17	2	1	0	-1	0	0
18	4	4	3	7	6	4
19	-10	-9	-10	-10	-11	-10
20	8	9	8	2	8	8
21	7	6	7	8	7	7
22	-3	-2	-4	-5	-5	-5
23	-5	-5	-7	-6	-8	-7
24	5	5	5	3	4	4
25	0	-1	-1	-4	-3	-4
26	7	6	6	4	4	4
27	-8	-8	-8	-10	-11	-10
28	1	1	2	1	-1	0
29	-3	-3	-3	-5	-4	-5

Os valores obtidos anteriormente são bastante idênticos, atendendo a que existe pouca variabilidade, quer nos avaliadores(goniómetro 1 e goniómetro 2) quer nas 3 repetições efetuadas.

Referências

Eliasziw, M., Young, S. L., Woodbury, M. G., & Fryday-Field, K. (1994). Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. *Physical Therapy*, *74*(8), 777–88. http://doi.org/10.1186/1471-2474-7-60

Gwet, K. L. (2010). *Handbook of Inter-Rater Reliability: the definitive guide to measuring the extent of agreement among raters*. *Gaithersburg, MD: STATAXIS Publishing Company*. Advanced Analytics, LLC.

Hayen, A., Dennis, R. J., & Finch, C. F. (2007). Determining the intra- and inter-observer reliability of screening tools used in sports injury research. *Journal of Science and Medicine in Sport*, *10*(4), 201–210. http://doi.org/10.1016/j.jsams.2006.09.002

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass Correlation: Uses in Assessing Rater Reliabitlity. *Psychological Bulletin*, *86*(2), 420–428.

Capítulo 6: Métodos não paramétricos para a estimação da concordância em variáveis quantitativas ou ordinais com várias categorias

Neste capítulo o enfâse será dado para os casos onde a variável medida é ordinal ou quantitativa, mas que não segue uma distribuição Normal ou que a sua distribuição não seja conhecida. Nestes casos utilizam-se técnicas não paramétricas que ordenam por ordem crescente os valores quantitativos, atribuindo-lhes posições (rankings). Todos os testes estatísticos apresentados neste capítulo seguem este princípio.

Os testes não paramétricos são considerados testes de distribuição livre, independentes da forma da distribuição da população de onde a amostra foi retirada. Se a dimensão da amostra é reduzida (por exemplo, n=7), não há alternativa senão usar os testes não paramétricos. Os pressupostos associados aos testes não paramétricos são reduzidos quando comparados com os dos testes paramétricos. Para alguns testes não paramétricos, o único pressuposto que se assume é que a distribuição subjacente aos dados seja contínua, característica partilhada com todos os testes paramétricos.

Estes testes são particularmente úteis quando os dados a analisar já são ordinais, ou seja, o resultado é apresentado por rankings (por exemplo, desde o melhor relatório até ao pior relatório de um conjunto de alunos) mas igualmente úteis para transformar dados quantitativos em rankings após ordenação prévia. Neste caso, os resultados dos testes não paramétricos tornam-se pouco sensíveis a valores outliers severos que possam existir na amostra, o que não acontece nos testes paramétricos. Estatísticas de ordem como as medianas e os percentis são utilizados em vez da média e o desvio-padrão (por exemplo).

No entanto, os testes não paramétricos também têm desvantagens. Se os pressupostos para o teste paramétrico tiverem sido validados, então este é preferível a um teste não paramétrico, dado que a potência de um teste paramétrico é superior à do correspondente teste não paramétrico. Os métodos não-paramétricos também não conseguem testar as interações entre fatores como a análise de variância consegue fazer. Por último, o seu desconhecimento muitas vezes resulta numa clara preferência pelos testes paramétricos (muito mais conhecidos) mesmo quando os pressupostos não são verificados. Neste caso, os resultados estatísticos e as conclusões podem estar errados.

6.1 Coeficiente de correlação de Spearman

Como (Siegel & Castellan, 1988) indicam, de todas as estatísticas baseadas em rankings, o coeficiente de correlação de Spearman (*Spearman rank correlation coefficient*) foi o primeiro a ser desenvolvido e é o mais conhecido a seguir ao coeficiente de correlação de Pearson para variáveis quantitativas com distribuição Normal bivariada.

Tal como o coeficiente de correlação de Pearson, o de Spearman (r_s) só apresenta valores no intervalo [-1; +1] e quanto mais próximos do extremo mais forte será o valor de dependência entre as duas variáveis, enquanto que mais próximo de 0, mais independentes serão as variáveis.

O coeficiente R_s pode ser calculado usando a formula do coeficiente de correlação de Pearson, substituindo os valores das observações de X_1 e X_2 pelas respetivas ordens r_1 e r_2:

$$R_S = \frac{\sum_{i=1}^{n}(r_{1i} - \bar{r}_1)(r_{2i} - \bar{r}_2)}{\sqrt{\sum_{i=1}^{n}(r_{1i} - \bar{r}_1)^2}\sqrt{\sum_{i=1}^{n}(r_{2i} - \bar{r}_2)^2}} \quad 6.1$$

Esta expressão pode simplificar-se, após manipulação algébrica, obtendo-se a expressão conhecida para este coeficiente (Zar, 2010)

$$R_S = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n^3 - n} \qquad 6.2$$

onde ${d_i}^2 = (r_{1i} - r_{2i})^2$, n representa o número de sujeitos e d_i a diferença das pontuações associadas ao sujeito *i*.

Quando existem empates nas pontuações, situação frequente, então a equação 6.2 tem de ser corrigida para:

$$R_S^* = \frac{A_1 + A_2 - \sum_{i=1}^{n} {d_i}^2}{2\sqrt{A_1 * A_2}} \qquad 6.3$$

$$A_i = \frac{(n^3 - n - T_i)}{12} \qquad 6.4$$

$$T_i = \sum_{j=1}^{s_i} {t_j}^3 - t_j \qquad 6.5$$

onde A_i representa a correção dos empates do avaliador *i* (para *i*=1,2), baseado no número de empates (s_i), onde T_i é a soma das diferenças ${t_j}^3 - t_j$ para todos os empates produzidos pelo avaliador *i*.

Exemplo 6.1. O seguinte exemplo foi retirado de Gwet (2010) sobre a avaliação de dois avaliadores sobre a capacidade pulmonar de 15 crianças. A tabela 6.1 apresenta os resultados necessários para calcular o coeficiente de correlação.

Tabela 6.1: Pontuações obtidas para a capacidade pulmonar em crianças

Sujeito	Avaliador 1	Avaliador 2	Rank 1	Rank 2	d_i	d_i^2
1	190	220	1	2	-1	1
2	220	200	3	1	2	4
3	260	260	**4.5**	4	0.5	0.25
4	210	300	2	**11.5**	-9.5	90.25
5	270	265	**6.5**	5	1.5	2.25
6	280	280	**9.5**	**7.5**	2	4
7	260	280	**4.5**	**7.5**	-3	9
8	275	275	8	6	2	4
9	280	290	**9.5**	**9.5**	0	0
10	320	290	**12.5**	**9.5**	3	9
11	300	300	11	**11.5**	-0.5	0.25
12	270	250	**6.5**	3	3.5	12.25
13	320	330	**12.5**	15	-2.5	6.25
14	335	320	14	**13.5**	0.5	0.25
15	350	320	15	**13.5**	1.5	2.25
Total						145

O valor de T_1 é dado pela série de 4 pares repetidos duas vezes (4.5 duas vezes; 6.5 duas vezes; 9.5 duas vezes; 12.5 duas vezes, indicado a negrito na tabela), enquanto o valor de T_2 também é dado pela série de 4 pares repetidos duas vezes (7.5 duas vezes; 9.5 duas vezes; 11.5 duas vezes; 13.5 duas vezes):

$$T_1 = \sum_{j=1}^{4} {t_j}^3 - t_j = 4(2^3 - 2) = 24$$

$$T_2 = \sum_{j=1}^{4} {t_j}^3 - t_j = 4(2^3 - 2) = 24$$

$$A_1 = \frac{(15^3 - 15 - 24)}{12} = 278$$

$$A_2 = \frac{(15^3 - 15 - 24)}{12} = 278$$

O valor da correlação de Spearman ajustada para os empates é dada por:

$$R_S^* = \frac{278 + 278 - 145}{2\sqrt{278 * 278}} = 0.7392$$

Ou seja, o valor da fiabilidade destes dois avaliadores é de aproximadamente 0.74.

6.2 Coeficiente de correlação de Kendall τ

O coeficiente de correlação Kendall τ , é uma medida de associação entre duas variáveis e foi proposto por Kendall em 1938. Tal como com o coeficiente de Spearman, este coeficiente é bivariado e é baseado em dados de ranking ordenados. Deste modo, as variáveis têm de estar numa escala pelo menos ordinal e estarmos na presença de dois avaliadores. (Gwet, 2010)(Siegel & Castellan, 1988)

O coeficiente de correlação de Kendall τ fornece uma medida do grau de associação ou de correlação entre os dois conjuntos de classificações, e para muitos investigadores é usado para quantificar o grau de concordância entre os rankings de 2 avaliadores.

Embora o coeficiente de correlação Kendall τ tenha surgido mais tarde do que o coeficiente de Spearman, e tenha um procedimento de calculo mais demorado (Gwet, 2010), possui algumas propriedades estatísticas interessantes, como a distribuição amostral de Kendall τ aproximar-se da distribuição normal para pequenas amostras. (Lindeman, Merenda, & Gold, 1980).

Como foi referido, o Kendall τ pode ser visto como uma medida do grau de concordância entre 2 conjuntos de rankings com respeito à ordenação relativa de todos os possíveis pares de sujeitos. Para qualquer par de sujeitos (*i,j*) podemos determinar os seus rankings (A_i, A_j) e (B_i, B_j), no que diz respeito aos 2 avaliadores A e B respetivamente. Se o sinal da diferença A_i - A_j for o mesmo que o sinal da diferença B_i - B_j então o par (*i,j*) está em concordância, de outra forma está em discordância.

Quando não há empates em 2 conjuntos de avaliações qualquer par de indivíduos serão concordantes ou discordantes. Considerando n_C o par de contagens concordantes dos indivíduos e n_D, o numero de pares discordantes e n o número de sujeitos que participam no estudo, então o Kendall τ é dado por:

$$\tau = \frac{n_C - n_D}{n(n-1)/2} \qquad 6.6$$

onde $n(n-1)/2$ representa o numero total de pares distintos de indivíduos. O seguinte exemplo mostra o cálculo do Kendall τ numa pequena amostra de 8 sujeitos.

Exemplo 6.2. Supondo que é efetuado um estudo para testar a eficiência de um novo inclinómetro digital, expresso em graus, de forma a medir a amplitude do movimento do ombro esquerdo em 8 doentes selecionados aleatoriamente, os dados foram obtidos por dois Médicos, os avaliadores A e B respetivamente, e estão representados na tabela 6.2.

Tabela 6.2 Classificação e rankings dos 8 pacientes classificados pelos médicos A e B

Indivíduos	Avaliador A	Avaliador B	Rank A	Rank B
1	79.8º	78.1º	7	5
2	65.1º	63.1º	2	1
3	78.8º	78.6º	5	6
4	65.4º	65º	4	4
5	80º	79.8º	8	8
6	65.3º	64.9º	3	3
7	64º	64.2º	1	2
8	79.3º	79	6	7

Na tabela 6.3 estão apresentados os cálculos necessários para obter o Kendall τ. A primeira linha contém os indivíduos, as linhas relativas ao ranking A e ao ranking B , representam os rankings dos avaliadores A e B respetivamente. Os números da diagonal representam os rankings do avaliador B replicados e as letras C e D indicam quando um par é concordante ou discordante respetivamente. A primeira letra da linha (i), está associada ao par de indivíduos (1,2). Este par é concordante porque os pares de ranks associadas (i.e. (7,2) do avaliador A e (5,1) do avaliador B) estão na mesma direção (i.e. 7-2=5>0 e 5-1=4>0). Na mesma linha a primeira letra D diz respeito ao par de indivíduos (1,3), pois os 2 pares de rankings variam em direções opostas (ou seja (7,5) do avaliador A e (5,6) do avaliador B), sendo estes pares discordantes. As restantes linhas da tabela 3.10 obtêm-se da mesma forma.

A ultima linha da tabela 6.3 contém o total de pares concordantes n_C e o total de pares discordantes n_D, desta forma e usando a equação 6.6, o Kendal τ é dado por:

$$\tau = \frac{25 - 3}{8 \times (8 - 1)/2} = \frac{22}{28} = 0.7857$$

Tabela 6.3. Rankings dos 8 doentes classificados pelos médicos A e B

Individuo	1	2	3	4	5	6	7	8		
Rank A	7	2	5	4	8	3	1	6		
Rank B	5	1	6	4	8	3	2	7	n_C	n_D
(i)	5	C	D	C	C	C	C	D	5	2
(ii)		1	C	C	C	C	D	C	5	1
(iii)			6	C	C	C	C	C	5	0
(iv)				4	C	C	C	C	4	0
(v)					8	C	C	C	3	0
(vi)						3	C	C	2	0
(vii)							2	C	1	0
(vii)								7		
Total									25	3

Quando os avaliadores ao produzir os resultados produzem empates, então o coeficiente Kendall τ terá de ser ajustado. A definição de pares concordantes e discordantes dos indivíduos não é compatível com a existência de empates, desta forma o Tau ajustado permite excluir os pares onde existe empate, no entanto o denominador terá de ser ajustado em conformidade, Este τ ajustado é calculado como se segue:

$$\tau^* = \frac{2(n_C - n_D)}{\sqrt{n(n-1) - T_A} \times \sqrt{n(n-1) - T_B}} \qquad 6.7$$

Exemplo 6.3. Esta situação é idêntica à do exemplo 6.2 mas com empates nas classificações. A tabela 6.4 mostra as amplitudes do movimento do ombro esquerdo de 8 pacientes medidas pelo novo inclinómetro digital.

Tabela 6.4. Classificação e rankings dos 8 pacientes classificados pelos médicos A e B.

Indivíduos	Avaliador A	Avaliador B	Rank A	Rank B
1	79.8º	78º	7	6.5
2	65º	65.2º	4.5	4
3	79.8º	79º	7	8
4	65º	63º	4.5	2
5	79.8º	78º	7	6.5
6	64º	67º	2	5
7	64.3º	65.1º	3	3
8	61º	60	1	1

Tabela 6.5 Rankings dos 8 pacientes classificados pelos médicos A e B

Individuo	1	2	3	4	5	6	7	8		
Ranking A	7	4.5	7	4.5	7	2	3	1		
Ranking B	6.5	4	8	2	6.5	5	3	1	n_C	n_D
(i)	6.5	C	0	C	0	C	C	C	5	0
(ii)		4	C	0	C	D	C	C	4	1
(iii)			8	C	0	C	C	C	4	0
(iv)				2	C	D	D	C	2	2
(v)					6.5	C	C	C	3	0
(vi)						5	D	C	1	1
(vii)							3	C	1	0
(vii)								1		
Total									20	4

Os valores de T_A e T_B, determina-se de forma idêntica á apresentada pelo método de Spearman dado pela equação 6.5.

