MÁRIO LUIZ PINTO FERREIRA

MENSURAÇÃO DA VARIAÇÃO EM SAÚDE, POR ESCORES ORDINAIS:

ASPECTOS TÉCNICOS E PRÁTICOS E PROPOSTA DE UM NOVO INDICADOR.

DISSERTAÇÃO DE MESTRADO APRESENTADA AO CORPO DOCENTE DO INSTITUTO DE ESTUDOS DA SAÚDE COLETIVA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA À OBTENÇÃO DO GRAU DE MESTRE EM SAÚDE COLETIVA.

ORIENTADOR: Prof. Dr. Ronir Raggio Luiz

RIO DE JANEIRO

2007

FICHA CATALOGRÁFICA

Ferreira, Mário Luiz Pinto.
Mensuração da Variação em Saúde, por Escores Ordinais: Aspectos Técnicos e Práticos e Proposta de um Novo Indicador / Mário Luiz Pinto Ferreira. – 2007.
62 f.: il.

Dissertação (Mestrado em Saúde Coletiva) – Universidade Federal do Rio de Janeiro, Faculdade de Medicina, Centro de Ciências da Saúde, Instituto de Estudos da Saúde Coletiva, Rio de Janeiro, 2007.

Orientador: Ronir Raggio Luiz

1. Mensuração da Variação. 2. Dados Ordinais. 3. *Responsiveness* (Sensibilidade). 4. Indicadores. 5. Melhora. I. Luiz, Ronir Raggio (Orient.). II. Universidade Federal do Rio de Janeiro. Instituto de Estudos da Saúde Coletiva. Curso Mestrado em Saúde Coletiva. III. Título.

MÁRIO LUIZ PINTO FERREIRA

MENSURAÇÃO DA VARIAÇÃO EM SAÚDE, POR ESCORES ORDINAIS:

ASPECTOS TÉCNICOS E PRÁTICOS E PROPOSTA DE UM NOVO INDICADOR.

DISSERTAÇÃO DE MESTRADO APRESENTADA AO CORPO DOCENTE DO INSTITUTO DE ESTUDOS DA SAÚDE COLETIVA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA À OBTENÇÃO DO GRAU DE MESTRE EM SAÚDE COLETIVA.

APROVADA EM 7 DE MARÇO DE 2007

Prof. Dr. Ronir Raggio Luiz – IESC/UFRJ

Prof. Dr. Antonio José Leal Costa – IESC/UFRJ

Prof. Dr. Renan Moritz V. R. Almeida – COPPE/UFRJ

Prof. Guillermo Coca Velarde - UFF

AGRADECIMENTOS

Ao meu orientador Ronir Raggio, pelo aprendizado obtido e apoio na execução deste trabalho.

A todos os professores do IESC pelo aprendizado transmitido durante as aulas.

Aos colegas do HCII, em especial ao Dr. Rondinelli, pela compreensão dos vários momentos de ausência.

Em especial a minha esposa Rita e filhos Davi e Tati, pelo apoio em todos os momentos.

RESUMO

A mensuração da variação com variáveis ordinais, de uma mesma unidade, entre dois momentos ao longo do tempo (antes e depois), no contexto da saúde, é um tema de muita controvérsia. Os pontos principais desta discussão são: a) a subjetividade (imprecisão) do dado categórico ordinal e b) o tratamento estatístico mais adequado a ser usado nas análises dos resultados.

Neste trabalho ressaltam-se os aspectos técnicos e práticos, que afetam diretamente a forma de mensuração da variação e a interpretação de seus resultados decorrente do tratamento estatístico aplicado. Além disso, é apresentado um novo indicador para ser aplicado na mensuração da variação com dados categóricos ordinais.

PALAVRAS-CHAVE: MENSURAÇAO DA VARIAÇAO, DADOS ORDINAIS, INDICADORES, MELHORA.

ABSTRACT

The measurement of the variation with ordinal variables, of a same unit, between two moments along the time before and later, in the context of health science, is a controversial theme. The main points of this discussion are: the subjectivity of the ordinal categorical data and the appropriate statistical treatment to be used in the analyses of the results.

This lecture emphasizes technical and practical aspects that affect the form to measure the variation and also affect the interpretation of the results directly due to the applied statistical treatment. Besides of that this work also presents a new indicator to be applied in the measurement of the variation with ordinal categorical data.

KEYWORDS: ORDERED CATEGORICAL-DATA, INDEXES, MEASURE OF CHANGE.

LISTA DE ILUSTRAÇÕES (QUADROS E FIGURAS)

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Com alguma freqüência, a subjetividade de certas variáveis usadas na avaliação de uma intervenção tende a falsamente sugerir sua impossibilidade de ser medida. Atribui-se a Galileu a afirmação de que se deve medir o mensurável e transformar em mensurável o que, à primeira vista, não for. Uma tradução livre de Bauman (1980) explicita bem a questão da mensuração de qualquer variável.

> Freqüentemente é dito que algumas variáveis não podem ser medidas [...] Se isto fosse verdade, então muitas questões importantes sobre a saúde e bem estar de comunidades não poderiam ser investigadas. Felizmente, todas as variáveis podem ser consideradas mensuráveis quando isto é reconhecido como um processo pelo qual pessoas as definem e descrevem. Isto é, variáveis existem como um resultado de suas próprias definições e descrições (apud Pereira, 2004, p.29).

Ante o exposto, medir a variabilidade de algum fenômeno é uma etapa fundamental na pesquisa científica, seja ela natural, social ou da saúde e, também, utilizada na investigação dos resultados de políticas e programas de saúde. Por meio desta prática, torna-se possível conhecer, por exemplo: a) as conseqüências decorrentes da aplicação, experimental ou não, de novas drogas ou intervenções clínicas e cirúrgicas; b) o quanto cada paciente se beneficiou após a realização de uma intervenção; c) resultados alcançados por diferentes programas sociais ou de saúde pública. Diante destas necessidades, o método mais utilizado na atividade de avaliação é a mensuração por meio de indicadores. Essa estratégia visa a demonstrar empiricamente os resultados alcançados decorrentes, por exemplo, de ações (intervenções) desenvolvidas por instituições de saúde ou por gestores de políticas de saúde.

Muitos autores entendem que a principal meta de qualquer intervenção é induzir um efeito no cuidado à saúde de indivíduos[1] ou de grupos. Assim, medir o padrão de

[1] O termo indivíduos deve ser entendido também como uma única unidade observada, como por exemplo: a) uma pessoa; b) um agregado ou coletivo; c) uma organização.

variabilidade longitudinal[2] deve ser encarado como uma das etapas do delineamento da investigação. Dessa forma, os indicadores utilizados na mensuração da variação[3] do indivíduo ou grupo, ao longo do tempo, são medidas mais sensíveis dos efeitos das intervenções do que mensurações que simplesmente avaliam o estado de saúde dos indivíduos ou grupo ao final da investigação.

A forma de mensuração permite delinear três tipos de investigação: a) conhecer diferenças nas quantidades de variação entre indivíduos, ou seja, quem variou mais e quem variou menos; b) identificar que outros fatores, além da intervenção, influenciaram na quantidade da variação; c) inferir efeitos da intervenção, a partir das diferenças encontradas nos grupos (Streiner e Norman, 2005).

Apesar de existir, segundo Streiner e Norman (2005), um estreito relacionamento entre a forma de mensuração e o delineamento da investigação, as práticas utilizadas nessa mensuração se apresentam como tema de muito debate e de pouco consenso. Um ponto de discussão é sobre o que seria um indicador sensível para detectar variação quando de fato ela ocorre e como essa sensibilidade deveria ser abordada. Outro ponto desse debate refere-se ao o uso de escores de variação (escores da diferença), obtidos aritmeticamente pela diferença entre mensurações pareadas, antes e depois, de um mesmo indivíduo, como um indicador bruto de variação (Cronbach e Furby, 1970).

Independentemente deste debate, quaisquer que sejam os resultados de mensuração eles freqüentemente exigirão um alto grau de precisão e validade, tanto para variáveis ainda pouco estudadas, como para aquelas já bem estabelecidas. Assim, profissionais que trabalham na atenção à saúde (pesquisadores, gestores públicos,

[2] O termo longitudinal caracteriza-se por observações repetidas de unidades em pontos (dois ou mais) no tempo. Estes pontos podem ser anos, meses, semanas, dias ou horas.
[3] A expressão mensuração da variação é uma tradução livre para o termo em inglês *measurement of change*.

políticos, diretores de unidades de saúde, entre outros) precisam colocar em prática, além da forma de mensuração apropriada para a avaliação, métodos que possam garantir a validade e a confiabilidade dos resultados de suas ações.

Com respeito à precisão e à validade dos resultados obtidos pelas intervenções, há uma discussão importante entre estatísticos, matemáticos e psicólogos no que se refere à mensuração e à análise da variação quando os dados são do tipo ordinal (Gaito, 1980; Merbitz e col. 1989; Knapp, 1990; Velleman e Wilkinson, 1993; Kampen e Swyngedouw, 2000). A partir do momento em que Stevens (1946) definiu a tipologia das variáveis como nominal, ordinal, intervalar e razão, a discussão sobre o método mais adequado para variáveis ordinais persiste entre os especialistas. Os temas mais relevantes são: a subjetividade da escala e o uso freqüente de análises paramétricas para variáveis mensuradas de forma categórica ordinal. Como exemplos de variáveis ordinais têm-se: grau de dor, qualidade de vida e estado funcional.

O volume de trabalhos que utilizam variáveis ordinais tem sido significativo. Por exemplo, aproximadamente 20% (32 de 168) dos artigos publicados no volume 36 do periódico *New England Journal of Medicine* incluíam variáveis ordinais em suas análises (Kampen e Swyngedouw, 2000). Com o objetivo de avaliar o uso correto dos métodos estatísticos utilizados em periódicos de Reumatologia, Lavalley e Felson (2002) investigaram 644 artigos de três periódicos: *Arthrits & Rheumatism - A&R* (n = 282), *Journal of Rheumatology - JR* (n = 322) e *Arthrits Care & Research - AC&R* (n = 40), todos em 1999. Os autores identificaram que 27,2% dos textos tinham variáveis ordinais como base das mensurações.

Este trabalho tem como objetivo discutir os métodos utilizados na mensuração da variação com variáveis ordinais, de uma mesma unidade, entre dois momentos ao longo do tempo (antes e depois), no contexto da saúde, explorando suas características e

controvérsias, e propor um indicador que permita que os resultados da mensuração da variabilidade sejam facilmente calculados e interpretados.

2 MENSURAÇÃO NO CONTEXTO DA SAÚDE

A mensuração no contexto da saúde trabalha basicamente com dois tipos de informação. O primeiro tipo é relacionado às características dos indivíduos, como raça, grau de dor, qualidade de vida, e, o segundo, evidenciado pelas medidas do tipo miligramas e milímetros. Como estabelecido por Stevens (1946), de acordo com as suas escalas de mensuração, os primeiros podem ser classificados como nominal ou ordinal e os segundos como intervalar ou razão.

A mensuração também permite aos profissionais conhecerem estatísticas vitais (nascimentos e mortes), prevalências de uma doença, sua variação e localização geográfica, avaliarem informações sobre os fármacos e equipamentos médicos disponíveis no mercado, identificarem os procedimentos diagnósticos mais adequados a serem utilizados no atendimento ao paciente, entre outras utilidades.

2.1 CONCEITOS BÁSICOS DE INDICADOR

Indicadores são formas de mensurar os resultados de objetivos definidos pela prática clínica diária ou por políticas e programas de saúde. Tradicionalmente, a interpretação de resultados permite estabelecer critérios para orientar a tomada de decisão. Exemplos de decisões: a) definição de recursos para políticas e programas; b) estabelecimento de diretrizes para os tipos de atenção que devem ser prestados (primário, secundário e terciário); c) divulgação aos usuários sobre o desempenho das instituições de saúde (Mainz, 2003).

Antes de apresentar a proposta do indicador, mencionado anteriormente, é necessário estabelecer diferenças entre duas palavras comumente utilizadas sem muito rigor, indicador e índice. A Organização para Cooperação e Desenvolvimento Econômico - OCDE (OCDE, 1999) considera indicador como um parâmetro, ou valor

derivado de um parâmetro, que fornece as informações sobre um fenômeno. Um indicador, por definição, deve ser capaz de alertar a ocorrência de fenômenos que necessitem de atenção e que evidenciem tendências no sentido de demonstrar níveis aceitáveis ou inaceitáveis de desempenho. O uso de indicadores está associado a sua capacidade de focar em análises de desempenho de organizações ou indivíduos.

O termo “indicador” pode ser entendido, também, como uma ferramenta para acompanhar e avaliar desempenho de processos clínicos, administrativos, científicos, entre outros. No contexto geográfico podem ser divididos em nacional, regional e local e, no contexto organizacional podem ser estratégicos, gerenciais e operacionais (Mant, 2001).

Na definição de indicadores, algumas características devem ser atendidas: a) precisa existir consenso e estabilidade, isto é, consenso da equipe, e a fórmula de medição não deve ser alterada frequentemente; b) deve permitir a realização de comparações; c) deve estar baseado em evidência; d) deve ser um bom discriminante; e) deve captar claramente os eventos de interesse; e f) deve possuir validade e confiabilidade (Mainz, 2003).

A definição de “índice” consiste em um grupo de indicadores que visa a sintetizar dados em uma única expressão. Ele pode resumir informações de um ou mais fenômenos e traz consigo um julgamento implícito de valor. Por exemplo, o Índice de Desenvolvimento Humano - IDH retrata as condições de vida de uma população, residentes em determinada região (Kazandjian e Vallejo, 2004). Uma dificuldade em adotar índices é o fato de que dados podem ser perdidos ou escondidos em razão da agregação de informações (Granados e Peterson, 1999).

Assim, enquanto um indicador é baseado em informações fortemente relacionadas ao objeto de investigação, um índice utiliza diversos indicadores agregados para evidenciar um fenômeno.

2.2 UTILIZAÇÃO DE INDICADORES NA MENSURAÇÃO EM SAÚDE

A prática do uso de indicadores, por parte de profissionais de saúde, gestores e instituições permite a monitorização e avaliação do que acontece com a atenção à saúde que é oferecida ao indivíduo ou grupo diante da sua real necessidade, além de evidenciar o quanto a equipe e o sistema de saúde foram eficazes durante a atenção prestada.

As estratégias de avaliação e monitoramento realizadas por meio de indicadores podem ser aplicadas para diversas finalidades (Mainz, 2003):

- subsidiar a avaliação dos serviços de saúde, os processos de qualidade e acreditação;
- subsidiar a tomada de decisão através da definição de prioridades;
- subsidiar as regulamentações do sistema de saúde;
- subsidiar na seleção dos melhores fornecedores de materiais de consumo;
- descrever o nível de saúde de indivíduos e populações;
- conhecer o nível de satisfação do paciente com o cuidado prestado;
- conhecer os custos diretos e indiretos do sistema de saúde;
- documentar a evolução da qualidade do cuidado;
- realizar estudos de comparações de desempenho com referenciais de excelência.

Como exemplos de indicadores, tipicamente encontrados no contexto da saúde, têm-se: a) taxa de infecção urinária; b) número de pacientes que tiveram queda do leito; c) percentual de rotinas revisadas em um determinado período; d) proporção de

pacientes tratados segundo o protocolo clínico; e) mortalidade materna e infantil; f) grau de morbidade hospitalar; g) prevalência da desnutrição; h) nível da qualidade de vida e grau de satisfação do paciente; i) proporção de pacientes com câncer de mama que estão vivos pelo menos três anos após o início do tratamento.

Quando o interesse é avaliar programas sociais desenvolvidos, no âmbito da saúde, tais como: a) Programa de Agentes Comunitários; b) Programa Nacional de Imunização; c) Programa de Atenção Integral à Saúde da Mulher, da Criança e do Adolescente; d) Programa de Combate às Carências Nutricionais, os indicadores procuram traduzir o desempenho das atividades desenvolvidas. Para o caso de um Programa de Imunização tem-se, por exemplo:

- proporção de pessoas imunizadas;
- proporção de pessoas imunizadas expostas após a imunização e que não desenvolveram a doença;
- proporção de pessoas sob alto risco que foram imunizadas.

Resultados pelos indicadores sofrem influência de fontes de variabilidade, tais como: diferenças nas características dos pacientes que são tratados por diferentes instituições; diferenças entre grupos populacionais; diferenças nas estratégias de coleta das informações e diferenças observadas por mero acaso (Mant, 2001). Diante da diversidade de fontes de variação, torna-se relevante e evidente a avaliação dos indicadores quanto ao atendimento às suas propriedades, visto que os resultados obtidos a partir de técnicas e as análises estatísticas, por mais robustas que sejam somente serão dignos de confiança se os dados de origem forem confiáveis.

2.3 GARANTIA DA QUALIDADE DO DESEMPENHO DOS INDICADORES.

A determinação das propriedades de um indicador reflete a credibilidade da mensuração na avaliação em saúde. A propriedade chamada de validade diz respeito à

capacidade da medida em mensurar o que ela se propõe a medir, mantendo-se fiel ao seu planejamento. Assim, a investigação da validade de uma medida torna-se dependente das características da população sobre a qual ela será utilizada (De Vet, 2003). Pode-se afirmar que, em função dos objetivos definidos para uma determinada intervenção, uma medida pode ser válida somente para essa intervenção e não para outra. Por exemplo, uma medida pode ser válida para certo grupo de pacientes ou para uma população em um ponto no tempo, mas ser ineficaz para medir variação longitudinal em alguns indivíduos ou em alguns grupos desta população. A propriedade conhecida como confiabilidade[4] refere-se a uma medida que sob as mesmas condições produz os mesmos resultados, quando aplicado duas ou mais vezes num mesmo dia ou quando aplicado em tempos distintos pelos mesmos ou por diferentes avaliadores ou em um mesmo momento, por avaliadores distintos.

Estas duas qualidades (validade e confiabilidade) também conhecidas como propriedades inerentes ao indicador, são utilizadas em larga escala, tornando-as extremamente relevantes do ponto de vista metodológico, e, sua avaliação necessária visando à confiança nos resultados.

Uma terceira qualidade, conhecida como sensibilidade[5] (*responsiveness*) é utilizada quando o objetivo da investigação é avaliar a magnitude da variação ao longo de um determinado tempo, de determinadas características dos indivíduos ou de populações, como conseqüência da aplicação de uma intervenção (Terwee e col., 2003). Em outros termos, a sensibilidade é entendida como a habilidade de uma mensuração detectar variações quando estas de fato ocorreram.

---

[4] Esta é uma tradução ampla e livre do termo reliability. Este termo pode ser encontrado, no contexto da saúde, como: objectivity, reproducibility, stability, agreement, association, sensitivity e precision.