A tabela 6.5 apresenta os passos para o calculo do Kendall τ corrigido. O procedimento é idêntico ao caso anterior, no entanto quando surgir um empate, quer para o avaliador A, quer para o avaliador B, não existe concordância nem

discordância, dai ser atribuído o valor 0, isto significa que estes pares serão excluídos da contagem, e da mesma forma obtemos o número de pares concordantes e o número de pares discordantes, representados por n_C e n_D, respetivamente.

Para determinar o Kendall τ corrigido, teremos de determinar todos os rankings do avaliador A e do avaliador B onde surgem os empates, deste modo, para o avaliador A temos 2 conjunto de empates {7,7,7} e {4.5,4.5}, o primeiro com 3 elementos e o segundo com 2 elementos. Desta forma $T_A=(3^3 - 3)+(2^3 - 2)=30$ e para o avaliador B temos 1 único conjunto de empates, com 2 elementos {6.5,6.5}, então $T_B=2^3-2=6$. Usando a equação 6.7, tem-se:

$$\tau^* = \frac{2 \times (20 - 4)}{\sqrt{8 \times (8 - 1) - 30} \times \sqrt{8 \times (8 - 1) - 6}} = 0.8875$$

Como se pode verificar o τ ajustado é mais elevado que o tau não ajustado, isto deve-se essencialmente à existência de um número significativo de empates.

6.3 Coeficiente Kendall W

O coeficiente de Kendall de concordância (KCC) é adequado para variáveis pelo menos ordinais, e avalia a extensão de concordância entre 2 ou mais avaliadores no que diz respeito à sua classificação de um mesmo grupo de indivíduos.

Este coeficiente é frequentemente denotado por W e o seu valor está compreendido entre 0 e 1, onde 0 representa ausência total de concordância e 1 representando uma concordância perfeita.(Siegel & Castellan, 1988). Valores negativos de W são impossíveis, atendendo a que não pode existir um desacordo total em mais de 2 avaliadores (Siegel & Castellan, 1988). A noção de associação negativa ou correlação negativa (ou seja, classificações com

direções opostas) não se aplica a um grupo de 3 ou mais avaliadores, mesmo que muitas vezes seja relevante no caso de dois avaliadores, o terceiro já não estará em discordância total com os outros 2.(Siegel & Castellan, 1988).

Ao longo desta seção, vamos supor que temos de analisar uma tabela de dados com as classificações organizados coluna a coluna, como mostra a Tabela 6.6. Nesta encontram-se as classificações numéricas atribuídas a 8 indivíduos por quatro avaliadores.

Tabela 6.6: Classificações atribuídas aos 8 indivíduos pelos avaliadores A, B,C e D.

Sujeitos	Avaliador A	Avaliador B	Avaliador C	Avaliador D
1	79.8º	78º	77º	75º
2	65º	65.2º	63.1º	67º
3	79.8º	79º	80º	79.1º
4	65º	63º	64º	67º
5	79.8º	78º	81º	80º
6	64º	67º	64º	65º
7	64.3º	65.1º	64º	65º
8	61º	60	63.5º	67º

Para formalizar o coeficiente de concordância de Kendall, vamos assumir que temos n indivíduos que serão avaliados por k avaliadores (na tabela 6.6, n=8 e k=4). O KCC é uma medida baseada nos rankings. Portanto em primeiro lugar temos de atribuir os rankings em ordem ascendente desde o 1 até ao número de indivíduos, exceto obviamente a 1ª coluna que diz respeito aos indivíduos.

Se existirem empates, então estes terão um ranking que resulta da média aritmética dos seus rankings. Por exemplo o avaliados A, que está representado na 2ª coluna da tabela 6.6 contem um empate {79.8;79.8;79.8}, e estas classificações são as 3 mais altas de todas as classificações atribuídas pelo avaliador A, então os seus rankings seriam 6, 7, e 8 e a sua média

aritmética (6+7+8)/3 =7, consequentemente a cada uma das classificações 79.8 será atribuído o ranking 7. A notação R_{ij} designará o ranking associado ao individuo *i* e ao avaliador *j.*

O KCC representa a proporção da variância associada com a soma dos rankings do individuo marginal R_i e o maior valor possível da variância dada pelo número de indivíduos e o número de juízes. Desta forma o KCC denotado por W é calculado usando uma das seguintes equações[1] (Gwet, 2010).

$$W = \frac{12S}{k^2 n(n^2-1) - kT} \quad 6.8$$

$$W = \frac{12S^* - 3k^2 n(n+1)^2}{k^2 n(n^2-1) - kT} \quad 6.9$$

onde *S* é a soma de todos os quadrados das diferenças entre as somas marginais dos rankings R_i e a sua média global $\bar{R}$ (i.e. $(R_i - \bar{R})^2$), e S^* é a soma de todos os quadrados (R_i^2). Finalmente, *T* é o fator de correção dos empates, e é definido por:

$$T = \sum_{l=1}^{m} (t_l^2 - t_l) \quad 6.10$$

onde *m* é o número total de empates da amostra de dados (por exemplo na tabela 6.6, existem 2 empates nas classificações dos avaliadores A e D, e 1 empate nas classificações dos avaliadores B e C , logo neste caso temos um total de 6 empates), e t_l é o número de indivíduos associados com o empate especifico (na tabela 6.6, t_l=2 ou , t_l=3).

Os resultados das classificações de cada um dos médicos, denotados por avaliadores A, B,C e D encontram-se representados na tabela 6.7. Na tabela

[1] Ambas as equações aparecem em livros didáticos e permitem encontrar o coeficiente de concordância de Kendall produzindo o mesmo resultado. No entanto algumas tabelas estatísticas, frequentemente utilizadas para avaliar a significância estatística do KCC são baseadas no valor de S, que não se encontra na equação 6.9

6.8 mostra-nos os rankings dos 8 indivíduos baseados nas classificações dadas pelos 4 avaliadores. A coluna R_i contém a soma dos rankings para cada individuo, as somas marginais dos rankings e a respetiva coluna R_i^2 contém os quadrados dessas somas e S é dado por:

$$S = \sum_{i=1}^{8} (R_i - \bar{R})^2 \quad e \quad \bar{R} = \sum_{i=1}^{8} R_i \Big/ 8 \tag{6.11}$$

Tabela 6.7.: Rakings dos 8 pacientes atribuídos pelos avaliadores A, B, C e D.

Sujeitos	Avaliador A	Avaliador B	Avaliador C	Avaliador D	R_i	R_i^2
1	7	6.5	6	6	25.5	650.25
2	4.5	4	1	4	13.5	182.25
3	7	8	7	7	29	841
4	4.5	2	4	4	22.5	210.25
5	7	6.5	8	8	29.5	870.25
6	2	5	4	1.5	12.5	156.25
7	3	3	4	1.5	11.5	132.25
8	1	1	2	4	8	64
Total	36	36	36	36	144	3106.5
S					514,5	
T	30	6	24	30	90	

Desta forma, pela equação 6.8, o coeficiente de concordância de Kendall W, é dado por:

$$W = \frac{12 \times 514.5}{4^2 \times 8 \times (8^2 - 1) - 4 \times 90} = 0.801$$

Usando a expressão 6.9 obtém-se o mesmo resultado:

$$W = \frac{12 \times 3106{,}5 - 3 \times 4^2 \times 8 \times (8+1)^2}{4^2 \times 8 \times (8^2 - 1) - 4 \times 90} = 0.801$$

Neste exemplo o coeficiente de concordância de Kendall é alto, o que significa que existe uma boa fiabilidade entre os 4 avaliadores.

O coeficiente de concordância de Kendall foi desenvolvido de forma independente por Kendall e Babington-Smith (1939) e Wallis (1939), e tem uma estreita relação com o coeficiente de correlação de Spearman, a qual é expressa pela seguinte expressão:

$$W = \bar{r} - \frac{\bar{r} - 1}{k} \qquad 6.12$$

onde $\bar{r}$ representa a média de todos os pares distintos do coeficiente de correlação de Spearman. A equação 6.12 sugere-nos que á medida que o número de avaliadores aumenta, o coeficiente de concordância de Kendell tende a ficar cada vez mais próximo da média do coeficiente de Spearman.[2]

Referências

Gwet, K. L. (2010). *Handbook of Inter-Rater Reliability: the definitive guide to measuring the extent of agreement among raters*. *Gaithersburg, MD: STATAXIS Publishing Company*. Advanced Analytics, LLC.

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). Introduction to Bivariate and Multivariate Analysis, Glenview IL: Scott, Foresman.

Siegel, S., & Castellan, J. (1988). *Nomparametric Statistics for the behavioral science*. (1988 McGraw-Hill, Ed.).

[2] O coeficiente de concordância de Kendall(W) pode ser calculado pela equação 6.12, mas só quando o numero de avaliadores é pequeno, pois para este procedimento será necessário calcular C_2^k coeficientes de correlação de Spearman e de seguida determinar a sua média.

Zar, J. H. (2010). *Biostatistical Analysis*. *Prentice Hall New Jersey USA*. http://doi.org/10.1037/0012764

Capitulo 7: Inferência estatística para os métodos de concordância e fiabilidade apresentados

Neste capítulo iremos apresentar alguns resultados de inferência estatística associada aos métodos apresentados nos capítulos anteriores. Os exemplos apresentados nos capítulos anteriores serão novamente abordados, apresentado a respetiva inferência e o código R produzido será apresentado em apêndice.

A inferência estatística considerada nesta dissertação consiste na estimação intervalar (intervalos de confiança) ou na realização de testes de hipótese para o parâmetro da população de interesse.

Um intervalo de confiança é um intervalo de valores, derivados das medições recolhidas, que se espera conter o verdadeiro valor do parâmetro da população com uma probabilidade pré-definida que se denomina por grau de confiança. Geralmente os valores standards são 90%, 95% e 99%.

Um teste de hipótese consiste em testar se uma determinada hipótese, H_0 (hipótese nula), é verdadeira ou não, quando colocada em alternativa a uma segunda hipótese, H_1 (hipótese alternativa). A realização de teste de hipóteses consiste basicamente em calcular a probabilidade de se observar um valor amostral tanto ou mais afastado do valor considerado na hipótese nula como aquele que foi observado na amostra recolhida. A esta probabilidade dá-se o nome de p-value. Se ela for reduzida, então quer dizer que a nossa amostra se afasta da situação considerada em H_0 e consequentemente devemos rejeitar essa hipótese em favor da hipótese alternativa H_1.Portanto, valores de p-values pequenos indicam que as diferenças entre o que foi observado e o que seria de esperar observar (sob H_0) são muito significativas. (Gwet, 2010)

Os resultados apresentados foram obtidos utilizando o software RStudio (https://www.rstudio.com, versão 0.99.896) e R (https://www.r-project.org, versão 3.2.3), utilizando os packages IRR e lpsolve e as rotinas implementadas por Gwet (2010). Todo o código R produzido no âmbito desta dissertação (que não foi possível encontrar nos packages consultados) encontra-se em anexo, no final da mesma.

7.1 Inferência estatística para variáveis nominais ou ordinais classificados por categorias

O intervalo de confiança (1-α)*100% para as estatísticas Kappa (ponderadas ou não ponderadas) é representado pela seguinte relação genérica (Sim & Wright, 2005):

$$\widehat{K} \pm z_{1-\frac{\alpha}{2}} * \frac{SD(\widehat{K})}{\sqrt{n}} = \widehat{K} \pm z_{1-\frac{\alpha}{2}} * SE(\widehat{K}) \quad 7.1$$

onde $\widehat{K}$ representa o estimador do parâmetro Kappa, SD e SE representam o estimador do seu desvio-padrão e do erro padrão amostral, respetivamente, $z_{1-\alpha/2}$ é o quantil de ordem 1-α/2 da distribuição Normal standard, N(0,1).

O intervalo de confiança apresentado só é válido se o valor da dimensão da amostra for elevado, sendo a distribuição Normal justificada pelo Teorema do Limite Central. No caso de dimensões reduzidas deve-se utilizar a abordagem proposta por Donner (1998). Esta abordagem é bastante complexa, saindo fora do âmbito desta dissertação.

Cohen (1960) apresentou a seguinte expressão para o desvio-padrão:

$$SD(\widehat{K}_{Cohen}) = \sqrt{\frac{P_a(1-P_a)}{(1-P_e)^2}} \quad 7.2$$

No entanto, Fleiss, Cohen, e Everitt (1969), explicam que a expressão 7.2 apresenta algumas limitações mas contudo, a sua simplicidade permite efetuar alguns cálculos, como por exemplo, o cálculo da dimensão da amostra. Sobre a hipótese nula de que a concordância entre os avaliadores é zero (H_0: K=0), a estimativa do erro padrão é dada por Fleiss, Cohen, e Everitt (Fleiss, Joseph L.; Cohen, Jacob; Everitt, 1969):

$$SE_0(\widehat{K}_{Cohen}) = \frac{1}{\sqrt{n}(1-P_e)}\sqrt{P_e + {P_e}^2 - \sum_{k=1}^{q} p_{k+}p_{+k}(p_{k+} + p_{+k})} \quad 7.3$$

onde $\widehat{K}_{Cohen}$ e P_e estão definidos pelas expressões 2.3 e 2.4, respetivamente apresentadas no capitulo 2.

Nos testes de hipóteses, considerando que a condição H_0 representa o caso de a concordância entre os avaliadores ser apenas devido ao acaso (ou seja que as suas avaliações não têm qualquer relação) (H_0:K=0; H_1:K≠0), a estatística de teste será dada por:

$$Z_{obs} = \frac{\widehat{K} - 0}{SE_0(\widehat{K})} \sim N(0,1) \quad 7.4$$

O cálculo do p-value é obtido através da sua definição, neste caso para a situação bi-lateral:

$$p - value = 2 * P\left(Z > |Z_{obs|}|H_0\right) \quad 7.5$$

No entanto se o teste de hipóteses for para um Kappa diferente do valor 0, o erro padrão do Kappa de Cohen terá de ser modificado para (Fleiss, Joseph L.; Cohen, Jacob; Everitt, 1969):

$$SE\left(\widehat{K}_{Cohen}\right) = \frac{1}{\sqrt{n}(1-P_e)^2}\sqrt{A + B - C} \quad 7.6$$

$$A = \sum_{i=1}^{q} p_{ii}\left[1 - (P_{+i} - P_{i+})\left(1 - \widehat{K}_{Cohen}\right)\right]^2 \quad 7.7$$

$$B = (1 - \widehat{K}_{Cohen})^2 \sum_{i=1}^{q} \sum_{\substack{j=1 \\ i \neq j}}^{q} p_{ij}\left(p_{+i} + p_{j+}\right)^2 \quad 7.8$$

$$C = \left(\widehat{K}_{Cohen} - P_e\left(1 - \widehat{K}_{Cohen}\right)\right)^2 \quad 7.9$$

A estatística de teste será dada por:

$$Z_{obs} = \frac{\widehat{K} - K_0}{SE(\widehat{K})} \sim N(0,1) \quad 7.10$$

As expressões para os desvios-padrões dos coeficientes Kappa apresentados nos capítulos 2 e 3 são bastante extensas, podendo encontrar-se a sua implementação em diversos softwares como o R, SAS, ou SPSS. A título de exemplo, para o Kappa de Fleiss, o seu SE é dado por (Fleiss, 1971):

$$SE\left(\widehat{K}_{Fleiss}\right) = \sqrt{\frac{2}{nr(r-1)} \frac{\sum_{k=1}^{q} p_k^2 - (2r-3)\left(\sum_{k=1}^{q} p_k^2\right)^2 + 2(r-2)\sum_{k=1}^{q} p_k^3}{\left(1 - \sum_{k=1}^{q} p_k^2\right)^2}} \quad 7.11$$

$$p_k = \frac{1}{nr} \sum_{i=1}^{n} n_{ik} \quad 7.12$$

onde n_{ik} é o números de avaliadores associado ao sujeito i da categoria k.