[5] A tradução de *responsiveness* por "sensibilidade" pode gerar confusão, pois este termo tem sido também, utilizado no campo dos testes de diagnóstico, que avalia a proporção de testes verdadeiros positivos entre todos os indivíduos que se encontram doentes. Também pode ser encontrada na abordagem chamada análise de sensibilidade, que avalia quão forte deve ser o efeito de uma variável confundidora não-mensurada para explicar uma aparente associação.

2.4 PLANEJAMENTO DA INVESTIGAÇÃO E AS FUNÇÕES DOS INDICADORES

Segundo Kirshner e Guyatt (1985), indicadores de avaliação da atenção à saúde referente a indivíduos ou grupos são delineados na etapa de planejamento da investigação com base em três tipos de objetivos, sendo classificados como preditivo, discriminativo e avaliativo. Portanto, a avaliação do desempenho de um indicador depende diretamente dos objetivos definidos para o mesmo.

- Se a intenção é a predição, os indicadores buscam classificar indivíduos segundo um padrão-ouro. São geralmente utilizados como medidas de rastreamento ou prognósticas, no sentido de identificar indivíduos que possuem ou poderiam desenvolver determinada doença. A validade neste enfoque avalia se de fato o indicador é capaz de prever a doença. A confiabilidade para indicadores preditivos é caracterizada pela concordância entre observações repetidas, isto é, se o indicador classifica cada indivíduo na mesma categoria, mesmo quando avaliado em dois tempos diferentes, desde que não haja alteração do estado de saúde.
- Se a intenção é discriminativa, os indicadores buscam distinguir indivíduos ou grupos, em um único momento no tempo, quando um padrão-ouro não está disponível ou não existe um critério externo como referência a ser usado para validar as medidas. A validade na avaliação discriminativa pode ser desdobrada em duas partes: a validade de conteúdo e a validade de construto. A primeira busca demonstrar se o instrumento cobre todos os aspectos que são importantes na avaliação de um domínio, por exemplo, saúde física, mental, social e estado funcional. A segunda busca demonstrar a relação existente entre o indicador discriminativo e outros,

que são consistentes com uma hipótese teórica. Por exemplo, um pesquisador espera que pacientes obesos respondam com maior freqüência à falta de exercícios físicos. A confiabilidade pode ser verificada pelas diferenças observadas entre indivíduos, após a realização de um teste e re-teste. Este procedimento avalia se a variação entre indivíduos permanece constante.

- Se o objetivo é avaliativo, os indicadores são desenvolvidos para medir a magnitude da variação longitudinal em um indivíduo ou grupo. A qualidade do indicador que deve ser mensurada é a sensibilidade, cujo objetivo é avaliar a capacidade do indicador em detectar variação quando esta realmente ocorreu. A validade de conteúdo, neste enfoque, se preocupa se as áreas (dimensões ou variáveis) de interesse para a avaliação da variação longitudinal estão identificadas e, quando necessário, segmentadas, para sinalizar as variações clinicamente importantes da intervenção. Para a validade do construto, o que importa é demonstrar que variações longitudinais intra-indivíduos, decorrentes de uma intervenção, implicam em variações esperadas em outras variáveis semelhantes. A confiabilidade avalia se as mensurações repetidas (pelo menos duas) de cada indivíduo permanecem estáveis ao longo do tempo, ou seja, é esperado que a magnitude da variância intra-indivíduos seja pequena. Essa interpretação da confiabilidade é evidenciada quando o coeficiente de confiabilidade do escore de mudança (*reliability of change score*) é zero, isto é, todos os indivíduos variaram uniformemente e, portanto, não seria possível separar quais indivíduos variaram mais ou menos.

O Quadro 1 apresenta uma síntese das definições estabelecidas por Kirshner e Guyatt (1985).

Quadro 1. Definições de Kirshner e Guyatt.

	Critérios		
	Preditivo	Discriminativo	Avaliativo
Objetivo	Rastreamento ou diagnóstico de indivíduos doentes	Distinguir indivíduos em um único tempo	Conhecer a magnitude da variação
Validade	Avalia se o indicador prevê de fato a doença	De conteúdo, avalia se todos os aspectos são mensurados; de construto, avalia a relação entre o indicador e uma hipótese teórica.	De conteúdo, avalia se os aspectos de interesse para a variação são mensurados; de construto, avalia se as variações intra-indivíduos implicam em alterações em outras variáveis relacionadas.
Confiabilidade	Avalia a concordância entre as observações repetidas	Avalia se a variação existente entre indivíduos é pequena.	Avalia se a variação entre as observações repetidas, de um mesmo indivíduo, é pequena.
Sensibilidade	Não é relevante	Não é relevante	Avalia a capacidade de identificar a variação existente

3 MENSURAÇÃO DA VARIAÇÃO

3.1 A IMPORTÂNCIA DE MENSURAR VARIAÇÃO

Mensurar variação é um conceito fundamental dentro das atividades desenvolvidas por pesquisadores, gestores de sistemas de saúde e médicos, de forma geral. Para pesquisadores e gestores o importante é olhar o desempenho dos diferentes projetos ou programas de intervenção ou, ainda, a possibilidade de reduzir custos no desenvolvimento das ações de saúde, sem comprometer o estado de saúde da população assistida. O programa promoveu algum impacto na saúde do grupo? Quanto? Em um ambiente clínico, a atividade de mensurar a variação é traduzida pelas questões transmitidas aos pacientes sobre a variação observada durante consultas de seguimento em um ambulatório. A intervenção promoveu um bem-estar ao paciente? Quanto? O paciente melhorou ou piorou? Quanto? Que variações eram esperadas em pacientes tratados de uma determinada maneira? (Beaton, 2000).

3.2 ABORDAGENS SENSÍVEIS SEGUNDO AS DEFINIÇÕES DE VARIAÇÃO

Na busca de entendimento sobre a melhor forma de medir a variação, autores estabeleceram uma relação entre a definição de variação e as abordagens metodológicas a serem usadas.

Neste sentido, Plewis (1985) diferenciou três tipos de abordagens para orientar a mensuração da variação: a) identificar a variação intra-individual; b) conhecer a variação dentro de um grupo de indivíduos; c) identificar variação dentro de um grupo de indivíduos comparada com a variação dentro de outro grupo de indivíduos.

Conforme visto anteriormente, a abordagem denominada avaliativa, apresentada por Kirshner e Guyatt (1985) busca identificar a magnitude da variação ao longo do

tempo em indivíduos ou grupos sobre uma dimensão de interesse. A intenção dos autores era quantificar o benefício da intervenção em estudos clínicos randomizados.

Liang (2000) diferencia as abordagens entre sensibilidade e sensibilidade à mudança (*sensitivity to change*). A primeira (sensibilidade) identifica variações clinicamente importantes, e, a segunda (sensibilidade para mudança), avalia a possibilidade de qualquer nível de variação.

Husted e colaboradores (2000) ao revisarem a literatura, com objetivo de examinar a aplicação da propriedade denominada sensibilidade (*responsiveness*), observaram que existem dois grandes aspectos da sensibilidade, cada um com suas próprias definições e estratégias para avaliação. Eles definiram-na como sensibilidade interna, caracterizada pela avaliação de um indicador em um contexto de ensaios clínicos randomizados, envolvendo uma intervenção demonstrada previamente eficaz, ou por avaliação de medidas repetidas de um grupo antes (pré-teste) e depois (pós-teste). O segundo aspecto, chamado por eles de sensibilidade externa, caracteriza-se pelo relacionamento entre o indicador sensível à variação considerando um referencial externo, como por exemplo, um padrão-ouro.

Norman e colaboradores (2001) definem os tipos de variação como: "*distribution-based*" e "*anchor-based*". O primeiro tipo tem como objetivo identificar a variação entre o grupo tratamento e grupo controle para alguma medida sensível à variabilidade. O segundo (“*anchor-based"*) examina a relação entre uma medida sensível e uma outra independente, como por exemplo, um padrão-ouro, no sentido de demonstrar o significado de um grau de variação. Este padrão-ouro pode ser a opinião do paciente ou do médico sobre se ocorreu ou não melhora no estado de saúde do avaliado. A medida que mostra uma alta correlação com o padrão-ouro é considerada a mais sensível.

Demonstrando a dificuldade semântica intrínseca à questão, Terwee e colaboradores (2003) em estudo de revisão sobre mensuração da variação longitudinal encontraram vinte e cinco definições para a sensibilidade (*responsiveness*), categorizadas pelas habilidades em: a) detectar variação em geral; b) detectar variação clinicamente importante e, c) detectar variação no atributo medido. Tais definições estão sistematizadas no Quadro 2.

Quadro 2. Definições de sensibilidade.

A – Como habilidade de detectar variação em geral
A habilidade para detectar variação ao longo do tempo. A habilidade de um indicador medir variação clínica. A habilidade de uma medida detectar o efeito geral dos tratamentos. A habilidade de detectar uma variação devido ao tratamento. Quando uma medida detecta diferenças entre intervenções. A habilidade para detectar pequena variação intra-pacientes ao longo do tempo. O tamanho da amostra requerido para observar uma pequena, média ou grande variação ou o tamanho do efeito na população. O poder para detectar uma diferença quando uma está presente. O efeito do erro aleatório e sistemático sobre o poder de um teste.
B – Como habilidade de detectar variação clinicamente importante
A habilidade de uma medida em detectar importante variação ao longo do tempo. A habilidade de detectar variações clinicamente importantes no estado de saúde ao longo do tempo. A habilidade de detectar variação clínica estatisticamente significante. Quando variações estatisticamente significantes ocorrem em uma direção concorrente em indivíduos com alterações significantes dos sintomas. A habilidade para detectar variações clinicamente importantes no estado de saúde ao longo do tempo, mesmo se estas variações são pequenas. A habilidade para detectar variações pequenas, mas importantes. A habilidade de medir variações intra-indivíduos pequenas, mas clinicamente importante após efetiva terapêutica de intervenção. A habilidade para detectar a diferença mínima clinicamente importante. Detectando variações clinicamente importantes e diferenças entre tratamentos.
C – Como habilidade de detectar variação no atributo medido
O nível para o qual uma medida está apta em detectar variações na variável medida. Se variações no atributo estão refletidas nas variações dos valores da medida. A capacidade de detectar variações no estado de saúde em qualquer momento que ele exista. A habilidade de discriminar entre aqueles que melhoraram e aqueles que não. Uma medida de associação entre variações nos valores observados e variações no verdadeiro valor do construto. A habilidade de uma medida em avaliar variação em resposta a uma intervenção.

O critério utilizado na categorização foi baseado no tipo de variação que o indicador deveria estar apto a detectar. O primeiro grupo definido como a capacidade de

detectar variações em geral se aproxima muito da definição estabelecida por Liang (2000), denominada "*sensitivity to change"*. Neste grupo não importaria se a variação era relevante ou não. A definição para o segundo grupo foi a capacidade de detectar variações clinicamente importantes. A diferença mais relevante entre os dois grupos, já descritos, é o fato de que no segundo é necessário um ponto que sinalize variação significativa, embora o julgamento seja freqüentemente subjetivo. O terceiro grupo é definido como a capacidade de detectar uma variação real na dimensão (conjunto de atributos) em que o estudo está interessado em avaliar. Este grupo, além de necessitar de um ponto que caracterize variação significativa, exige também um padrão-ouro para comparação do resultado (Terwee e col., 2003).

Norma e Streiner (2005) argumentam que tipos de definição da variação direcionam o delineamento da investigação para três tipos de abordagens: a) avaliar quem variou mais ou menos; b) identificar fatores que contribuíram para a variação; e c) inferir efeitos da intervenção, a partir das diferenças encontradas nos grupos.

Ante o exposto, entende-se, baseado nos autores, que a abordagem mais adequada deve ser aquela que está intimamente correlacionada com a definição da variação descrita no delineamento da investigação.

3.3 INDICADORES SENSÍVEIS UTILIZADOS NA MENSURAÇÃO DA VARIABILIDADE

O estabelecimento de um indicador sensível de ser aplicado na mensuração da variação se tornou um tema de muita controvérsia e ainda não está consolidado no contexto da saúde. Provavelmente, a causa está relacionada à dificuldade de se obter um consenso entre os profissionais que trabalham no cuidado ao paciente.

Diversos indicadores surgiram em virtude da falta deste consenso sobre o que seria um indicador sensível em detectar variação e como ele deveria ser avaliado. Em

seu estudo de revisão, Terwee e col. (2003) apresentou para cada tipo de categoria de variação, descritos na tabela 4, um indicador para o cálculo da sensibilidade.

Mirjam e col. (2002) relataram, também, indicadores, mais comuns utilizados para avaliar a sensibilidade da variação. Estes indicadores estão listados abaixo em ordem cronológica:

1. Estatística t pareada, de Liang e col. (1985);
2. Tamanho do Efeito (*Effect Size – ES*), de Cohen (1988);
3. Média de Resposta Padronizada (*Standardized Response Mean*), de Liang e col. (1990);
4. Sensibilidade Estatística, de Guyatt e col. (1992).

A Estatística *t* pareada tem sido usada na avaliação de variação entre medidas repetidas em estudos com um único grupo. O indicador se utiliza da diferença entre as médias de duas avaliações distintas, no início ($\overline{X_1}$) e no final ($\overline{X_2}$), ou seja, verifica, em uma distribuição, onde se localiza a diferença entre os escores da linha de base (início da investigação) e os escores finais. Esta estratégia não é a mais adequada para avaliarmos a sensibilidade de uma medida, pois a influência do tamanho da amostra sobre a significância estatística, representada pelos valores de p, provoca conclusões diferentes sobre a sensibilidade (Terwee e col., 2003).

O Tamanho do Efeito (*Effect Size – ES*), apresentado por Cohen (apud Norman e Streiner, 2005, p. 199) fornece uma mensuração direta sobre a magnitude da variação, pois é calculado pela razão entre a média ($\overline{D_x}$) dos escores da diferença ($D_X = X_2 - X_1$) entre duas avaliações distintas no tempo, a da linha de base (X_1) e a do final (X_2), e o desvio padrão dos escores da linha de base ($SD\ (X_1)$). Dessa forma, podemos inferir que, se a medida possui alta variabilidade na linha de base, o tamanho do efeito será pequeno, e que o indicador torna-se sensível à homogeneidade da amostra.

Valores de referências têm sido sugeridos para facilitar a interpretação da variação observada considerada clinicamente importante. O valor de ***ES*** menor ou igual a 0,2 significa que a variação longitudinal representa um quinto do desvio padrão dos escores da linha de base e é considerado pequeno. Um valor de ***ES*** = 0,5 significa que a variação representa 50% do desvio padrão dos escores da linha de base. A obtenção de valores de ***ES*** iguais ou maiores que 0,8, significa que a variação representa quatro quintos do desvio padrão da linha de base e é considerada alta (Husted, 2000).

O indicador pode ser aplicado para avaliar a variação entre grupos (tratamento e controle), em um único indivíduo ou entre indivíduos. Embora estudos sobre escores de diferença tenham iniciado em 1962, por McNemar (apud Maassen, 2000, p. 187), este indicador foi publicado por Cohen, em 1988, no seu livro "*Statistical power analysis for the behavioral sciences*".

A média de resposta padronizada é definida como a razão entre a média ($\overline{D_x}$) dos escores da diferença ($D_X = X_2 - X_1$) de duas observações no tempo e o desvio padrão dos escores da diferença $SD(D_x)$. Da mesma maneira que o Tamanho do Efeito, se existir grande variabilidade nos escores da diferença, teremos valores pequenos desta medida. A vantagem que esta medida e o Tamanho do Efeito têm em relação a Estatística *t* Pareada é o fato de que elas não dependem do tamanho da amostra. Os valores de referências de 0,2, 0,5 e 0,8 são interpretados como pequena, média e alta sensibilidade, respectivamente (Husted, 2000).

A Sensibilidade Estatística proposta por Guyatt (1992), indicador 3, foi desenvolvida em virtude da falta de uma avaliação crítica, por parte dos responsáveis pelas investigações, da necessidade de se desenvolver novos indicadores para avaliar a prática clínica. A explicação para essa falta de avaliação é que o desenvolvimento de novos indicadores, freqüentemente, não está associado aos objetivos (preditivo,

discriminativo e avaliativo) estabelecidos no delineamento da investigação. O indicador está estruturado pela razão entre a variação mínima considerada clinicamente importante (Δ_x)[6] e o desvio padrão dos escores da diferença de individuos que foram considerados clinicamente estáveis. O numerador descreve a associação entre a diferença mínima importante dos escores na linha de base e no final como um benefício significativo para um grupo de pacientes. A abordagem usada no denominador busca ajustar a medida para variações espúrias que podem surgir devido ao erro de mensuração.

Leffondré e colaboradores (2004) realizaram um estudo no sentido de propor indicadores estatísticos simples de calcular e que permitissem identificar diferentes padrões de variação longitudinais, no contexto da saúde, classificados como: a) estáveis e não-estáveis; b) crescentes e decrescentes; c) lineares e não-lineares; d) com padrões de variação monotônico e não-monotônico. Eles escolheram vinte e sete indicadores e os classificaram em cinco categorias, a saber:

- estatística descritiva básica;
- indicadores de não-linearidade ou de inconsistência de variação;
- indicadores elementares de variação;
- indicadores que buscam alguma relação entre o escore da linha de base e o escore final.
- indicadores sensíveis para a não-monotonicidade e para diferenças bruscas no curto prazo

A estratégia denominada *Receiver Operating Characteristic (ROC)*, ou simplesmente curva ROC, também tem sido utilizada com o objetivo de distinguir os escores de variação dos indivíduos que melhoraram daqueles que não melhoraram. A

[6] A variação mínima clinicamente importante é definida como a menor diferença entre os escores da linha de base e os escores do final associados a um benefício para um grupo de pacientes (Husted e col., 2000).

metodologia utiliza a probabilidade da medida em classificar corretamente os pacientes que demonstraram melhora e a probabilidade da medida em classificar corretamente os pacientes que não mostraram melhora quando comparados a um referencial externo. A área da curva ROC representa a probabilidade de uma medida classificar corretamente os indivíduos quanto à melhora ou não. O uso da curva ROC permite verificar o relacionamento entre a medida e um referencial externo. A desvantagem da estratégia ROC é o fato de termos que dicotomizar a informação entre melhora ou não melhora ou piora ou não piora. Esta dualidade de resultados resume bastante a informação da magnitude da variação em relação ao critério estabelecido (Husted e col., 2000).

Para mensurações com mais de dois pontos no tempo deve-se usar estratégias como modelos de regressão e curvas de crescimento, apresentadas por McCullagh (1980) e Rogosa e col. (1982), respectivamente. Cabe destacar que a primeira referência possuía, em janeiro de 2007, mais de 900 citações na *ISI Web of Knowledge*. Este volume evidencia a importância desse artigo para o tema.