Nas tabelas seguintes são apresentados os resultados completos para os exemplos 2.1, 2.2 e 2.3 obtidos pelos diferentes coeficientes Kappa apresentados no capítulo 2. Nestas tabelas é também apresentado o coeficiente de concordância de AC_1 desenvolvido por Gwet (2010). Este coeficiente não foi apresentado no capítulo 2 devido á sua complexidade matemática. Para mais informação consulte (Gwet, 2010). No entanto apresentamos os seus valores que foram obtidos através das rotinas do R.

Tabela 7.1. Coeficientes de concordância estimados para o exemplo 2.1 (dois avaliadores com uma escala binária). O valor de Pa é idêntico em todos os coeficientes (Pa=0.75)

Estatísticas Kappa	Pe	Kappa	SE_0	95% IC	p-value
Cohen	0.49	0.51	0.08	[0.35;0.67]	p<0.001
Scott	0.50	0.50	0.09	[0.33;0.67]	p<0.001
Brenan-Prediger	0.50	0.50	0.09	[0.33;0.67]	p<0.001
AC_1	0.50	0.50	0.67	[0.33;0.67]	p<0.001

Tabela 7.2. Coeficientes de concordância estimados para o exemplo 2.2 (dois avaliadores com uma escala multinomial). O valor de Pa é idêntico em todos os coeficientes (Pa=0.87)

Estatísticas Kappa	Pe	Kappa	SE_0	95% IC	p-value
Cohen	0.34	0.81	0.05	[0.71;0.90]	p<0.001
Scott	0.35	0.81	0.05	[0.71;0.90]	p<0.001
Brenann-Prediger	0.33	0.81	0.05	[0.71;0.91]	p<0.001
AC_1	0.33	0.81	0.05	[0.71;0.91]	p<0.001

Como se pode constatar das tabelas 7.1 e 7.2, todos os valores para os diferentes coeficientes de Kappa são considerados moderados a quase perfeitos (ver classificação da Tabela 2.5) e significativos, e apresentam amplitudes semelhantes para o IC a 95%. Este resultado não é surpreendente tendo em consideração que estes coeficientes apresentam pequenas variações para o cálculo de Pe.

Os resultados da tabela 7.3 apresentam coeficientes classificados como moderados, sendo todos significativos, e apresentam amplitudes elevadas para o IC95%. A amplitude elevada deve-se á combinação de uma dimensão reduzida com a existência de discordâncias entre os avaliadores. Novamente a variação existente nos resultados é devido ao cálculo de Pe, sendo o maior valor obtido para o Kappa de Fleiss.

Tabela 7.3. Coeficientes de concordância estimados para o exemplo 2.3 (múltiplos avaliadores com uma escala multinomial). O valor de Pa é idêntico em todos os coeficientes (Pa=0.69).

Estatísticas Kappa	Pe	Kappa	SE_0	95% IC	p-value
Fleiss	0.24	0.60	0.13	[0.30;0.89]	p<0.001
Conger	0.23	0.60	0.13	[0.31;0.89]	p<0.001
Brenann-Prediger	0.20	0.62	0.12	[0.34;0.89]	p<0.001
AC_1	0.19	0.62	0.12	[0.35;0.89]	p<0.001

Para os Kappas ponderados iremos apresentar os resultados de inferência para os exemplos 3.1 e 3.2. Para a análise dos valores em falta numa situação de dois avaliadores, serão utilizados os exemplos 3.3 e 3.4. Para mais do que dois avaliadores e com valores em falta será utilizado o exemplo 3.2 para a ilustração dos resultados.

Tabela 7.4. Coeficientes de concordância estimados para o exemplo 3.1 (dois avaliadores com uma escala ordinal com 3 categorias) com ponderação linear e quadrática.

Estatísticas	Pa	Pe	Kappa	SE_0	95% IC	p-value
Cohen (não ponderado)	0.64	0.36	0.44	0.23	[-0.08,0.95]	0.090
Cohen (linear)	0.82	0.60	0.54	0.20	[0.10,0.99]	0.020
Cohen (quadrático)	0.91	0.73	0.67	0.16	[0.31,1.00]	0.002
Brenan-Prediger (não ponderado)	0.64	0.33	0.45	0.22	[-0.03,0.94]	0.063
Brenan-Prediger (linear)	0.82	0.56	0.59	0.16	[0.23,0.95]	0.005
Brenan-Prediger (quadrático)	0.91	0.67	0.73	0.11	[0.48,0.97]	p<0.001

Os resultados dos coeficientes não ponderados (Cohen e Brenan-Prediger) são moderados no entanto estes resultados não são significativos ao nível de 5%. Os seus respetivos intervalos de confiança a 95%, apresentam amplitudes muito elevadas. Note-se que o 0 pertence ao intervalo de confiança, o que significa que concordância concordância não é diferente da que se obtém só por acaso, para o referido grau de significância. Os resultados dos coeficientes com ponderação linear continuam moderados, mas já são significativos, no entanto apresentam amplitudes elevadas para os respetivos intervalos de confiança a 95%. Quando de consideram as ponderações quadráticas, ambos os coeficientes são bastante mais elevados, substanciais e significativos, os intervalos de confiança têm uma amplitude grande mas mais reduzida que as anteriores. A amplitude ainda elevada já foi justificada anteriormente. A variação existente nos resultados é devido ao cálculo de Pe, sendo para qualquer ponderação, o maior valor obtido para o Kappa de Cohen, daí este apresentar sempre um coeficiente mais reduzido.

Tabela 7.5. Coeficientes de concordância estimados para o exemplo 3.2 (quatro avaliadores com uma escala ordinal com 5 categorias) com ponderação linear e quadrática e com valores em falta.

Estatísticas	Pa	Pe	Kappa	SE_0	95% IC	p-value
Fleiss (não ponderado)	0.56	0.31	0.36	0.16	[0.02,0.71]	0.038
Fleiss (linear)	0.83	0.70	0.44	0.19	[0.05,0.84]	0.030
Fleiss (quadrático)	0.92	0.84	0.51	0.23	[0.03,0.99]	0.039
Brenan-Prediger (não ponderado)	0.56	0.20	0.45	0.12	[0.21,0.70]	0.001
Brenan-Prediger (linear)	0.83	0.60	0.58	0.12	[0.32,0.85]	p<0.001
Brenan-Prediger (quadrático)	0.92	0.75	0.68	0.14	[0.38,0.99]	p<0.001

Os valores obtidos para as estatísticas encontradas são significativas, e obtivemos melhores valores para o Brenan-Prediger do que para o Fleiss para cada uma das ponderações, devido á forma como é calculado P_e. Note-se que o calculo de P_e no Brenan-Prediger só leva em conta o número de categorias. As amplitudes dos intervalos de confiança continuam elevadas para todas as estatísticas obtidas.

7.2 Cálculo da dimensão da amostra com base nas estatísticas Kappa

O cálculo da dimensão da amostra é fundamental no planeamento de um estudo. Nesta secção iremos apresentar o processo do cálculo da dimensão da amostra baseado na amplitude do intervalo de confiança. Alternativamente, para o Kappa de Cohen com respostas binárias, o valor da dimensão da amostra pode ser calculado baseado no trabalho de (Cantor, 1996). Uma abordagem mais geral, pode ser encontrada em (Altaye, Donner, & Eliasziw, 2001).

Supondo que se pretende uma determina amplitude do intervalo de confiança para o parâmetro da população, a expressão 7.1 pode ser reescrita da seguinte forma:

$$2 * z_{1-\frac{\alpha}{2}} * \frac{SD(\widehat{K})}{\sqrt{n}} \leq amp \qquad 7.13$$

onde amp representa a amplitude do intervalo de confiança pretendida. No caso da proporção de concordância ser conhecida, a expressão 7.13 é reescrita como:

$$n \geq \left(\frac{2 * z_{1-\frac{\alpha}{2}} * SD(\widehat{K})}{amp} \right)^2 \qquad 7.14$$

No caso da proporção de concordância ser desconhecida, P_a assume o valor 0.5 ou o valor da sua estimativa. Para o cálculo da dimensão da amostra irá ser utilizada a expressão 7.2 por questão de simplicidade de cálculo. Nesta situação é necessário conhecer/fornecer os valores da estimativa de Kappa e da proporção observada de concordância (P_a) para estimar o valor de P_e:

$$P_e = \frac{P_a - Kappa}{(1 - Kappa)} \qquad 7.15$$

Supondo que pretendíamos com um intervalo de confiança de 95% com uma amplitude inferior a 0.1. Voltando ao exemplo 2.1, onde Pa=0.75 e Kappa=0.51, a dimensão da amostra será calculada de acordo com a expressão 7.14:

$$\mathrm{n} \geq \left(\frac{2 * 1.96 * 0.5134}{0.1}\right)^2 \approx 405.027$$

Desta forma, concluímos que a dimensão da amostra aumenta consideravelmente de 102 para 406 indivíduos, o que já esperávamos pois o intervalo de confiança tem uma menor amplitude, com a mesma confiança, desta forma os dados tratados terão uma maior precisão.

Baseado nos trabalhos (Cantor, 1996) e (Flack, Afifi, Lachenbruch, & Schouten, 1988) é possível calcular a dimensão da amostra para o Kappa de Cohen com dois avaliadores e utilizando uma escala binária ou uma escala multinomial (desde que as distribuições marginais sejam idênticas para os dois avaliadores). Para mais do que dois avaliadores, não foi encontrada literatura sobre este assunto.

No package IRR do R são apresentadas duas rotinas que implementam os métodos acima descritos. O cálculo da dimensão da amostra irá ser feito para uma escala binária (Tabela 7.6), em que a probabilidade de um diagnóstico positivo por parte do avaliador 1 é de 0.6 e do avaliador 2 é de 0.5. Um diagnóstico positivo é a frequência marginal da categoria “1” da tabela 2.1. A tabela 7.7 apresenta os resultados para uma escala multinomial, com as

seguintes frequência marginais: 0.31, 0.45 e 0.24 (idênticas para ambos os avaliadores). Estes valores foram escolhidos apenas para efeitos de representação. Os valores para o erro tipo I e para a potência (π) do teste estatístico serão de 5% (α=0.05, bilateral) e de 80%, respetivamente.

Tabela 7.6. Cálculo da dimensão da amostra para variáveis binárias e dois avaliadores. O valor de K_1 representa o afastamento da hipótese nula (H_0:K_0=0), com probabilidade de um diagnóstico positivo de 0.6 e de 0.5 para os avaliadores 1 e 2, respetivamente

	Valores para K_1			
	0.1	0.3	0.6	0.9
Dimensão da amostra	752	82	19	7

α=0.05 (bilateral) e uma potência de 80%

Como esperado, quanto menor for o afastamento K_1, ou quanto menor for o valor de α, ou quanto maior for o valor da potência, maior será a dimensão da amostra necessária para se obter um diferença significativa..

Tabela 7.7. Cálculo da dimensão da amostra para variáveis multinomiais. O valor de K_1 representa o afastamento da hipótese nula (H_0:K0=0), com probabilidade marginais idênticas (0.31, 0.45 e 0.24) para os dois avaliadores.

	Valores para K_1			
	0.1	0.3	0.6	0.9
Dimensão da amostra	581	64	15	5

α=0.05 (bilateral) uma potência de 80%, respetivamente

Como esperado, quanto maior for o afastamento K_1, menor será o valor da dimensão da amostra necessário para se obter um diferença significativa. As probabilidades marginais têm de ser obrigatoriamente fixas para ambos os avaliadores representando uma restrição no cálculo da dimensão da amostra.

7.3 Inferência estatística para variáveis quantitativas numa situação inter-avaliador e intra-avaliador sem medidas repetidas

No capítulo 4 descrevemos métodos para o cálculo do ICC nas situações inter-avaliadores e intra-avaliadores. Nesta secção, iremos apresentar as respetivas expressões para os limites inferiores e superiores dos intervalos de confiança do ICC (ρ) com um grau de confiança $1-\alpha$., bem como a estatística do teste sob H_0.

O mais comum é testar a hipótese H_0: ρ=0 (não há fiabilidade), contra a hipótese alternativa H_1:ρ>0 (existe fiabilidade). Como indicado por Shrout & Fleiss (1979) testes de significância em relação ao ICC podem ser efetuados através do cálculo da estatística F e do valor da correspondente probabilidade (p-value) de uma tabela ANOVA. As expressões que a seguir apresentamos, foram retiradas de (McGraw & Wong, 1996)

O p-value é definido como a probabilidade da variável aleatória F, que segue uma distribuição F com n-1 e n(k-1) graus de liberdade, exceder F_{obs}, isto é:

$$p-value = P(F \geq F_{obs}) \qquad 7.16$$

Embora os testes de hipótese para ρ=0 sejam comuns, eles não são particularmente informativos. Uma forma de dar resposta a esta limitação consiste em calcular um IC para o ICC e ter em conta a amplitude desse intervalo na interpretação dos resultados. O extremo inferior (LCB) e o extremo

superior (UCB) do intervalo de confiança para o ICC(1,1) são definidos pelas seguintes expressões:

$$LCB = \frac{MS_S - F_1 MS_W}{MS_S + F_1(k-1)MS_W} \quad 7.17$$

onde F_1 representa o percentil de ordem (1-α/2)x100% da distribuição F com n-1 e n(k-1) graus de liberdade

$$UCB = \frac{F_2 MS_S - MS_W}{F_2 MS_S + (k-1)MS_W} \quad 7.18$$

em que F_2 representa o percentil de ordem (1-α/2)x100% da distribuição F com n(k-1) e n-1 graus de liberdade

No caso de estarmos perante o ICC(1,k), os limites inferior e superior do intervalo de confiança são dados respetivamente por:

$$LCB = \frac{MS_S - F_1 \times MS_W}{MS_S} \quad e \quad UCB = \frac{F_2 \times MS_S - MS_W}{F_2 \times MS_S} \quad 7.19$$

onde, F_1 e F_2 estão definidos anteriormente e o respectivo F_{obs} para as classificações individuais é definido por:

$$F_{obs} = \frac{MS_S(1-\rho_0)}{MS_W(1+(k-1)\rho_0)} \quad 7.20$$

No caso de se considerar a média das classificações, o F_{obs} é dado por:

$$F_{obs} = \frac{MS_S \cdot (1-\rho_0)}{MS_W} \quad 7.21$$

Nos modelos de dois fatores, sejam aleatórios ou mistos as expressões para os limites inferiores e superiores são as mesmas. O que irá diferir é em

relação á unidade de medida (individual ou em média) e em relação ao tipo de ICC (concordância absoluta ou consistência).