3.4 INTERPRETAÇÃO DOS RESULTADOS DA MENSURAÇÃO DA VARIAÇÃO

A diversidade de indicadores, bem como as interpretações diferenciadas dos pontos de referência e a falta de consenso sobre a melhor abordagem para a mensuração da variação geram problemas nas interpretações dos resultados obtidos.

Na Psicometria, Cronbach e Furby (1970) consideraram como um quebra-cabeça a mensuração da variabilidade. Eles alertaram para o fato de que a mensuração mais básica, até então empregada, o escore da diferença entre pré-intervenção e pós-intervenção ($D_x = X_2 - X_1$) provocava, em determinadas situações, conclusões falaciosas. A questão que justifica esta imprecisão refere-se à correlação existente entre as parcelas das mensurações (X_1) e (X_2), geralmente altas, que influencia a

confiabilidade ou estabilidade do escore da diferença (D_x). Isto é, se as parcelas possuírem alta correlação tem-se baixa confiabilidade do escore da diferença. Por exemplo, quando as parcelas têm confiabilidade de 0,80 isoladamente, as confiabilidades dos escores da diferença serão 0,6, 0,5, 0,33 e 0,0 associadas as correlações entre (X_1) e (X_2) de 0,5, 0,6, 0,7 e 0,8 respectivamente (Bergh e Fairbank, 2002). Entretanto, alguns autores durante os anos 80, entre eles, Rogosa e col. (1982), demonstraram que o escore da diferença não era uma estimativa tão imprecisa e deveria ser usado quando as parcelas têm: alta confiabilidade, baixa correlação e variâncias são heterogêneas.

Os escores da diferença são frequentemente correlacionados com os valores da linha de base. Essa correlação é negativa, pois a variância de (X_2) é normalmente menor do que a variância de (X_1). Este fenômeno implica em que indivíduos com valores altos do escore da diferença provavelmente possuem valores baixos de (X_1), enquanto indivíduos com valores altos de (X_1) tendem a atingir valores baixos de (X_2). Esse fenômeno é conhecido como regressão à média. Isto quer dizer que, indivíduos com valores altos para (X_1) migram para valores baixos em (X_2) e indivíduos com baixos valores de (X_1) migram para valores maiores em (X_2), todos em direção a um valor médio (Bergh e Fairbank, 2002).

Esta correlação pode levar a interpretações equivocadas, pois determinados efeitos podem ser entendidos como associados à intervenção quando na verdade estão relacionados à correlação alta entre (D_x) e (X_1). Esta correlação, mais tarde, foi definida como "*mathematical coupling*", que será abordada posteriormente.

Wright e Young (1997) avaliaram cinco indicadores da sensibilidade, entre eles a Sensibilidade Estatística, a Média de Resposta Padronizada e o Tamanho do Efeito com o objetivo de verificar se os diferentes indicadores forneciam resultados

semelhantes, quando aplicados em quatro diferentes instrumentos de mensuração. Os resultados para os indicadores foram diferentes, dependendo do instrumento. Utilizando os resultados da Média de Resposta Padronizada como exemplo, os autores evidenciaram que os resultados do indicador variaram de 0,2 a 4,9.

No trabalho de revisão feito por Terwee e colaboradores (2003), os autores alertam para o fato de que os indicadores de sensibilidade não apresentavam resultados similares nem próximos, quando utilizados para determinados instrumentos.

Twisk e Proper (2004) avaliaram dois ensaios clínicos randomizados com o objetivo de demonstrar como as diferentes definições de variabilidade podem afetar os resultados de uma investigação. Os achados evidenciaram que, dependendo do tipo de definição estabelecida para a variação, há uma influência diferente nos resultados da investigação.

Beaton (2000) pontuou que, para o caso de comparação entre grupos, deve-se ficar atento com o tempo de acompanhamento dos grupos, pois, se os tempos forem diferentes, o grupo que possuir um seguimento menor pode apresentar uma sensibilidade inferior ao outro, apenas pelo fato de que a duração do seguimento não ter sido igual.

Tu Y e col. (2004) demonstraram a existência de um relacionamento estreito entre os resultados de uma intervenção e o grau de severidade dos indivíduos na linha de base. A intenção dos autores foi estudar a influência do fenômeno chamado "acoplamento matemático" (*mathematical coupling*) na associação entre a linha de base e os resultados das investigações. Este fenômeno ocorre quando a variável, diretamente ou indiretamente, absorve o todo ou parte de outra variável. Isto é, o valor da linha de base (X_1) está contido no valor do escore da diferença ($D_x = X_2 - X_1$), onde (X_2) representa o escore final, após o término do seguimento. Portanto, os valores da linha de

base (X_1) e a diferença ($D_x = X_2 - X_1$) estão, segundo os autores, algebricamente correlacionados.

Clarke (2005) argumenta que, se as condições existentes no momento das mensurações de (X_1) e (X_2) forem diferentes, um confundimento com relação à variação entre os tempos devido aos efeitos das condições, pode surgir. Para o caso de uma intervenção em que as condições estejam associadas com a variação, tem-se um indicador enviesado. Um exemplo do impacto das condições é o efeito da aprendizagem, no qual cada indivíduo é avaliado em cada ponto no tempo, porém, ele aprende por completo no tempo (X_2). Para este caso, em especial, o confundimento deve ser tratado na etapa de análise dos resultados.

Portanto, ao mensurar variação é necessário conhecer o maior número possível de aspectos relativos às variáveis utilizadas, no sentido de definir o que seria variação para o investigador. Em seguida, estabelecer indicadores pertinentes à definição de variação estabelecida e, por conseguinte, interpretar os resultados dos indicadores com cautela, considerando os problemas descritos na seção 3.4.

4 VARIÁVEIS ORDINAIS NA MENSURAÇÃO DA VARIAÇÃO EM SAÚDE

Há mais de 100 anos, escalas de graduação são usadas nas áreas de psicologia e, nos dias atuais, estão disseminadas por diversas disciplinas. Instrumentos de avaliação da mensuração com escalas de graduação, tais como questionários, normalmente, utilizam avaliações categóricas ordinais em seus escores para mensurar a variabilidade de variáveis tais como grau da dor, percepção da saúde, comportamento, habilidades físicas e qualidade de vida (Svensson, 2000). Assim, esforços têm sido empregados no sentido de estabelecer indicadores que avaliam o comportamento de tais variáveis, conforme visto em 3.3. Entretanto, há uma falta de preocupação, por parte dos profissionais da saúde, em reconhecer que as características dos dados representam ponto relevante, para as análises estatísticas produzirem resultados confiáveis.

4.1 O USO DE ESCORES ORDINAIS E SUAS CARACTERÍSTICAS

No contexto da saúde encontram-se exemplos de mensurações com escores ordinais. Os tumores são estagiados de acordo com o grau de desenvolvimento. A classificação internacional para o estadiamento do carcinoma cervical é uma escala ordinal de 0 a 4. A artrite reumatóide é classificada, de acordo com a gravidade da doença, variando da atividade normal (classe 1) até a restrição a cadeira de rodas (classe 4). Os valores de Apgar, que descrevem a maturidade de lactantes recém-nascidos variam em uma escala de 0 a 10, entre outros exemplos.

A escala ordinal mede atributos que se distinguem em grau ou intensidade. Promove relações do tipo “maior (ou menor) que”, estabelece uma hierarquia entre os valores, definindo um sentido de orientação de forma que a ordem aritmética dos valores não reflete necessariamente a ordem hierárquica. Por exemplo, estadiamento em

câncer que pode ser codificada como 1, 2, 3 e 4, com estabelecimento de um sentido de orientação tal que 1 > 2 > 3 > 4, ou seja, estádio 1 tem um melhor prognóstico e estádio 4 tem um pior prognóstico. Além disso, a escala é estruturada por classes separadas, mutuamente exclusivas. Essas escalas podem ser simplesmente representadas por números ou dicotomizadas por um ponto de corte que estabelece que valores estejam acima ou abaixo deste ponto.

A escala ordinal de mensuração de atitudes apresentada por Likert (1932), é a mais utilizada e estabelece cinco pontos do tipo "ótimo", "bom", "regular", "ruim" e "péssimo" (apud Pereira, p. 64). A categoria "regular" pode ser entendida como a classificação de situação intermediária ou de indiferença. Esta escala apresenta as seguintes vantagens: estabelece uma ordem, define uma situação de indiferença e oposição entre contrários, ótimo e bom contra ruim e péssimo. A desvantagem é que não podemos afirmar que a diferença entre "ótimo" e "bom" é a mesma entre "ruim" e "péssimo".

Outra questão relevante é o fato de que a escala permite diferentes interpretações, de uma mesma variação, para variáveis diferentes. Se transformássemos a escala Likert em uma escala de 0 a 10, a medida ganharia em precisão, entretanto perderíamos em acurácia, visto que dificilmente um pesquisador consegue estabelecer onze categorias diferentes, mutuamente exclusivas, para classificar o evento de interesse, como por exemplo, ótimo, quase ótimo, muito bom, bom, quase bom, regular, quase ruim, ruim, muito ruim, quase péssimo e péssimo (Pereira, 2004).

Exemplos de outras escalas do tipo ordinal utilizadas podem ser vistas nos seguintes instrumentos de avaliação: o *Roland-Morris Disability Questionnaire* e o *Physical Health Scales of the Medical Outcomes Study 36-Item Short-Form Health Survey - SF-36*. Esses instrumentos são utilizados na mensuração da qualidade de vida

do paciente relacionada com dimensões (conjunto de atributos) do cuidado médico. O primeiro adota escores de avaliação de 0 a 24 e o segundo escores de 0 a 100.

A Tabela 1 exemplifica por uma escala de dor, a estrutura mais utilizada para a análise longitudinal de dados, do tipo ordinal, bem como apresenta os indicadores resumo, conforme discutido na seção 3.3.

A última coluna (escore da diferença) é obtida pela diminuição entre o grau de dor na consulta subseqüente e o grau de dor da consulta anterior. A coluna “média” mostra o valor médio entre o grau de dor da consulta anterior e subseqüente de cada paciente. A última linha da tabela demonstra a média dos graus de dor de todos os pacientes na consulta anterior e subseqüente.

Tabela 1. Grau de dor, segundo a Escala Visual Analógica (EVA), de seis pacientes em acompanhamento ambulatorial.

Paciente	**Grau de Dor Consulta Anterior**	**Grau de Dor Consulta Subseqüente**	**Média**	**Escore da Diferença**
1	8	4	6	-4
2	4	5	4,5	+1
3	6	5	5,5	-1
4	6	1	3,5	-5
5	9	8	8,5	-1
6	4	0	2	-4
Média	6,16	3,83	5	-2,33

4.2 A CONTROVÉRSIA ORDINAL

Além dos problemas já descritos, que afetam a mensuração da variação, entre eles, as diferentes definições de variações e seus respectivos indicadores e a utilidade dos escores de diferença, há dois temas que permanecem em discussão: a) subjetividade (imprecisão) do dado categórico ordinal e b) o tratamento estatístico (análise paramétrica e não-paramétrica) mais adequado a ser usado nas análises dos resultados.

4.2.1 A SUBJETIVIDADE DA ESCALA

Kampen e Swyngedouw (2000) descreveram um problema que retrata bem essa subjetividade da escala. Os pesquisadores perguntaram a dois respondentes sobre como eles preferem o café: 1 = fraco, 2 = normal e 3 = forte e obtiveram as respostas 1 e 3. Com o objetivo de calibrar o que os respondentes consideravam fraco e forte, eles estipularam que café fraco seria com menos de cinco colheres de chá de café, entre cinco e oito colheres de chá de café consideraram normal e acima de 9 colheres de chá de café seria forte. A conclusão dos autores é que não há razão para esperar a priori que um outro respondente adotará a mesma categorização nas quantidades de colheres de chá. Nesta situação, o critério para ser forte ou fraco pode ser utilizado de formas diferentes entre os respondentes e, portanto, os dados ordinais não podem ser comparados nem interpretados adequadamente.

Outro ponto para reflexão é que dados ordinais utilizados em saúde são frequentemente extraídos de questionários. Neste tipo de mensuração cada questão corresponde a uma variável, apresentada de forma ordenada, segundo as alternativas referentes às categorias. Tradicionalmente, essas alternativas são representadas, arbritariamente, por números. Ante o exposto, não se conhece exatamente a distância entre as classes, causando dificuldades de interpretação. Por exemplo, qual o significado de uma variação de 3 pontos em uma escala de avaliação da dor? Será que indivíduos que variaram 3 pontos acima da diferença mínima importante têm a mesma interpretação, sabendo que eles iniciaram o estudo com escores diferentes? Uma variação de 5 pontos de um indivíduo é muito diferente de uma variação de 7 pontos para outro indivíduo? (Merbitz, 1989).

No sentido de minimizar o problema das distâncias imprecisas entre valores da escala, uma equipe de revisores dos níveis de escala para reabilitação (*Level of*

Rehabilitation Scale – LORS II) argumentou que a definição de cada nível da escala foi baseada por consenso entre especialistas da área de reabilitação. Eles chamaram este tipo de escala de "*quasi-ratio scales*".

Segundo Merbitz (1989, p. 309, tradução nossa), "valores ordinais devem ser chamados de *ordinal nonnumber*, pois as categorias da escala podem ser mais consideradas como símbolos do que números".

Sintetizando o problema relacionado ao primeiro pilar da controvérsia, que diz respeito à subjetividade da escala, tem-se a imprecisão da distância entre as classes, dificultando assim a interpretação dos resultados.

4.2.2 O USO DAS ANÁLISES PARAMÉTRICAS E NÃO-PARAMÉTRICAS

Stevens (1946) foi o primeiro autor a sugerir que a tipologia das variáveis deveria ser um requisito para o uso de determinados procedimentos estatísticos de análise de dados. Esta proposta foi documentada, tempos depois, por Siegel (1956) e Senders (1958). Ele determinou a análise estatística apropriada para cada tipo de variável, e deste modo, análises não-paramétricas seriam adequadas para variáveis nominais e ordinais, enquanto análises paramétricas seriam adequadas para variáveis intervalar e razão (Gaito, 1980). Pela sugestão de Stevens (apud Velleman e Wilkinson, 1993, p. 66), análises estatísticas para variáveis categóricas nominais deveriam ser apresentadas por números de casos, moda e tabelas de contingência. Para dados ordinais, as análises possíveis seriam as anteriores mais medianas, percentis e correlações ordinais. Para as intervalares seriam as anteriores mais médias, desvios-padrão, correlação de Pearson. Por fim, para as variáveis do tipo razão seria todas as análises descritas anteriormente, acrescentando média geométrica e coeficientes de variação.

Segundo Knapp (1990), existem dois grupos que discutem as regras definidas por Stevens, os liberais (anti-Stevens) e os conservadores (pro-Stevens). Para os liberais, as diferenças entre as categorias da escala ordinal são iguais, portanto operações matemáticas são possíveis. Entretanto, autores considerados conservadores (Marcus-Roberts & Roberts, 1987), demonstraram resultados “estranhos” quando usam médias, desvios padrão e coeficientes de Pearson com escalas ordinais. Os conservadores entendem que essa decisão metodológica pode causar erros de inferência, principalmente pela inadequação das estratégias de análise utilizadas. Para os liberais, não importa qual o tipo de variável definido no delineamento da investigação, pois para eles a estratégia paramétrica sempre é mais adequada e robusta para as análises. Eles referenciam com freqüência dois trabalhos, Baker, Hardyk e Petrinocich (1966) e Labovitz (1967). Ambos mostraram empiricamente que tratar uma variável ordinal como numérica durante as análises é o que menos importa. Segundo Gaito (1980), “os números não sabem de onde eles vêm”. Essa afirmativa tinha como objetivo explicar o que os liberais queriam provar.

Segundo Kampen e Swyngedouw (2000), para alguns metodologistas, dados ordinais não permitem a avaliação de variação por meio de análises paramétricas. Outro grupo de metodologistas entende que mensurações derivadas de escalas são irrelevantes para análises estatísticas. Por fim, um terceiro grupo advoga que, subjacente a uma variável ordinal observada, existe uma variável contínua não observada e, para essa última, seria possível aplicar uma análise paramétrica.

Talvez, duas explicações pelas quais profissionais de saúde usem análises paramétricas em lugar das não-paramétricas, para ordinais, sejam que a maioria das análises não-paramétricas não possui procedimentos que trata de questões multivariadas, isto é, questões com mais de duas variáveis ordinais e, também, o

desleixo por parte dos profissionais (médicos, pesquisadores, gestores entre outros), em não valorizar as limitações das variáveis ordinais (Kampen e Swyngedouw, 2000).

Resumindo, o debate apresentado anteriormente cria uma barreira ao entendimento sobre o uso mais adequado das estratégias estatísticas na análise de variação com dados ordinais.

4.3 MENSURAÇÃO DA VARIAÇÃO LONGITUDINAL PARA DADOS ORDINAIS

A forma mais básica para analisar a variação dos atributos ordinais de indivíduos, levando em consideração o tempo decorrido após uma intervenção, é caracterizada pela diferença entre pelo menos duas medidas repetidas (antes e depois). Leffondré e colaboradores (2004) classificaram esse tipo de indicador como de não-linearidade ou de inconsistência de variação, conforme descrito em 3.3. Para Rogosa e Zimowski (1982), a avaliação da variação definida como antes e depois é referenciada como *Two-wave* [7] *Longitudinal Panel Design*.

Deste modo, as informações utilizadas na análise são as duas distribuições marginais, antes e depois, derivadas de tabelas de contingência, conforme modelo da Tabela 2.

Tabela 2. Modelo de tabela de contingência.

Depois / **Antes**	1	...	m	Total
1				Linha 1
...				...
m				Linha m
Total	Coluna 1	...	Coluna m	

[7] O termo *two-wave*, no contexto da mensuração da variação, significa que foram realizadas duas mensurações das características dos indivíduos, ao longo do tempo.

A primeira coluna representa as ***m*** classes ou categorias de um dado ordinal mensurado na linha de base (início da intervenção). A primeira linha representa as mesmas ***m*** classes ou categorias da variável ordinal, porém, mensurados após algum tempo decorrido. As últimas, linha e coluna, representam os totais de indivíduos em cada classe nos dois pontos mensurados, antes e depois da intervenção, por exemplo. E, no "miolo" da tabela, está a quantidade de indivíduos que migram ou não de uma classe para a outra ao longo do tempo.

É importante lembrar que os problemas relatados anteriormente, tais como: a subjetividade da escala, os impactos dos escores da linha de base e a controvérsia das análises, estão, também, presentes nesta forma básica de avaliação.