No caso do ICC concordância absoluta (*absolute agreement*), os limites inferior e superior do respetivo intervalo de confiança nos modelos 2 e 3 ($ICC_A(2,1)$ e $ICC_A(3,1)$) são dadas por:

$$LCB = \frac{n(MS_S - F3 \quad MS_E)}{F3 \quad [kMS_R + (kn - k - n)MS_E] + nMS_S} \tag{7.22}$$

$$UCB = \frac{n(F4 \quad MS_S - MS_E)}{kMS_R + (kn - k - n)MS_E + nF4 \quad MS_S} \tag{7.23}$$

onde F3 é o percentil de ordem $\left(1 - \frac{\alpha}{2}\right).100\%$ da distribuição F com n-1 graus de liberdade para o numerador e *v* graus de liberdade para o denominador, definido por:

$$v = \frac{(aMS_R + bMS_E)^2}{\frac{(aMS_R)^2}{k-1} + \frac{(bMS_E)^2}{(n-1)(k-1)}} \quad e \quad a = \frac{k(\hat{\rho})}{n(1-\hat{\rho})}, b = 1 + \frac{k\hat{\rho}(n-1)}{n(1-\hat{\rho})} \tag{7.24}$$

O valor de F4 é o percentil de ordem $\left(1 - \frac{\alpha}{2}\right).100\%$ da distribuição F com *v* graus de liberdade para o numerador (expressão 7.24) e *n-1* graus de liberdade para o denominador.

No caso do ICC concordância absoluta para a média de classificações ($ICC_A(2,k)$ e $ICC_A(3,k)$), as referidas expressões são dadas por:

$$LCB = \frac{n(MS_S - F3 \quad MS_E)}{F3 \quad (MS_R - MS_E) + nMS_S} \tag{7.25}$$

$$UCB = \frac{n(F4 \quad MS_S - MS_E)}{MS_R - MS_E + nF4 \quad MS_S} \tag{7.26}$$

com $F3$ e $F4$ definidos anteriormente.

No caso do $ICC_A(2,1)$ e $ICC_A(3,1)$, o F_{obs} é dado por:

$$F_{obs} = \frac{MS_S}{aMS_R + bMS_E} \quad onde\ a = \frac{k(\rho_0)}{n(1-\rho_0)} \quad e \quad b = 1 + \frac{k\rho_0(n-1)}{n(1-\rho_0)} \tag{7.27}$$

No caso do $ICC_A(2,k)$ e $ICC_A(3,k)$, vem:

$$F_{obs} = \frac{MS_S}{cMS_R + dMS_E} \quad onde\ c = \frac{\rho_0}{n(1-\rho_0)} \quad e \quad d = 1 + \frac{\rho_0(n-1)}{n(1-\rho_0)} \tag{7.28}$$

onde MS_S ,MS_R e MS_E são os quadrados médios das linhas das colunas e dos erros, respetivamente.

As expressões que a seguir apresentamos referem-se aos limites para o ICC consistency, quando se consideram avaliações individuais ($ICC_C(2,1)$ e $ICC_C(3,1)$):

$$LCB = \frac{MS_S - F_5 MS_E}{MS_S + F_5(k-1)MS_E} \tag{7.29}$$

$$UCB = \frac{F_6 \times MS_S - MS_E}{F_6 \times MS_S + (k-1)MS_E} \tag{7.30}$$

onde F_5 é o percentil $\left(1 - \frac{\alpha}{2}\right).100\%$ da distribuição F com n-1 e (n-1)(k-1) graus de liberdade, para o numerador e denominador, respectivamente e F_6 é o percentil $\left(1 - \frac{\alpha}{2}\right).100\%$ da distribuição F com (n-1)(k-1) e n-1 graus de liberdade, para o numerador e denominador, respetivamente.

No caso de estarmos perante o $ICC_C(2,k)$ e $ICC_C(3,k)$, os limites inferior e superior do intervalo de confiança são dados respetivamente por:

$$\frac{MS_S - F_5 MS_E}{MS_S} \quad e \quad \frac{F_6 MS_S - MS_E}{F_6 MS_S} \tag{7.31}$$

onde F_5 e F_6 foram definidos como anteriormente no ICC consistência. O F_{obs} é dado por:

$$F_{obs} = \frac{MS_S(1-\rho_0)}{MS_E(1+(k-1)\rho_0)} \quad 7.32$$

No caso de se considerar a média das classificações, no $ICC_C(2,k)$ e no $ICC_C(3,k)$, o F_{obs} é dado por:

$$F_{obs} = \frac{MS_S \cdot (1-\rho_0)}{MS_E} \quad 7.33$$

onde MS_S e MS_E são os quadrados médios das linhas e dos erros como já foi definido no capitulo 4.

As formulas apresentadas estão escritas para qualquer valor de ρ. No caso de H_0:ρ=0, situação usual na prática, então as fórmulas anteriores podem ser simplificadas.

Nos *softwares* estatísticos (R, SPSS, SAS), os cálculos dos ICC são feitos através das estimativas dos valores das variâncias apresentadas nos modelos do ICC apresentados no capítulo 4. Este procedimento é alternativo ao uso das tabelas da ANOVA para o cálculo do ICC. Quando existe um valor em falta associado a um determinado sujeito, esse sujeito é retirado da amostra. Gwet (2010) apresenta um método que é robusto á existência de dados em falta, mas a sua implementação em R sai fora do âmbito desta dissertação.

Utilizando o exemplo 4.1, irão ser apresentados os cálculos para o ICC inter-avaliador nos 3 modelos (Tabela 7.8). O mesmo exemplo pode ser aplicado para os três modelos, mas os seus resultados terão interpretações e generalizações diferentes. Estes cálculos foram obtidos usando as rotinas implementadas no R, estando o código produzido no apêndice A.

Tabela 7.8. Resultados para o ICC, considerando uma situação inter-avaliador, considerando H0:ρ=0 e H1:ρ>0.

Modelos	ICC	IC 95%	Fobs	GL1	GL2	p-value
Modelo 1						
ICC(1,1)	0.17	[-0.13,0.72]	1.79	5	18	0.165
ICC(1,k)	0.44	[-0.89,0.91]	1.79	5	18	0.165
Modelo 2						
$ICC_A(2,1)$	0.29	[0.02,0.76]	11	5	4.79	0.011
$ICC_A(2,k)$	0.62	[0.04,0.93]	11	5	4.19	0.017
$ICC_C(2,1)$	0.72	[0.34,0.95]	11	5	15	<0.001
$ICC_C(2,k)$	0.91	[0.68,0.99]	11	5	15	<0.001

GL1: Graus de liberdade do numerador; GL2: Graus de liberdade do denominador.

Na tabela 7.8 não aparece o modelo 3 (de efeitos mistos), porque os resultados obtidos são iguais aos do modelo 2 (de efeitos aleatórios), o que difere é a interpretação dos resultados dos mesmos. Por isso a rotina do R não permite o seu cálculo, mas no SPSS, esta opção aparece, mas os resultados são iguais ao modelo 2, como não podia deixar de ser.

Os valores do ICC para o modelo 1 (apenas os sujeitos são considerados um fator aleatório), quer na unidade de medida individual, quer seja na média são baixos e não significativos. O intervalo de confiança para o caso das avaliações serem individuais apresenta uma maior amplitude, refletindo desta forma uma maior incerteza. Em ambos, o valor 0 pertence ao referido intervalo, o que significa que não existe fiabilidade entre avaliadores.

No modelo 2 (os sujeitos representam um fator e os avaliadores representam um segundo fator), os resultados do ICC são mais elevado (ver tabela 2.5 para a classificação) e significativos. O valor do $ICC_A(2,1)$ é bastante inferior ao $ICC_C(2,1)$ e as amplitudes do intervalo de confiança são elevadas. Isto significa que a variabilidade das pontuações entre os avaliadores

é bastante elevada. Quando a unidade de medida é a média dos resultados obtidos, as estimativas dos ICCs são sempre superiores.

7.4 Cálculo da dimensão da amostra para variáveis quantitativas numa situação inter-avaliador

O cálculo amostral para o ICC inter-avaliador foi apresentado por Zou (2012) e pode ser realizado através do package *ICC.sample.size* do R como iremos mostrar. A tabela 7.9 apresenta os resultados do cálculo amostral do ICC inter-valiador, para um número fixo de avaliadores (4), sob a hipótese nula (H0:ρ0=0), para um α=0.05 (bilateral) e uma potência de 80%, fazendo variar o valor de ρ entre 0 e 1.

Tabela 7.9. Cálculo da dimensão da amostra para o ICC inter-valiador. O valor de ρ representa o afastamento da hipótese nula (H0:ρ0=0), com um número de avaliadores iguais 4.

	Valores para ρ (H1:ρ0>0)			
	0.1	0.3	0.6	0.9
Dimensão da amostra	156	22	7	3

α=0.05 (bilateral) e uma potência de 80%.

Como esperado, quanto maior for o afastamento ρ da hipótese nula, menor será o valor da dimensão da amostra necessário para se obter um diferença significativa. O valor de ρ0 pode assumir valores diferentes de 0, estando a rotina do R (ICC.sample.size) preparada para o efeito.

7.5 Inferência estatística para variáveis quantitativas numa situação inter-avaliador e intra-avaliador de medidas repetidas

No capítulo 5 descrevemos métodos para o cálculo do ICC para as situações inter-avaliador e intra-avaliador com medições reptidas. Nesta secção, iremos apresentar as fórmulas para os limites inferiores dos intervalos de confiança para o ICC (ρ) com um nível de confiança $1-\alpha$., bem como a estatística do teste sob H_0 apresentados por Eliasziw et al(1994) (Eliasziw, Young, Woodbury, & Fryday-Field, 1994) .

No caso de haver repetições, ou seja, m>1, o intervalo de confiança do inter-avaliador de efeitos aleatórios para concordância absoluta fica:

$$IC_{1-\alpha} = \left(\frac{n(MS_S - F_7 MS_E)}{F_7[k(MS_R - MS_E) + n(k-1)MS_E + nk(m-1)MS_{RE}] + nMS_S}, 1\right) \quad 7.34$$

onde F_7 é o percentil (1-α).100% da distribuição F com n-1 e v_1 graus de liberdade, onde v_1 é dado por:

$$v1 = \frac{(n-1)(k-1)\{k\rho(MS_R - MS_E) + n[1+(k-1)\rho]MS_E + nk(m-1)\rho MS_{RE}\}^2}{(n-1)(k\rho)^2 MS_R^2 + \{n[1+(k-1)\rho] - k\rho\}^2 MS_E^2 + (n-1)(k-1)[nk(m-1)]\rho^2 MS_{RE}^2} \quad 7.35$$

sendo ρ, o coeficiente de fiabilidade apresentado no capitulo 5, expressão 5.2.

O intervalo de confiança para a situação inter-avaliador de efeitos fixos para a consistência fica:

$$IC_{1-\alpha} = \left(\frac{n(MS_S - F_8 MS_E)}{F_8[n(k-1)MS_E + nk(m-1)MS_{RE}] + nMS_S}, 1\right) \quad 7.36$$

onde F_8 é o percentil (1-α).100% da distribuição F com *n-1* e v_2 graus de liberdade, onde v_2 é dado por:

$$\mathrm{v2} = \frac{(n-1)(k-1)\{n[1+(k-1)\rho]MS_E + nk(m-1)\rho MS_{RE}\}^2}{\{n[1+(k-1)\rho]\}^2 MS_E^2 + (n-1)(k-1)[nk(m-1)]\rho^2 MS_{RE}^2} \quad 7.37$$

sendo ρ, o coeficiente de fiabilidade apresentado no capitulo 5, expressão 5.3.

Para ambos os casos, a estatística do teste é dada pela mesma relação:

$$F_{inter} = \frac{MS_S \times (1-\rho_0)}{MS_E \times (1+(k-1)\rho_0)} \quad 7.38$$

com uma distribuição F com (n-1) e n(k-1) graus de liberdade.

Quando estamos perante um desenho de medidas repetidas, em ambos os modelos, efeitos fixos ou efeitos aleatórios, a estatística de teste usada, quando é estimada a fiabilidade intra-avaliador em avaliações gerais (overall) e o seu intervalo de confiança é dado por:

$$F_{intra} = \frac{MS_S/k \times (1-\rho_0)}{MS_{RE}(1+(m-1)\rho_0)} \quad 7.39$$

$$IC_{1-\alpha} = \left(\frac{MS_S/k - F_8 MS_{RE}}{MS_S/k + F_9(m-1)MS_{RE}}, 1\right) \quad 7.40$$

onde F_9 é o percentil (1-α).100% da distribuição F com (n-1) e n(m-1) graus de liberdade.

A estatística do teste quando são consideradas avaliações individuais é dada por:

$$F_{intra,j} = \frac{MS_S/k \times (1-\rho_0)}{MS_{REj}(1+(m-1)\rho_0)} \quad 7.41$$

e o seu intervalo de confiança para avaliações individuais é dado por:

$$IC_{1-\alpha} = \left(\frac{MS_S/k - F_6 MS_{REj}}{MS_S/k + F_9(m-1)\,MS_{REj}}, 1\right) \quad 7.42$$

onde F_9 foi definido anteriormente.

Os intervalos de confiança apresentados são unilaterais, sendo a parte inferior a mais relevante em problemas de fiabilidade. Note-se que o limite inferior nunca excede 1, pois o coeficiente de fiabilidade assume sempre 1 como valor máximo (Eliasziw et al., 1994).

Voltando ao exemplo 5.1, onde os dados representam um estudo de teste/re-teste, para avaliar o nível de fiabilidade de 2 goniómetros utilizados na medição de um ângulo associado a uma articulação com uma amostra de 29 doentes e com três repetições consecutivas pelos dois goniómetros. Os resultados apresentados foram obtidos a partir da rotina relIntraInter do package IRR

Tabela 7.10 Resultados para o ICC, considerando uma situação inter-avaliador (H0:ρ=0.0 e H1:ρ>0.0) e intra-avaliador (H0:ρ=0.0 e H1:ρ>0.0).

Modelos	ICC	IC 95%	Fobs	GL1	GL2	p-value
Inter-avaliador						
$ICC_A(2,1)$	0.9451	[0.85;1]	17.4	28	9.02	<0.001
$ICC_C(2,1)$	0.9612	[0.94;1]	17.4	28	51.07	<0.001
Intra-avaliador						
$ICC_A(2,1,m)$	0.9842	[0.97;1]	14.0	28	58	<0.001
$ICC_A(2,1,m,av1)$	0.9864	[0.98;1]	16.4	28	58	<0.001
$ICC_A(2,1,m,av2)$	0.9820	[0.96;1]	12.3	28	58	<0.001
$ICC_C(2,1,m)$	0.9840	[0.97;1]	14.0	28	58	<0.001
$ICC_C(2,1,m,av1)$	0.9862	[0.98;1]	16.4	28	58	<0.001
$ICC_C(2,1,m,av2)$	0.9818	[0.96;1]	12.3	28	58	<0.001

Como indicado no artigo do Eliasziw et al (1994), os valores para o F_{obs} numa situação inter-avaliador dão o mesmo resultado, acontecendo o mesmo para a situação intra-avaliador. Os valores dos ICC para ambas as situações são classificados como quase perfeitos (ver tabela 2.5) e significativos. Os

intervalos de confiança mostram pequenas amplitudes indicando que a qualidade das estimativas produzidas é bastante boa.