Com o objetivo de buscar alternativas de avaliação da mensuração com variáveis ordinais, alinhadas com os pressupostos dos conservadores (pro-Stevens), que não sofressem dos problemas relatados, e que, ao mesmo tempo, utilizassem tabelas de contingência, estratégia aplicada pelo Indicador de Melhora (seção 5), duas abordagens foram encontradas. A primeira descrita por Chen e Yang (1979) e a segunda apresentada por Sonn e Svensson (1997).

4.3.1 O MÉTODO DE SVENSSON

Sonn e Svensson (1997) apresentaram métodos estatísticos não-paramétricos que consideram os diferentes aspectos da variabilidade (inter- e intra-indivíduos), existentes entre mensurações pareadas, para variáveis categóricas ordinais. A autora entende que a principal propriedade de uma variável ordinal é que as classes ou categorias representam uma ordenação gradual de acordo com a intensidade de um específico fenômeno. Assim, as distâncias entre as classes são indeterminadas, isto é, a ordenação não representa nenhum valor matemático e sim uma ordem. Como conseqüência deste entendimento, Sonn e Svensson não recomendam a utilização de escores da diferença

($D_x = X_2 - X_1$), pois operações matemáticas não seriam permitidas. Esta posição alinha-se ao pensamento dos conservadores.

A metodologia busca analisar a variação existente na distribuição marginal[8] entre duas mensurações, X_1 e X_2, realizada ao longo do tempo, em um grupo, identificando em separado as variações individuais. A análise da variação do grupo na distribuição marginal apresentada na Tabela 3 é demonstrada por um indicador chamado de "*Relative Position – RP"*. Este indicador é definido pela diferença entre duas probabilidades: a probabilidade de a mensuração inicial ser menor do que a final e a probabilidade da mensuração final ser menor do que a inicial ($P(X_1 < X_2) - P(X_2 < X_1)$). Este indicador varia de -1 a 1.

As classes, da escala ordinal, estão indexadas por ***i***, que varia de **0** a ***m***, os números de observações na *i*-ésima classe nos tempos *X* e *Y* estão definidos como $\boldsymbol{x_i}$ e $\boldsymbol{y_i}$. O total de todas as classes está representado por ***n*** e, **C(X)** e **C(Y)** são as freqüências acumuladas relativas aos tempos *X* e *Y*.

Tabela 3. Distribuição marginal de indivíduos avaliados antes e depois, segundo uma variável ordinal.

Classes	0	1	2	3	4	5	6	7	...	*m*
X (Antes)										
Total (Dist. Marginal)	$\boldsymbol{x_0}$	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	$\boldsymbol{x_4}$	$\boldsymbol{x_5}$	$\boldsymbol{x_6}$	$\boldsymbol{x_7}$	**…**	$\boldsymbol{x_m}$
Freqüência Acumulada	$C(X)_0$	$C(X)_1$	$C(X)_2$	$C(X)_3$	$C(X)_4$	$C(X)_5$	$C(X)_6$	$C(X)_7$	**…**	$C(X)_m$
Freq. Acumu. Relativa										
Y (Depois)										
Total (Dist. Marginal)	$\boldsymbol{y_0}$	$\boldsymbol{y_1}$	$\boldsymbol{y_2}$	$\boldsymbol{y_3}$	$\boldsymbol{y_4}$	$\boldsymbol{y_5}$	$\boldsymbol{y_6}$	$\boldsymbol{y_7}$	**…**	$\boldsymbol{y_m}$
Freqüência Acumulada	$C(Y)_0$	$C(Y)_1$	$C(Y)_2$	$C(Y)_3$	$C(Y)_4$	$C(Y)_5$	$C(Y)_6$	$C(Y)_7$	**…**	$C(Y)_m$
Freq. Acumu. Relativa										

Fonte: Tabela adaptada de Sonn e Svensson (1997).

[8] A expressão "distribuição marginal" refere-se as distribuições de observações, com escores ordinais, mensuradas em dois momentos.

A fórmula de *RP* pode também ser escrita da seguinte maneira: $RP = p_0 - p_1$ onde

$$p_0 = \frac{1}{n^2}\sum_{i=1}^{m}\left[y_i \cdot C(X)_{i-1}\right] \qquad p_1 = \frac{1}{n^2}\sum_{i=1}^{m}\left[x_i \cdot C(Y)_{i-1}\right]$$

Resultados de *RP* próximos de zero significam que não ocorreu modificação na distribuição marginal ao longo do tempo. Para os casos em que X_2 apresenta valores mais altos que X_1, o indicador *RP* apresentará resultados positivos.

Uma das maneiras, utilizada por Svensson, de se observar os resultados graficamente é pela curva ROC. Nos casos positivos ($X_2 > X_1$) a curva tenderia a ficar abaixo da diagonal, conforme Figura 1. E, para os casos negativos ela ficaria acima da diagonal.

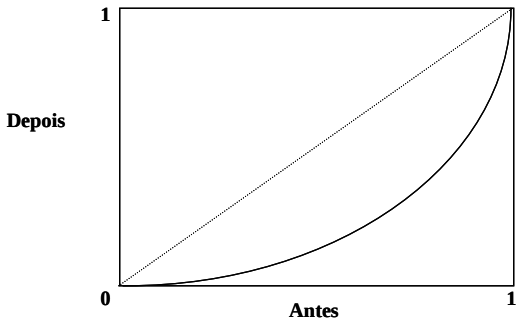

Figura 1. Curva ROC para a variação geral do grupo entre duas medidas (antes e depois), para casos de *RP* positivos.

As variações individuais são evidenciadas pelo indicador chamado de "*Relative Rank-Variance – RV*". Essa avaliação pode ser entendida como uma mensuração da dispersão individual, isto é, variações individuais que não são explicadas e que fazem parte da variação total do grupo, ou simplesmente, uma parte aleatória da discordância dos pares de observações. Valores baixos de *RV* significam que a quantidade de variação individual é pequena, dentro da variação do grupo.

Os valores de *RV* variam de 0 a 1 e a sua fórmula é dada por:

$$RV = \frac{6}{n^3} \sum_{i=1}^{m} \sum_{j=1}^{m} x_{ij} \left[x_{ij}^{ae} - x_{ij}^{bd} \right]$$

A metodologia de cálculo de *RV* está baseada nas observações localizadas nas células acima-esquerda (***ae***) e abaixo-direita (***bd***) relativas a uma célula específica ($\mathbf{C_{ij}}$), conforme Figura 2.

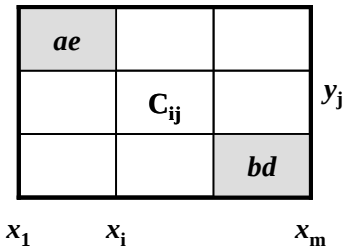

Figura 2. Áreas que contribuem para a mensuração da dispersão individual, *RV*.

A Tabela 4 e a Figura 3 apresentam um exemplo adaptado, retirado de Sonn e Svensson (1997), para demonstrar a aplicabilidade do método. A investigação se dá em uma população geral, residente em uma comunidade. Entre as variáveis levantadas, a Atividades da Vida Cotidiana (*Activities of Daily Living – ADL*) foram mensuradas por uma escala ordinal com 10 classes, conhecida como "*Staircase of ADL*", em grupo de indivíduos com 70 anos (antes), que foram seguidos e reavaliados aos 76 anos (depois).

Tabela 4. Distribuição de 371 indivíduos aos 70 e 76 anos, segundo a escala *ADL*.

Classes	0	1	2	3	4	5	6	7	8	9
X = 70 (Antes)										
Total (Dist. Marginal)	338	17	2	8	3	2	1	0	0	0
Freqüência Acumulada	338	355	357	365	368	370	371	371	371	371
Freq. Acumu. Relativa	0,91	0,96	0,96	0,98	0,99	1,00	1,00	1,00	1,00	1,00
Y = 76 (Depois)										
Total (Dist. Marginal)	246	34	21	20	17	7	12	3	9	2

Freqüência Acumulada	246	281	301	321	338	345	357	360	360	371
Freq. Acumu. Relativa	0,66	0,76	0,81	0,87	0,91	0,93	0,96	0,97	0,99	1,00

Fonte: Tabela adaptada de Sonn e Svensson (1997).

A forma de avaliação da variação da escala *ADL*, no nível do grupo (*RP*), foi calculada da seguinte maneira:

$p_0 = (1 / 371^2)$ **x** (34**x**338 + 21**x**355 + 20**x**357 + 17**x**365 + 7**x**368 + 12**x**370 + (3+9+2)**x**371)

$p_0 = 44.502 / 371^2 = 0,3233$

$p_1 = (1 / 371^2)$ **x** (17**x**246 + 2**x**281 + 8**x**301 + 3**x**321 + 2**x**338 + 1**x**345 + 0)

$p_1 = 9.136 / 371^2 = 0,0664$

RP = 0,3233 – 0,0664 = 0,257

Para valores positivos, entende-se que há variação para o grupo como um todo, isto é, há um impacto favorável da intervenção, evidenciado pela quantidade maior de indivíduos nos escores mais altos da classe. Assim, existe mais chance (25,7%) de se obter valores altos da escala aos 76 anos do que aos 70 anos.