7.6 Cálculo da dimensão da amostra para variáveis quantitativas numa situação inter-avaliador e intra-avaliador de medidas repetidas

Como para qualquer estudo, uma amostra pequena conduz a uma estimativa imprecisa do coeficiente de fiabilidade e a uma grande amplitude do intervalo de confiança. Os autores (Walter, Eliasziw, & Donner, 1998) e (Donner & Eliasziw, 1987) propuseram dois gráficos (Figura 7.1 e Figura 7.2) para determinar a dimensão da amostra. Neste tipo de estudos de medidas repetidas, a fiabilidade inter-avaliador exige uma maior dimensão da amostra do que a fiabilidade intra-avaliador, devido ao erro associado ao inter-avaliador ser superior ao erro associado ao intra-avaliador (medidas repetidas). Neste último, como é o mesmo avaliador a realizar as várias repetições, conduz automaticamente a uma redução de variabilidade e por conseguinte a um menor número de sujeitos necessários.

Os autores atrás referidos estabeleceram valores mínimos aceitáveis para as hipóteses nulas. Para a fiabilidade inter-avaliador, as hipóteses estatísticas são definidas da seguinte forma: $H0:\rho=0.6$ e $H1:\rho>0.6$ enquanto que para a fiabilidade intra-avaliador, as hipóteses estatísticas são: $H0:\rho=0.8$ e $H1:\rho>0.8$. Esta diferença está relacionada com o facto do ICC para o intra-avaliador ter resultados mais elevados do que numa situação inter-avaliador. Estes também são baseados na tabela 2.5 (Landis JR, 1977), onde os valores do ICC são moderados para inter-avaliador e substanciais para intra-avaliador.

A figura 7.1 apresenta uma estimativa para o cálculo do tamanho da amostra, para a situação de se testar H0: $\rho=0.6$ vs H1: $\rho>0.6$ (inter-avaliador) com um nível de significância 5% e 80% de potência do teste. Neste gráfico é apresentado duas possíveis cenários: para o caso do verdadeiro valor do ICC ser 0.8 (considerando uma distância de 0.20 em relação ao valor de H0) ou

para o caso do verdadeiro valor do ICC ser 0.75 (nesta situação a distância será de 0.15 em relação ao valor de H0). Por exemplo, para o caso de dois avaliadores, observa-se que são necessários 35 sujeitos para um estudo inter-avaliador, quando o verdadeiro valor do ICC é de 0.8.

A figura 7.2 apresenta uma estimativa para o cálculo do tamanho da amostra, para a situação de se testar H0: ρ=0.8 vs H1: ρ>0.8 (intra-avaliador) com um nível de significância 5% e 80% de potência do teste. Como na figura anterior, são apresentados duas possíveis cenários: para o caso do verdadeiro valor do ICC ser 0.95 (considerando uma distância de 0.15 em relação ao valor de H0) ou para o caso do verdadeiro valor do ICC ser 0.90 (nesta situação a distância será de 0.10).

Por exemplo, partindo do número necessário de sujeitos do exemplo anterior (35 sujeitos para um estudo inter-avaliador), observa-se seriam necessárias 3 medições por avaliador (estudo intra-avalaidor) quando o verdadeiro valor do ICC é 0.9.

7.7 Inferência estatística em estudos com variáveis classificadas por ratings

7.7.1 Correlação de Spearman

O estudo da inferência estatística para o coeficiente de Spearman (r_S) irá depender da dimensão da amostra. Para amostras grandes (n≥10), podemos determinar a significância obtida de r_S observado sob a hipótese nula (H0:ρ_s=0; H1:ρ_s≠0) através da expressão:

$$T = R_S \sqrt{\frac{n-2}{\sqrt{1-R_S^2}}} \qquad 7.43$$

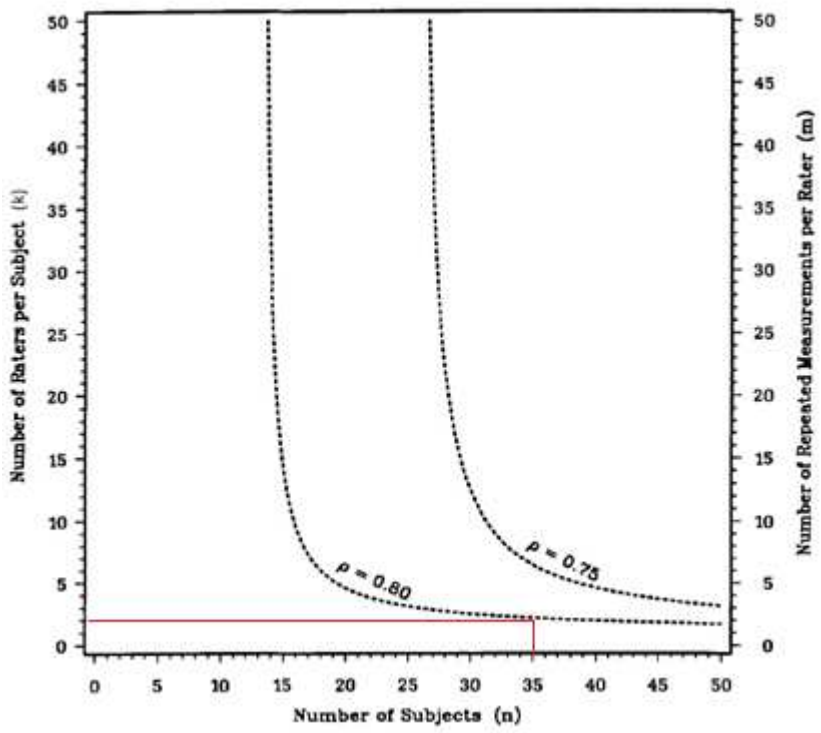

Figura 7.1.Estimativa do tamanho da amostra para testar H0: ρ=0.6 vs H1: ρ>0.6 com um nível de significância 5% e 80% de potencia do teste.

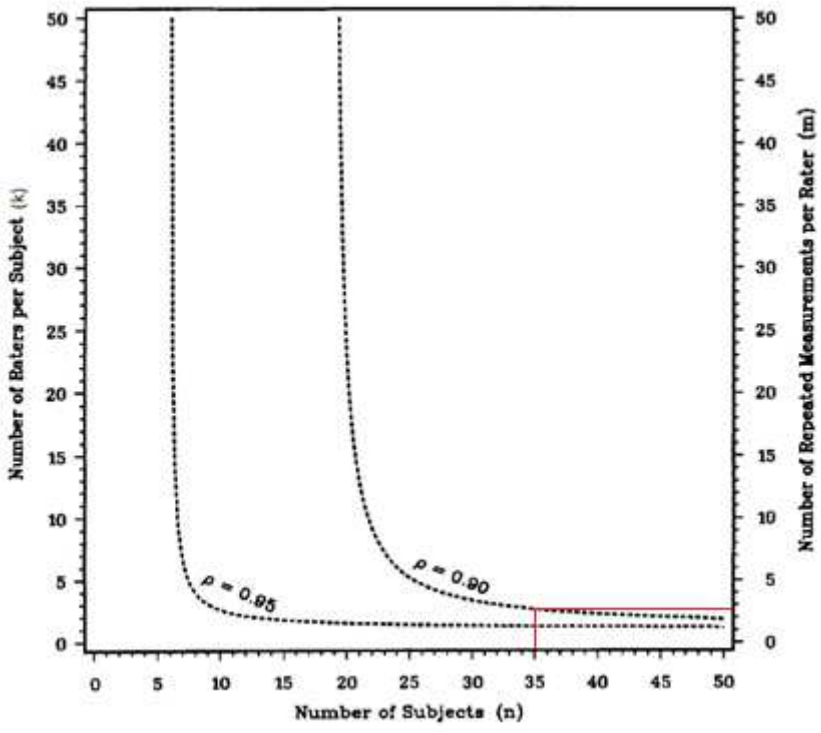

Figura 7.2. Estimativa do tamanho da amostra para testar H0: ρ=0.8 vs H1: ρ>0.8 com um nível de significância 5% e 80% de potencia do teste.

onde a estatística T segue uma distribuição T de Student com n-2 graus de liberdade.

Para amostras de dimensão pequenas (4≤n≤30), é possível determinar a significância obtida através de um tabela especifica onde são apresentados os valores críticos para os níveis de significância 0.05 e 0.01. Mais detalhes sobre este procedimento podem ser encontrados em Siegel & Castellan (1988).

No exemplo 6.1, obtivemos o coeficiente de Sperman R_S=0.74 e o p-value associado é 0.002. Desta forma o valor obtido para o referido coeficiente é estatisticamente significativo. A rotina utilizada (ver apêndice A: cor.test) não permite o cálculo do intervalo de confiança.

7.7.2 Correlação de Kendall tau

Quando a dimensão da amostra é pequena, o processo de cálculo é bastante moroso, à medida que n aumenta. No entanto, para n≥ 8, a distribuição de amostragem de Kendall τ é praticamente indistinguível da distribuição normal standard.

Para amostras em que n>10, a distribuição amostral do Kendall τ , sob a hipótese nula é conhecida e, portanto, Kendall τ é sujeito a testes de significância. Consideremos a estatística Z que segue uma distribuição normal standard, definida da seguinte forma:

$$Z = \frac{3\hat{\tau}\ \sqrt{n(n-1)}}{\sqrt{2(2n+5)}} \qquad 7.44$$

onde $\hat{\tau}$ é o coeficiente de correlação de Kendall τ e n é o número de sujeitos dessa amostra.

No exemplo 6.1, obtivemos o coeficiente de Kendal τ =0.79 e o p-value associado é 0.04, desta forma o valor obtido para o referido coeficiente é estatisticamente significativo para um nível de significância de 0.05. A rotina utilizada (ver apêndice A: cor.test) não permitiu o cálculo do intervalo de confiança.

7.5.3 Coeficiente Kendall W

Para amostras grandes (N>7), a estatística de teste de Kendall W é dada por:

$$X^2 = k(n-1)W, \qquad 7.45$$

onde k é o número de avalidadores e W é o coeficiente de concordância de Kendall (KCC apresentado no capitulo 6), seguindo uma distribuição do Qui-quadrado com n-1 graus de liberdade.

No caso da dimensão da amostra ser pequena (3≤n≤7), podemos testar o significado do valor observado de W através do cálculo da probabilidade, sob $H_{0,}$ deste valor ser tão elevado quanto a soma de todos os quadrados das diferenças entre as somas marginais dos rankings R_i e a sua média global $\bar{R}$. Mais detalhes podem ser encontrados em Siegel & Castellan (1988).

O p-value representa a probabilidade da variável aleatória X^2 exceder o valor observado:

$$p_{value} = P\left(X^2 > X^2_{obs} | H_{0)}\right) \qquad 7.46$$

Se o número de avaliadores ou o número de sujeitos for pequeno demais para a distribuição Qui-quadrado proporcionar uma aproximação adequada, então os autores Siegel & Castellan (1988) sugerem avaliar o significado de W com um valor de Qui-quadrado ajustado. A rotina utilizada para o cálculodo

Kendall W foi a Kendall. Voltando ao exemplo 6.3, a tabela 7.11 apresenta os resultados para os coeficientes de correlação de Spearman, Kendall τ, e Kendall W ajustados para empates.

Tabela 7.11. Cálculo dos coeficientes de Spearman, Kendall τ e Kendall W para o exemplo 6.3.

Método	Coeficiente	p-value
Spearman	0.79	0.020
Kendall tau	0.63	0.039
Kendall W	0.90	0.084

Desta forma podemos concluir que os 2 médicos estão em concordância em relação á eficiência do novo inclinómetro digital para medir a amplitude do movimento do ombro esquerdo nos seus pacientes, quando se considera o coeficientes de Sperman e o Kendall τ . O mesmo não se pode afirmar quando se utiliza o Kendall W, este apresenta um p-value superior a 0.05. O resultado da não significância é justificado pela dimensão da amostra reduzida deste exemplo.

Para o exemplo 6.4, apenas é possível calcular o Kendall W dado que existem mais do que dois avaliadores. O valor do coeficiente de Kendall W é 0.80 e o p-value associado é 0.002, existindo evidência estatística de concordância entre os avaliadores

Referências

Altaye, M., Donner, A., & Eliasziw, M. (2001). A general goodness-of-fit approach for interference procedures concerning the kappa statistic. *Statistics in Medicine*, *20*(16), 2479–2488.

Cantor, A. B. (1996). Sample-Size Calculations for Cohen's Kappa.

Psychological Methods, *1*(2), 150–153.

Donner, a, & Eliasziw, M. (1987). Sample size requirements for reliability studies. *Statistics in Medicine*, *6*(4), 441–448. http://doi.org/10.1002/sim.4780060404

Eliasziw, M., Young, S. L., Woodbury, M. G., & Fryday-Field, K. (1994). Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. *Physical Therapy*, *74*(8), 777–88. http://doi.org/10.1186/1471-2474-7-60

Flack, V. F., Afifi, A. A., Lachenbruch, P. A., & Schouten, H. J. A. (1988). Sample size determinations for the two rater kappa statistic. *Psychometrika*, *53*(3), 321–325.

Fleiss, Joseph L.; Cohen, Jacob; Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, *72*, 323–327.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*. http://doi.org/10.1037/h0031619

Gwet, K. L. (2010). *Handbook of Inter-Rater Reliability: the definitive guide to measuring the extent of agreement among raters*. *Gaithersburg, MD: STATAXIS Publishing Company*. Advanced Analytics, LLC.

Landis JR, K. G. (1977). *The measurement of observer agreement for categorical data.* Biometrics.

McGraw, K. O., & Wong, S. P. (1996). “Forming inferences about some intraclass correlations coefficients”: Correction. *Psychological Methods*, *1*(4), 390–390. http://doi.org/10.1037/1082-989X.1.4.390

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass Correlation: Uses in Assessing Rater Reliabitlity. *Psychological Bulletin*, *86*(2), 420–428.

Siegel, S., & Castellan, J. (1988). *Nomparametric Statistics for the behavioral science*. (1988 McGraw-Hill, Ed.).

Sim, J., & Wright, C. C. (2005). Interpretation, and Sample Size Requirements The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *PHYS THER. Physical Therapy*, *85*(3), 257–268. http://doi.org/15733050

Walter, S. D., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine*, *17*(1), 101–110. http://doi.org/10.1002/(SICI)1097-0258(19980115)17:1<101::AID-SIM727>3.0.CO;2-E

Zou, G. Y. (2012). Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. *Statistics in Medicine*, *31*(29), 3972–3981. http://doi.org/10.1002/sim.5466

Capitulo 8. Discussão e conclusões

Neste capítulo são apresentados as principais linhas de discussão e conclusões desta dissertação. A última secção será dedicada ao trabalho futuro.

8.1 Discussão

Como referido anteriormente, a fiabilidade é definida como a capacidade de um instrumento de medição diferenciar sujeitos, enquanto que, a concordância é definida como o grau em que scores ou pontuações medidas no mesmo sujeito são idênticas (Kottner et al., 2011). Os parâmetros de concordância são muitas vezes omitidos na literatura clínica, havendo uma preferência pelos parâmetros de fiabilidade. No entanto estes conceitos são diferentes.

Os parâmetros de fiabilidade são necessários para instrumentos usados com objetivos discriminatórios enquanto os parâmetros de concordância são utilizados para fins de avaliação relacionados com o erro desse instrumento. Por exemplo, um instrumento que seja discriminativo requer um elevado nível de fiabilidade, ou seja, o erro de medida deve ser relativamente pequeno quando comparado com a variabilidade entre sujeitos que o instrumento necessita de distinguir. Se a diferença entre os sujeitos é grande, um certo nível de erro de medida é aceitável. Para o erro de medida de um instrumento, a variabilidade entre sujeitos não é relevante ou sequer contabilizada. Por exemplo, numa situação de medidas repetidas, não é relevante a variabilidade entre sujeitos (*between subjects*), mas sim a variabilidade do mesmo sujeito que é relevante ao longo das medições (*within subjects*). Quanto menor for este erro de medição, maior será a precisão da medida efetuada. Se este

erro de medição for grande, então pequenas mudanças não serão detectadas pelo instrumento.

No artigo “When to use agreement versus reliability measures” (Vet, 2006), Vet apresenta um exemplo onde dois fisioterapeutas medem a amplitude do movimento do ombro com um inclinómetro digital expresso em graus. O estudo é efetuado com pacientes com o ombro afetado e pacientes com o ombro não afetado. Em ambos os casos o erro de medição é idêntico e baixo, no entanto o valor do coeficiente de fiabilidade, o ICC, é de 0.83, no caso do ombro afetado e 0.28, se o ombro dos pacientes não estiver afetado. O valor obtido para o ICC é bom no caso do ombro estar afetado (ICC=0.83), o que indica que o instrumento tem uma boa capacidade de discriminar os pacientes, mas é fraco, quando o ombro é bom (ICC=0.28), que nos indica que nesta situação o instrumento não tem uma boa capacidade para discriminar os pacientes. Os resultados obtidos parecem contraditórios, mas de facto, como o erro de medição do instrumento é idêntico nos dois casos (parâmetro de concordância), o que difere é o parâmetro de fiabilidade. No caso do ombro afetado, esta é elevada, o que indica que existe uma boa capacidade para discriminar os sujeitos enquanto no caso do ombro não afetado a fiabilidade é baixa, o que significa que o instrumento não tem capacidade paro os discriminar.(Vet et al., 2006)

Nos seguintes fluxogramas são apresentados os métodos de concordância e fiabilidade mais usuais em função do tipo de dados e do número de avaliadores. Estes fluxogramas foram baseados no trabalho apresentado por McGraw(McGraw & Wong, 1996). Estes autores apenas apresentam um fluxograma para a escolha do coeficiente de correlação intraclasse apropriado para a situação inter-avaliador.

Na figura 8.1 é apresentado o fluxograma geral para o estudo de um problema de fiabilidade ou concordância. Se o tipo de dados for nominal deveremos optar pelos métodos baseados na correcção de chance apresentados no capítulo 2. Por outro lado, se os dados forem ordinais, temos duas opções de análise, se estes estiverem agrupados em poucas categorias (por exemplo, menos ou iguais a 5 categorias) devemos utilizar os métodos baseados na correcção de chance ponderados (capítulo 3), caso contrário, deve optar-se por métodos baseados em rankings (capítulo 6). Por último, se os dados forem quantitativos então também

existem duas opções para a escolha do método mais apropriado. Se optarmos pela análise dos valores em absoluto então os métodos de fiabilidade serão os métodos baseados no rácio das variâncias (capítulos 4 e 5). Se optarmos pelas posições (rankings) que esses valores numéricos ocupam então estaremos numa análise baseada em rankings (capítulo 6).

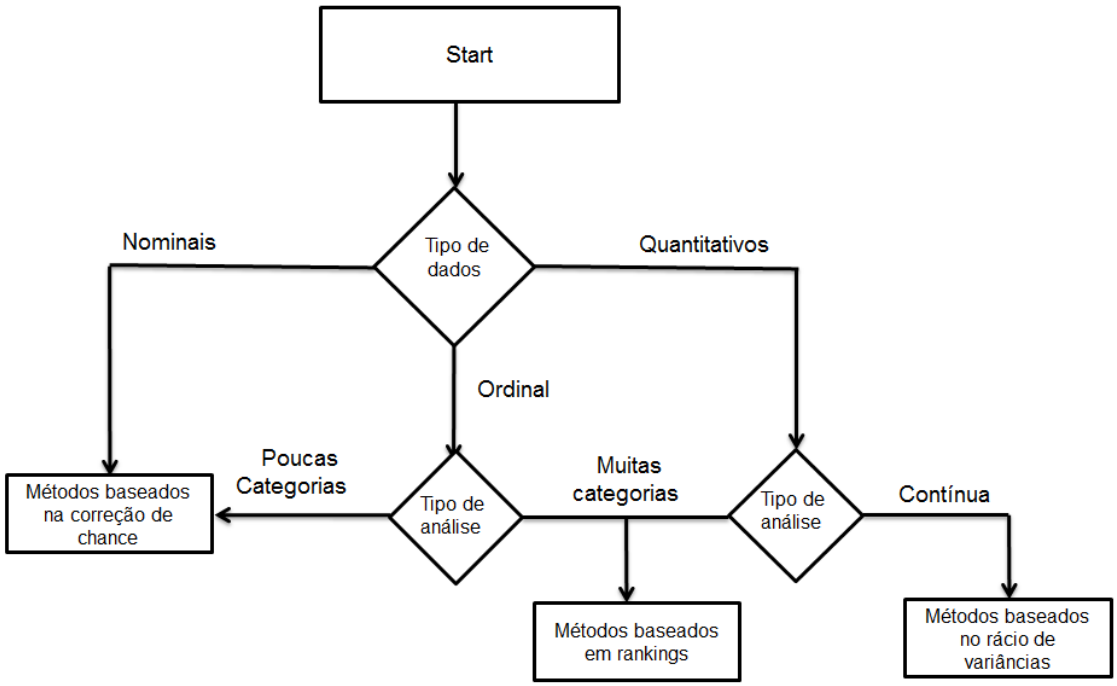

Figura 8.1. Fluxograma geral para um estudo fiabilidade ou concordância baseado no tipo de dados medidos.

Nas figuras seguintes, cada folha do fluxograma geral irá ser apresentada com mais detalhe. Na figura 8.2 estão apresentados os métodos mais usuais baseados na correcção de chance (ponderados ou não ponderados) em função do número de avaliadores. No caso dos dados se apresentarem numa escala nominal então devemos utilizar os coeficientes kappa não ponderados, como os que foram apresentados no capítulo 2. Se os dados forem ordinais (mas com um número de categorias baixo), então deve-se utilizar pesos nos coeficientes kappa para "penalizar" as discordâncias obtidas. Este raciocínio deve ser aplicado

independentemente do número de avaliadores utilizados, mudando apenas os testes estatísticos.

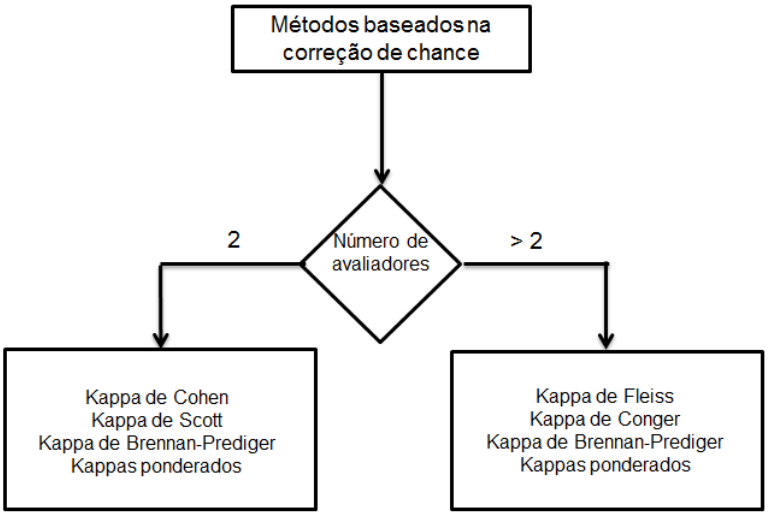

Figura 8.2. Fluxograma para métodos baseados na correção de concordância.

Na tabela 8.3 é apresentado o fluxograma dos métodos baseados em rankings. No caso de variáveis ordinais com muitas categorias, os métodos ponderados apresentados na figura 8.2 irão exibir valores de fiabilidade ou concordância baixos, devido á possibilidade de haver muitas categorias de resposta. Por exemplo, respostas em categorias adjacentes irão ser classificadas como discordâncias. Por outro lado, o recurso aos métodos baseados no rácio de variâncias para analisar dados ordinais com muitas categorias pode ser fortemente influenciado pela existência de valores atípicos, como por exemplo *outliers* moderados ou severos . Desta forma os métodos baseados em rankings apresentam-se como uma alternativa para este tipo de dados. Como no caso anterior, este raciocínio deve ser aplicado independentemente do número de avaliadores utilizados, mudando apenas os testes estatísticos.

Para a questão se os métodos de concordância/fiabilidade para variáveis nominais ou ordinais classificados em categorias se podem aplicar a situações de intra-avaliadores, não conseguimos obter uma resposta clara. No entanto no artigo onde são propostas orientações (guidelines) para estudos de fiabilidade e

concordância, os métodos apresentados para a análise inter-avaliador são os mesmos para uma situação intra-avaliador (Kottner et al., 2011).

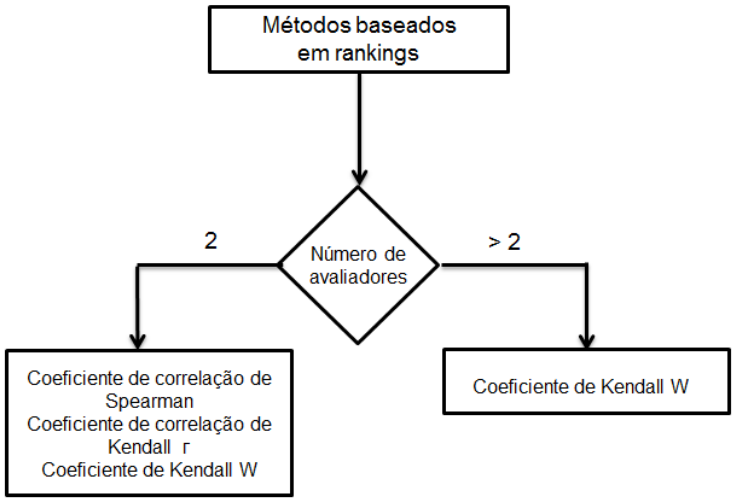

Figura 8.3. Fluxograma para métodos baseados em rankings.

Quando as variáveis são quantitativas os métodos a utilizar são os métodos baseados no rácio das variâncias. Em primeiro lugar teremos de decidir se os dados serão tratados por um modelo da ANOVA de um fator ou de dois fatores. A utilização de um ou dois fatores está relacionado com a inclusão ou não do efeito dos avaliadores no estudo. Na literatura médica, os modelos de dois fatores são bastante mais utilizados que os modelos de apenas um fator.

Se o único fator a ser avaliado é o fator sujeito e considerando que o fator avaliador possa ser desprezado da análise (por exemplo, nas situações onde cada sujeito é avaliado por um conjunto de avaliadores diferentes) então deve-se optar por um modelo de um fator (figura 8.4). Nesta situação, a primeira análise deve estar relacionada com o tipo de avaliação, se é uma avaliação intra-avaliador (mesmo avaliador em instantes de tempo diferentes) ou se representa uma situação inter-avaliador (diferentes avaliadores no mesmo instante de tempo). Para esta última situação uma possível divisão pode ser aplicada. Os resultados podem ser apresentados em forma de média (quando o erro de medição é dividido pelo número de avaliadores) ou de uma forma individual (quando o erro de medição não é

corrigido pelo número de avaliadores). Os valores de fiabilidade apresentados num formato de média, têm tendência a ser superiores, dado que o valor do erro de medição é menor, como foi referido no capítulo 4.

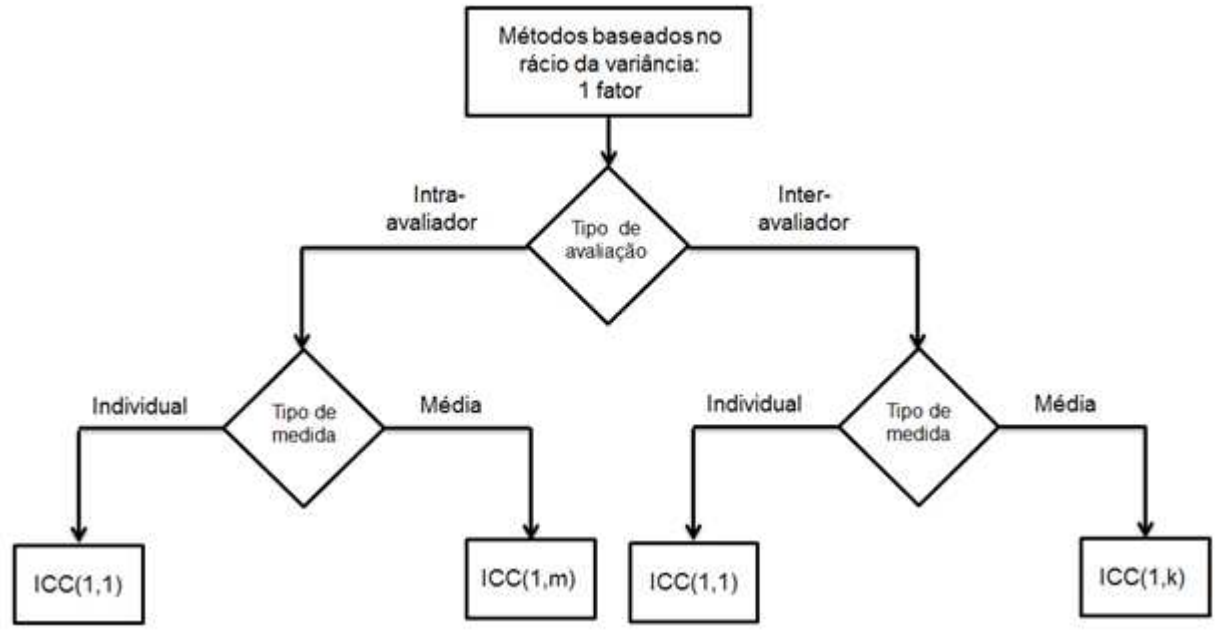

Figura 8.4. Fluxograma para métodos baseados no rácio da variância para 1 fator.

Para o caso em que o fator avaliador é considerado relevante para o estudo, então deve-se utilizar um modelo de dois factores (fator sujeito e o fator avaliador). Por questões de simplicidade de análise, os casos sem réplicas (figura 8.5) e com réplicas (figura 8.6) são apresentados e descritos separadamente.

No modelo de dois fatores (seja numa situação inter-avaliador ou intra-avaliador) é importante decidir pelo tipo de fiabilidade para o nosso estudo. A opção consistência deve ser considerada quando a variabilidade dos avaliadores não é considerada para o estudo e portanto não será incluída. Com a opção concordância absoluta esta variabilidade é considerada importante sendo desta forma incluída no cálculo da fiabilidade. Como nos modelos de um fator, a última questão será relativa á forma de apresentação dos resultados, podendo-se optar por duas diferentes unidade de medida de interesse (individual ou em média).

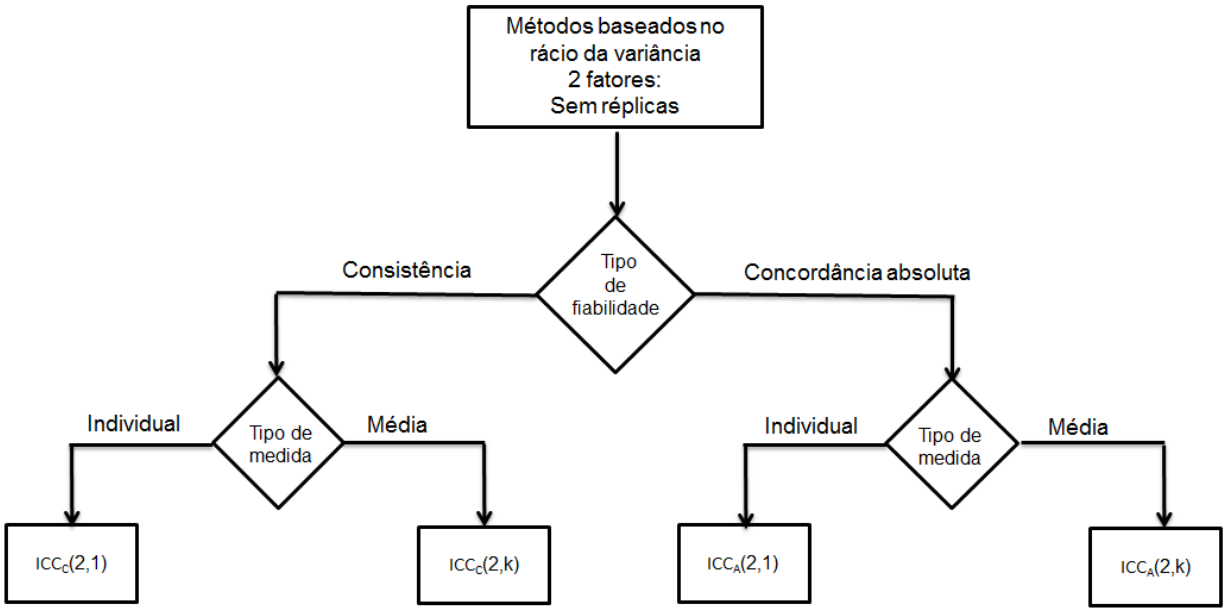

Figura 8.5. Fluxograma para métodos baseados no rácio da variância para 2 fatores sem réplicas.

Como se mostrou nos capítulos anteriores, a escolha dos avaliadores ser aleatória ou fixa conduz ás mesmas expressões para o ICC, apesar de conceptualmente provirem de modelos matemáticos distintos. Os resultados obtidos para a fiabilidade serão idênticos diferindo apenas na sua generalização. Se a escolha dos avaliadores for aleatória permite a generalização para outros avaliadores enquanto se a escolha dos avaliadores for fixa, esta generalização não poderá ser efectuada. Do nosso ponto de vista, para a figura 8.5, acrescentar mais um nível que separe os métodos de dois fatores em efeitos aleatórios e efeitos mistos aumenta a complexidade de uma forma desnecessária.

A figura 8.6 apresenta o fluxograma para uma situação intra e inter-avaliador com múltiplas medições. Como foi referido para a figura 8.5, não é necessário separar os modelos apresentados em efeitos aleatórios ou efeitos mistos, dado que as fórmulas dos ICCs são idênticas em ambos os casos, mudando apenas a interpretação dos resultados obtidos. Novamente, o que faz sentido na nossa opinião é dividir os modelos em consistência ou em concordância absoluta. Mostrou-se no capítulo 5, que os modelos para o inter-avaliador são idênticos aos modelos do ICC apresentados no capítulo 4 para o tipo de medida individual. Por

último, nos modelos intra-avaliadores é possível descriminar o resultado geral dos resultados individuais obtidos por cada avaliador.

A tabela 8.1 apresenta de forma sumária os métodos estatísticos utilizados nesta dissertação, distinguindo níveis de medidas, testes para fiabilidade e testes para concordância. Esta tabela foi inspirada no trabalho realizado por Kottner et al. (2011) onde são propostas orientações para estudos de fiabilidade e de concordância.

Tabela 8.1. Métodos estatísticos para estudos de fiabilidade e de concordância intra-avaliador e inter-avaliador

Nível de medida	**Medidas de fiabilidade**	**Medidas de concordância**
Nominal	Kappa de Cohen Kappa de Scott Kappa de Fleiss Kappa de Conger Kappa Brennan-Prediger	Proporção de concordância Proporção de concordância específica
Ordinal com poucas categorias (≤5)	Estatísticas kappa ponderadas	Idêntica às nominais
Ordinal com muitas categorias (>5)	Coeficiente de Spearman Coeficiente de Kendall tau Coeficiente Kendall W	Idêntica às nominais
Contínuas	Coeficientes de correlação intraclasse	Proporção de concordância em intervalo de amplitudes Erro padrão da medida (SEM) Limites de concordância de Bland-Altman*

* Não foi descrito nesta dissertação

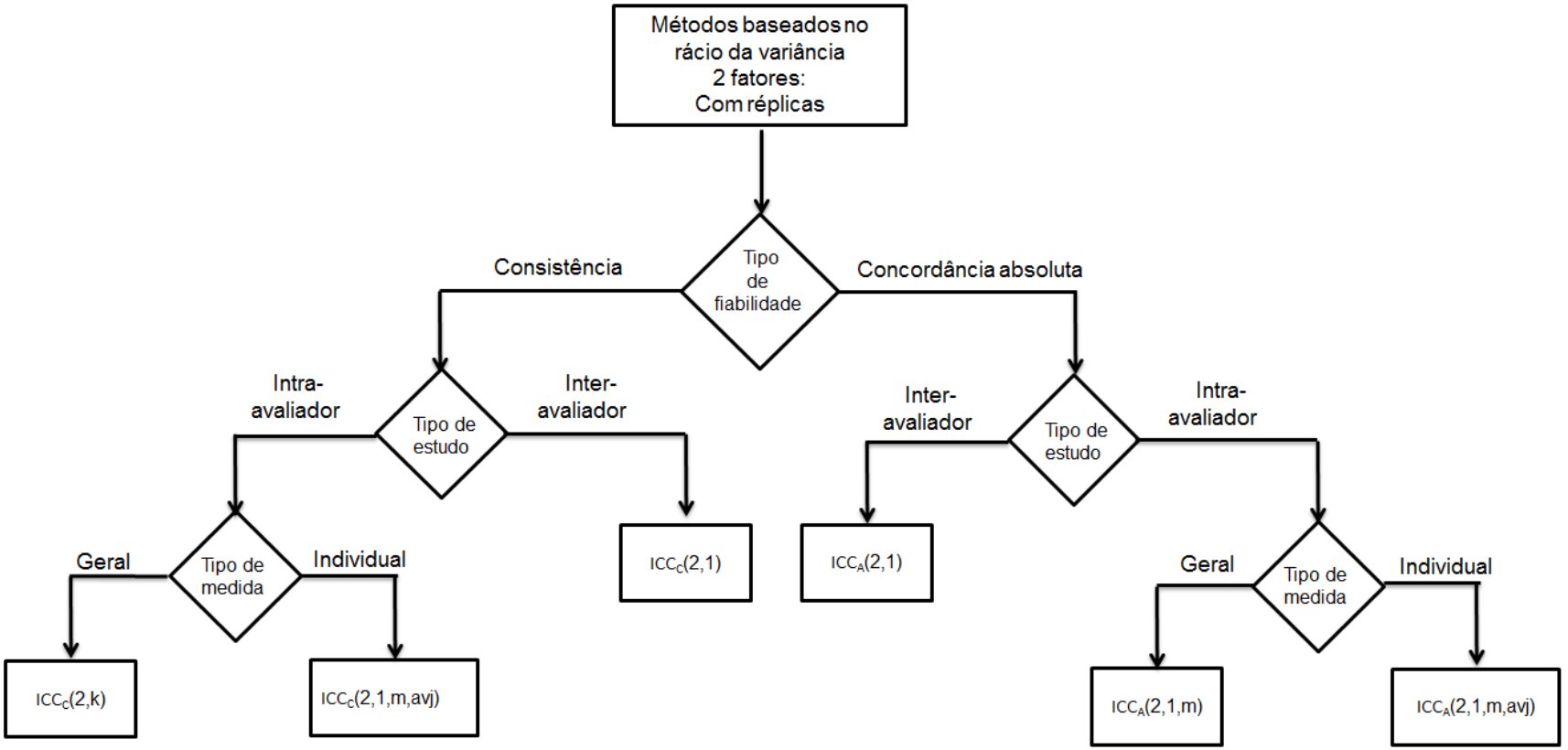

Figura 8.6. Fluxograma para métodos baseados no rácio da variância para 2 fatores com réplicas.

8.2 Conclusões

Os métodos de fiabilidade/concordância que são descritos na literatura são numerosos, com diferentes objetivos e aplicações, e por vezes confusos, dispersos e contraditórios. Nesta dissertação, o objetivo principal é clarificar em que situações se devem aplicar um determinado conjunto de métodos de fiabilidade/concordância. Para cada método foi feita apresentação do respetivo estimador e sua inferência estatística. .bem como a exemplificação da sua aplicação a casos concretos.

O diagrama inicial (Figura 8.1) apresentado neste capítulo permite de uma forma clara estabelecer conjuntos de métodos de fiabilidade/concordância, baseados no tipo de variável medida no estudo. As Figuras 8.2 a 8.6 introduzem as noções de nº de avaliadores, tipo de análise (inter-avaliador ou intra-avaliador), modelos de 1 fator ou dois fatores, sem réplicas ou com réplicas, o tipo de fiabilidade (concordância absoluta ou consistência) e finalmente a unidade de medida (singular ou em média). Nas folhas destes diagramas estão os testes estatísticos abordados nesta dissertação. Apesar de alguns destes diagramas estarem já apresentados na literatura (a situação inter-avaliador é conhecida e apresentada por McGraw & Wong (1996), uma apresentação deste problema de uma forma tao completa é original, segundo o nosso conhecimento do estado-da-arte.

Para cada método apresentado nos diagramas anteriores foram apresentados e derivados os respetivos coeficientes de fiabilidade e de concordância. A sua inferência estatística, um exemplo ilustrativo e a validação da análise dos resultados é outra mais-valia desta dissertação. O cálculo amostral também foi realizado para os métodos mais usuais na prática.

Para os métodos baseados em variáveis qualitativas nominais ou ordinais classificadas por categorias, foi possível identificar os principais métodos para dois ou mais avaliadores. No entanto, subsiste a dúvida se estes métodos podem ser aplicados numa situação intra-avaliador.

Para os métodos baseados em variáveis quantitativas, numa situação sem réplicas, foi identificado, apresentado e demonstrado a existência de 3 modelos (1 fator, 2 fatores aleatórios e 2 fatores mistos), bem como as diferenças entre eles. Para uma situação com réplicas, processo similar foi produzido, com a inclusão nos modelos anteriores da variabilidade existente devido ás repetições das medições.

A aplicabilidade do modelo de um fator num contexto de situações reais levanta algumas questões, levando os investigadores a optarem “quase sempre” pelos modelos de dois fatores. Na área da Saúde, raramente são apresentados exemplos de um fator. Para os modelos do ICC de dois fatores, distinções como concordância absoluta e consistência, como unidade de medida individual ou em média foram apresentadas e discutidas num contexto inter-avaliador e intra-avaliador.

Para cada exemplo selecionado (capítulo 7) foi produzido um código R para o cálculo da estimação dos coeficientes e respetiva inferência estatística. Estes códigos estão disponíveis no final desta dissertação.

Nos próximos parágrafos, apresentar-se-ão as principais conclusões por capítulo.

No capítulo 1 são apresentados conceitos relacionados com validade. Neste capítulo foram apresentados resumidamente os métodos estatísticos associados a cada conceito de validade. A principal conclusão sobre este capítulo é que a fiabilidade não implica validade de um instrumento ou medição. A falta de fiabilidade coloca problemas sobre a validade de um teste e portanto um teste que não seja fiável não pode ser válido.

No capítulo 2 são apresentados os coeficientes de concordância mais usuais para variáveis nominais para dois avaliadores (como o kappa de Cohen, Scott, e do Brennan-Prediger) ou mais do que dois avaliadores (Kappa de Fleiss, Conger, e do Brennan-Prediger). Estes métodos procuram medir a concordância, tendo em consideração a proporção de acordo devido ao acaso. A conclusão principal é que estes coeficientes apresentam variações para o cálculo da proporção esperada de acordo devido ao acaso (Pe) através das diferentes formas de utilização das probabilidades marginais. Além disso, todos eles apresentam comportamentos

anómalos (paradoxais) em determinadas situações o que levanta algumas questões na sua aplicação generalizada. Estes paradoxos são também apresentados e ilustrados.

No capítulo 3 são apresentados os coeficientes de concordância mais usuais para variáveis ordinais com poucas categorias para dois ou mais avaliadores, através da apresentação dos kappas ponderados. Estes coeficiente são baseados nos coeficientes apresentados no capítulo 2 com a inclusão de ponderações que procuram ter em conta a natureza ordinal das variáveis, atribuindo assim pesos diferentes às discordância, dependendo do afastamento entre categorias. A conclusão mais interessante é que a inclusão da ponderação permite obter resultados mais corretos sobre a medição realizada pelos avaliadores. Um estudo sobre valores em falta para dois avaliadores é também apresentado.

No capítulo 4 são apresentados os coeficientes de fiabilidade mais usuais para variáveis quantitativas para dois ou mais avaliadores, numa situação inter-avaliador e intra-avaliador sem réplicas, que são os três modelos existentes do coeficiente de correlação intraclasse. Uma clara distinção em como utilizar estes modelos, bem como a sua detalhada formulação matemática são os principais resultados deste capítulo.

No capítulo 5 são apresentados os coeficientes de fiabilidade mais usuais para variáveis quantitativas para dois ou mais avaliadores, mas com réplicas, novamente para os modelos de dois fatores existentes do ICC. Como no capítulo anterior, a clara forma de utilização destes modelos, bem como a sua detalhada formulação matemática são os principais resultados deste capítulo.

No capítulo 6 são apresentados os coeficientes de fiabilidade/concordância mais usuais para variáveis classificadas em rankings para dois ou mais avaliadores, numa situação inter-avaliador. A aplicabilidade destes métodos numa situação intra-avaliador contínua em análise.

No capítulo 7 é apresentado um estudo de inferência estatística relativo a todos os métodos apresentados nos capítulos anteriores. Resultados sobre o cálculo amostral para os métodos mais usuais também são apresentados. Todos os resultados

apresentados, a sua validação e o respetivo código R são as principais conclusões deste capítulo.

Desta forma, esta dissertação vem agrupar e completar muita da informação disponível na literatura, constituindo um contributo para uma mais correta aplicação destes métodos de fiabilidade e concordância na construção ou adaptação de instrumentos de medida.

8.3 Trabalho futuro

Apesar da extensão desta dissertação, apresentamos algumas pistas de como tornar este trabalho mais completo. A primeira está relacionada com a identificação dos coeficientes de concordância para variáveis nominais ou ordinais para uma situação intra-avaliador. Para variáveis quantitativas, o modelo de um fator para a análise intra-avaliador não é totalmente esclarecedor. Uma análise do cálculo da dimensão da amostra para o Kappa de Fleiss ou para kappa de Brennan-Prediger será muito útil devido á sua grande utilização na investigação atual.

Outra linha de investigação será a relação entres os kappas ponderados e os modelos do ICC e inclusão dos métodos propostos por Bland and Altman neste contexto. Os coeficientes de concordância propostos por Lin (1989) para dois avaliadores e por Barnhart et al. (2002) para mais do que dois avaliadores, também deverão ser analisados como uma alternativa aos métodos de correlação intra-classe.

Por último, métodos para avaliar a consistência interna associado a questionários, como o alpha de Cronbach (para variáveis quantitativas) ou Kuder–Richardson Formula 20 (KR-20 para variáveis qualitativas binárias) devem ser outros passos a seguir.

Referências

Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hroóbjartsson, A., … Streiner, D. L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *International Journal of Nursing Studies*, *48*(6), 661–671. http://doi.org/10.1016/j.ijnurstu.2011.01.016

McGraw, K. O., & Wong, S. P. (1996). “Forming inferences about some intraclass correlations coefficients”: Correction. *Psychological Methods*, *1*(4), 390–390. http://doi.org/10.1037/1082-989X.1.4.390

Vet, H., Terwee, C., Knol, D., & Bouter, L. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology*, *59*(10), 1033–9. http://doi.org/10.1016/j.jclinepi.2005.10.015

Lin LI. (1989). A concordance correlation coefficient to evaluate reproducibility. Biometrics; 45: 225–268.

Barnhart, H.X., Haber M. & Song J. (2002) Overall concordance correlation coefficient for evaluating agreement among multiple observers. Biometrics; 58: 1020–1027.

Apêndice A- Códigos utilizados nos exemplos do capítulo 7

Neste apêndice serão apresentados os códigos produzidos utilizando o software RStudio (https://www.rstudio.com, versão 0.99.896) e R (https://www.r-project.org, versão 3.2.3) , utilizando os packages IRR e Ipsolve e as rotinas implementadas pelo autor GWET (Gwet, 2010).

Códigos utilizados para os cálculos apresentados na tabela 7.1 do capitulo 7.

```
#Usando as funções desenvolvidas por Gwet
#Cria uma matriz de 2X2
ratings<-matrix(c(35, 20, 5, 40),ncol=2,byrow=TRUE)
# define pesos lineares
weights=diag(ncol(ratings))
# Calcular o Kappa de Cohen
kappa2.table(ratings, weights, conflev=0.95, N=Inf,print=TRUE)
# calcular o kappa de Scott
scott2.table(ratings, weights, conflev=0.95, N=Inf,print=TRUE)
#calcular o kappa de Brennan-Prediger
bp2.table(ratings, weights, conflev=0.95, N=Inf,print=TRUE)
#calcular o kappa de Gwet
gwet.ac1.table(ratings, weights, conflev=0.95, N=Inf,print=TRUE)
```

Códigos utilizados para os cálculos apresentados na tabela 7.2 do capitulo 7.

```
#Exemplo 2.2

#Usando as funções desenvolvidas por Gwet

#Cria uma matriz de 3x3

ratings<-matrix(c(31, 1,2,3,37,4,2,1,21),ncol=3,byrow=TRUE)

# define pesos lineares

weights=diag(ncol(ratings))

# Calcular o Kappa de Cohen

kappa2.table(ratings, weights, conflev=0.95, N=Inf,print=TRUE)

# calcular o kappa de Scott

scott2.table(ratings, weights, conflev=0.95, N=Inf,print=TRUE)

#calcular o kappa de Brennan-Prediger

bp2.table(ratings, weights, conflev=0.95, N=Inf,print=TRUE)

#calcular o kappa de Gwet

gwet.ac1.table(ratings, weights, conflev=0.95, N=Inf,print=TRUE)
```

Códigos utilizados para os cálculos apresentados na tabela 7.3 do capitulo 7.

```
#Exemplo 2.3

#Usando as funções desenvolvidas por Gwet

#Cria uma matriz de 12X5

ratings<-matrix(c(3,1,0,0,0,0,3,1,0,0,0,0,4,0,0,0,0,4,0,0,0,4,0,0,0,1,1,1,1,0,

        +0,0,0,4,0,3,1,0,0,0,0,4,0,0,0,

        0,0,0,0,4,2,0,0,0,2,0,3,1,0,0),ncol=5,byrow=TRUE)

# calcular o kappa de  fleiss
```

```
fleiss.kappa.dist(ratings)

#calcular o kappa de Brennan-Prediger

bp.coeff.dist(ratings)

 #calcular o kappa de Gwet

gwet.ac1.dist(ratings)


# Cria matriz 12x4 (para calcular o kappa de conger sujeitos x avaliadores)

ratings<-matrix(c(0,0,1,0,1,1,2,1,2,2,2,2,2,2,2,2,1,1,1,1,0,1,
                  2,3,3,3,3,3,0,0,1,0,1,1,1,1,4,4,4,4,4,4,0,0,1,1,2,1),ncol=4,byrow=T)

#calcular o kappa de Conger

conger.kappa.raw (ratings)


# calcular o kappa de  fleiss para confirmar o anterior

fleiss.kappa.raw(ratings)
```

Códigos utilizados para os cálculos apresentados na tabela 7.4 do capitulo 7.

```
# Exemplo 3.1

#Usando as funções desenvolvidas por Gwet

#Cria uma matriz de 3x3

ratings<-matrix(c(2,2,0,1,3,1,0,0,2),ncol=3,byrow=T)

# kappa de Cohen ponderado

kappa2.table(ratings)

# kappa de Cohen com pesos lineares

kappa2.table(ratings,linear.weights(1:3))

# kappa de Cohen com pesos quadraticos

kappa2.table(ratings,quadratic.weights(1:3))

#calcular o kappa de Brennan-Predigerccom pesos lineares

bp2.table(ratings)
```

```
#calcular o kappa de Brennan-Predigerccom pesos lineares

bp2.table(ratings,linear.weights(1:3))

#calcular o kappa de Brennan-Prediger com pesos quadráticos

bp2.table(ratings,quadratic.weights(1:3))
```

Códigos utilizados para os cálculos apresentados na tabela 7.5 do capitulo 7.

```
#Exemplo 3.2

#Cria uma matriz de 3x3

ratings1<-
matrix(c(1,1.5,1,NA,2,2,2,2,0.5,1,1.5,1.5,1,1,1,1,1,1,1.5,NA,1,2.5,NA,2.5,2.5,2.5,2.5,1,1,NA,1,
NA,1,2,1,1,1,0.5,1,1.5,1.5,1.5,1.5,1,1.5,1,NA ,1,1,1.5, NA,1,2,2.5,2,
NA,1,1.5,1,0.5,0.5,0.5,0.5),ncol=4,byrow=T)

#calcular o kappa de Fleiss não ponderado

fleiss.kappa.raw(ratings1,weights="unweighted",conflev=0.95,N=Inf,print=TRUE)

#calcular o kappa de Fleiss com pesos lineares

fleiss.kappa.raw(ratings1,weights="linear",conflev=0.95,N=Inf,print=TRUE)

#calcular o kappa de Fleiss com pesos quadraticos

fleiss.kappa.raw(ratings1,weights="quadratic",conflev=0.95,N=Inf,print=TRUE)

#calcular o kappa de Brennan-Prediger não ponderado

bp.coeff.raw(ratings1,weights="unweighted",conflev=0.95,N=Inf,print=TRUE)

#calcular o kappa de Brennan-Prediger com pesos lineares

bp.coeff.raw(ratings1,weights="linear",conflev=0.95,N=Inf,print=TRUE)

#calcular o kappa de Brennan-Prediger com pesos quadráticos

bp.coeff.raw(ratings1,weights="quadratic",conflev=0.95,N=Inf,print=TRUE)

ratings2<-matrix(c(22,10,2,3,6,27,11,2,2,5,17,3,3,1,6,0),ncol=4,byrow=T)

kappa2.table(ratings2,weights=diag(ncol(ratings2)),conflev=0.95,N=Inf,print=TRUE)
```

Códigos utilizados para os cálculos apresentados na tabela 7.6 e 7.7 do capitulo 7.

```
#exemplos para o calculo da dimensão da amostra para o kappa de Cohen( Tabela 7.6 e 7.7)

#k1=0.1

N.cohen.kappa(0.6, 0.5, 0.1, 0, alpha=0.05, power=0.8, twosided=TRUE)

#k1=0.3

N.cohen.kappa(0.6, 0.5, 0.3, 0, alpha=0.05, power=0.8, twosided=TRUE)

#k1=0.6

N.cohen.kappa(0.6, 0.5, 0.6, 0, alpha=0.05, power=0.8, twosided=TRUE)

#k1=0.9

N.cohen.kappa(0.6, 0.5, 0.9, 0, alpha=0.05, power=0.8, twosided=TRUE)

#k1=0.1

N2.cohen.kappa(c(0.31,0.45,0.24), 0.1, 0, alpha=0.05, power=0.8, twosided=TRUE)

#k1=0.3

N2.cohen.kappa(c(0.31,0.45,0.24), 0.3, 0, alpha=0.05, power=0.8, twosided=TRUE)

#k1=0.6

N2.cohen.kappa(c(0.31,0.45,0.24), 0.6, 0, alpha=0.05, power=0.8, twosided=TRUE)

#k1=0.9

N2.cohen.kappa(c(0.31,0.45,0.24), 0.9, 0, alpha=0.05, power=0.8, twosided=TRUE).
```

Códigos utilizados para os cálculos apresentados na tabela 7.8 do capitulo 7.

```
#Exemplo 4.1

pontuacao<-matrix(c(

 9,2,5,8,6,1,3,2,8,4,6,8,7,1,2,6,

 10,5,6,9,6,2,4,7), ncol=4,byrow=TRUE)
```

```
#ICC(1,1):

icc(pontuacao, model="oneway",  unit = "single", r0 = 0.0, conf.level = 0.95)

icc(pontuacao, model="oneway",type="agreement",  unit = "single", r0 = 0.0, conf.level = 0.95)

#ICC(1,4)

icc(pontuacao, model="oneway", type="agreement", unit = "average", r0 = 0.0, conf.level =
0.95)

#ICCa(2,1): agreement

icc(pontuacao, model="twoway", type="agreement", unit = "single", r0 = 0.0, conf.level = 0.95)

#ICCc(2,1): consistency

icc(pontuacao, model="twoway", type="consistency", unit = "single", r0 = 0.0, conf.level =
0.95)

#ICCc(2,4): consistency

icc(pontuacao, model="twoway", type="consistency", unit = "average", r0 = 0.0, conf.level =
0.95)

#ICCa(2,4): agreement

icc(pontuacao, model="twoway", type="agreement", unit = "average", r0 = 0.0, conf.level =
0.95)
```

Códigos utilizados para os cálculos apresentados na tabela 7.9 do capitulo 7.

```
# calculo da dimensão da amostra para p=0.80, p=0.60, Two ratings, alpha=0.05 bilateral e
potência 0.80

calculateIccSampleSize(p=0.01,p0=0.0,k=4,alpha=0.05,tails=2,power=0.80).
```

Códigos utilizados para os cálculos apresentados nas tabelas 7.10 e 7.11 do capitulo 7.

```
# Fiabilidade inter e intra avaliador para o desenho de medidas repetidas do exemplo 5.1
table4<-matrix(c(
  -2,16,5,11,7,-7,18,4,0,0,-3,3,7,-6,1,-13,2,4,-10,8,7,-3,-5,5,0,7,-8,1,-3,
  0,16,6,10,8,-8,19,5,-3,0,-2,-1,9,-7,1,-14,1,4,-9,9,6,-2,-5,5,-1,6,-8,1,-3,
  1,15,6,10,6,-8,19,5,-2,-2,-2,1,9,-6,0,-14,0,3,-10,8,7,-4,-7,5,-1,6,-8,2,-3,
  2,12,4,9,5,-9,17,5,-7,1,-4,-1,4,-8,-2,-12,-1,7,-10,2,8,-5,-6,3,-4,4,-10,1,-5,
  1,14,4,7,6,-10,17,5,-6,2,-3,-2,4,-10,-2,-12,0,6,-11,8,7,-5,-8,4,-3,4,-11,-1,-4,
  1,13,4,8,6,-9,17,5,-5,1,-3,1,2,-9,-3,-12,0,4,-10,8,7,-5,-7,4,-4,4,-10,0,-5 ),ncol=6)
# caso em que inter rho=0.6 e intra rho=0.8
relInterIntra(x=table4, nrater=2, raterLabels=c('goniometro1','goniometro2'), rho0inter=0.6,
        rho0intra=0.8, conf.level=.95)
```

Códigos utilizados para os cálculos apresentados na tabela 7.11 do capitulo 7.

```
# caso em que inter rho=0.0 e intra rho=0.0
relInterIntra(x=table4, nrater=2, raterLabels=c('goniometro1','goniometro2'), rho0inter=0.0,
        rho0intra=0.0, conf.level=.95)
```

Códigos utilizados para os cálculos apresentados na tabela 7.12 do capitulo 7

```
#Exemplo 6.3(com empates)
#Sperman
x<-c(79.8,65,79.8,65,79.8,64,64.3,61)
y<-c(78,65.2,79,63,78,67,65.1,60)
```

```
cor.test(x, y,
        alternative = "two.sided",
        method = "spearman",exact=FALSE,
        conf.level = 0.95, continuity = FALSE)

#Kendall tau
cor.test(x, y,
        alternative = "two.sided",
        method ="kendall",exact=FALSE,
        conf.level = 0.95, continuity = FALSE)

#Kendall W
ratings<-matrix(c(79.8,78,65,65.2,79.8,79,65,63,79.8,78,64,67,64.3,65.1,61,60)
, ncol=2,byrow=TRUE)
kendall(ratings, correct = T)
```