A Figura 3 é elaborada com o objetivo de identificar variações individuais divergentes do esperado para o nível do grupo. Na primeira linha da figura têm-se as classes da ADL em ordem crescente, mensuradas aos 70 anos (antes) e na primeira coluna têm-se as mesmas classes, porém em ordem decrescente, mensurada aos 76 anos (depois). Os valores dentro da figura representam as variações individuais, a partir da primeira mensuração (70 anos). Isto é, o indivíduo que variou de 4, os 70 anos, para 8, aos 76 anos, corresponderia à célula $C_{48} = 1$. Os valores da diagonal indicam indivíduos que permaneceram nas mesmas classes, isto é, não apresentaram nenhuma variação.

Antes / Depois	0	1	2	3	4	5	6	7	8	9	Total
9	1	1									2
8	5	2		1	1						9
7	2				1						3
6	8			2	1	1					12
5	6					1					7
4	14	3									17
3	14	2	1	3							20
2	14	4	1	2							21
1	31	2					1				34
0	243	3									246
Total	338	17	2	8	3	2	1	0	0	0	371

Figura 3. Distribuição de freqüência das variações individuais entre os grupos antes (X=70 anos) e depois (Y=76 anos).

A forma de avaliação da variação da escala *ADL*, no nível individual (*RV*), foi calculada da seguinte maneira:

$$\boldsymbol{RV} = (6 / 371^3)\ \mathbf{x}\ [31\mathbf{x}3^2 + 14\mathbf{x}6^2 + 14\mathbf{x}13^2 + 14\mathbf{x}19^2 + 6\mathbf{x}22^2 + 8\mathbf{x}23^2 + 2\mathbf{x}27^2 + 5\mathbf{x}28^2 + 1\mathbf{x}32^2 + 3\mathbf{x}95^2 + 2\mathbf{x}64^2 + 4\mathbf{x}(50-1)^2 + 2\mathbf{x}(36-4)^2 + 3\mathbf{x}(22-8)^2 + 2\mathbf{x}(1-14)^2 + 1\mathbf{x}16^2 + 1\mathbf{x}(58-1)^2 + 1\mathbf{x}(42-3)^2 + 2\mathbf{x}(59-1)^2 + 3\mathbf{x}(42-1)^2 + 2\mathbf{x}(11-2)^2 + 1\mathbf{x}(2-5)^2 + 1\mathbf{x}(12-2)^2 + 1\mathbf{x}(10-3)^2 + 1\mathbf{x}(2-3)^2 + 1\mathbf{x}(25-1)^2 + 1\mathbf{x}(14-1)^2 + 1\mathbf{x}91^2] = 0{,}0112$$

Para valores pequenos de *RV*, próximos de zero, entende-se que há homogeneidade de variação dentro do grupo, portanto, não haveria necessidade de intervenções individualizadas.

Cabe ressaltar que a autora apresenta, ainda, em seu artigo a estimativa do erro padrão para os indicadores.

As vantagens e desvantagens dessa abordagem são:

Vantagens

- não requer pressupostos de tipos de distribuições estatísticas;
- não tem limite de classes;
- utiliza uma medida para mensurar desvios individuais;
- possibilita estimar o erro padrão dos indicadores.

Desvantagem

- dificuldade de interpretação;
- não discute o impacto dos escores da linha de base.

Os artigos de Sonn e Svensson (1997) e Svensson (1998) são publicações de referência para o tema mensuração da variação com dados categóricos ordinais e úteis aos interessados em aplicar o método. Ainda sobre os artigos, em uma avaliação feita, no "site" *Web of Science*, em janeiro de 2007, foram encontradas 19 citações para o primeiro artigo e 41 citações para o segundo.

4.3.2 O MÉTODO DE CHEN E YANG

Chen e Yang (1979) apresentaram um indicador quantitativo, para comparação entre grupos, baseado no conceito de estágio da doença. Esse conceito estabelece que uma determinada doença possa ser classificada de um a três estágios mutuamente exclusivos. Por exemplo: estágio I, sem complicação; estágio II, com complicação local; estágio III, complicações sistêmicas. O indicador é próprio para estudos comparativos (dois grupos), em um contexto onde os dados são categóricos e pareados. O modelo permite o uso de análises paramétricas tais como: z ou teste *t* para testar hipóteses e pode ser aplicado para qualquer outro aspecto além de doença, que pode ser classificado, também, em três categorias como por exemplo estado funcional ou estado de saúde mental

O indicador utiliza uma tabela de contingência, representada pela Tabela 5.

Tabela 5. Distribuição de indivíduos entre os estágios (normal, médio e severo) no tempo 1 e no tempo 2.

Tempo 1 \ **Tempo 2**	Normal	Médio	Severo	Total
Normal	Célula $_{11}$	Célula $_{12}$	Célula $_{13}$	Normal
Médio	Célula $_{21}$	Célula $_{22}$	Célula $_{23}$	Médio
Severo	Célula $_{31}$	Célula $_{32}$	Célula $_{33}$	Severo
Total	Normal	Médio	Severo	Total

A última coluna representa a distribuição marginal das mensurações realizadas na linha de base, enquanto a última linha representa a distribuição marginal ao final da intervenção. No miolo da tabela, as células que representam o movimento dos indivíduos são $\mathbf{C_{12}}$, $\mathbf{C_{13}}$, $\mathbf{C_{21}}$, $\mathbf{C_{23}}$, $\mathbf{C_{31}}$ e $\mathbf{C_{32}}$. Nesta representação, já definida na Tabela 5, é esperada que a distribuição dos indivíduos, de cada grupo, migre em direção ao valor normal ou para valores menos severos, depois de realizada a segunda avaliação. As células da diagonal, $\mathbf{C_{11}}$, $\mathbf{C_{22}}$ e $\mathbf{C_{33}}$, indicam indivíduos que permaneceram inalterados para cada grupo avaliado.

Além da estrutura apresentada na Tabela 5, a abordagem utiliza pesos para as células. Esses pesos podem significar o grau de eficácia dos sistemas de saúde, os custos relativos aos cuidados prestados, à qualidade do cuidado, ou até mesmo o ponto de vista da sociedade ou dos pacientes. Os valores para os pesos foram assim definidos: peso 1, para máxima eficácia e peso 7 para a mínima eficácia, conforme apresentado na Tabela 6.

Tabela 6. Distribuição dos pesos alocados nas nove células de uma tabela de contingências.

Tempo 2 / Tempo 1	Normal	Médio	Severo
Normal	**3**	**6**	**7**
Médio	**2**	**4**	**6**
Severo	**1**	**2**	**5**

Assim, o indicador é definido da seguinte maneira:

$$I = \frac{1}{N}\sum_{j=1}^{k} w_j n_j$$

No qual w_j é o peso alocado para cada célula e n_j o número de indivíduos da j-ésima célula (Tabela 5). As células com pesos iguais ($\mathbf{C_{12}}$ e $\mathbf{C_{23}}$; $\mathbf{C_{21}}$ e $\mathbf{C_{32}}$) indicam eficácias iguais e, portanto, devem ser tratadas como uma única célula. Valores próximos de um significam um grande impacto da intervenção na população avaliada.

As vantagens e desvantagens dessa abordagem são:

Vantagens

- controla os efeitos dos valores da linha de base, na comparação de dois grupos.

Desvantagens

- dificuldade de interpretação;
- subjetividade na definição dos pesos;
- limitado em três tipos de classificação;
- foco na comparação de grupos;
- não discute há necessidade de pressupostos de tipos de distribuições estatísticas.

O método foi descrito principalmente por apresentar a construção de seu indicador com base em uma tabela de contingência. Essa estratégia norteou a construção da proposta do Indicador de Melhora apresentado na seção 5.

Em virtude de o método não se preocupar com pressupostos ressaltados pelos conservadores (pro-Stevens), conforme relatado na seção 4.3.2, entende-se que a estratégia se tornou apenas mais uma abordagem, ainda sem consenso, sobre a melhor maneira de analisar variação com dados ordinais.

Para identificar a difusão do método, foi realizado um levantamento das citações existentes pelo “site” *Web of Science*, em janeiro de 2007, sendo obtido que nenhuma citação foi feita desde a sua publicação.

5 UM NOVO INDICADOR

O Indicador de Melhora (IM) proposto visa oferecer mais uma estratégia de análise da variação com mensurações repetidas (duas ocasiões distintas) de um mesmo indivíduo, assumindo que a variável mensurada é do tipo ordinal.

A nova abordagem está baseada na distribuição conjunta dos escores “antes” e “depois”, representada pelas células (***i***,***j***) de uma tabela de contingência ***m*** x ***m***, seguindo o mesmo raciocínio de Chen e Yang. Nesse caso, ***m*** significa a quantidade de classes utilizadas pela escala de mensuração ordinal. As linhas e colunas são representadas por ***i*** e ***j***, que podem variar de 1 a ***m***. Para um melhor entendimento, suponha que as classes mais elevadas são os mais desfavoráveis da escala, portanto, é esperado que unidades que se situam em classes altas migrem para classes menores, dependendo do objetivo da intervenção. A Tabela 7 apresenta a distribuição hipotética dessa matriz. Diferentemente de Svensson, a tabela abaixo dispõe as classes sempre em ordem crescente. A diferença relacionada ao método de Chen e Yang está na disposição da avaliação antes (primeira linha) e depois (primeira coluna).

Tabela 7. Distribuição hipotética de indivíduos participantes de duas avaliações realizadas durante um intervalo de tempo.

Tempo 2 \ **Tempo 1**	**1**	**2**	**3**	**...**	***j***	***m***
1	**Célula $_{11}$**	**Célula $_{12}$**	**Célula $_{13}$**	**...**	**Célula $_{1j}$**	**Célula $_{1m}$**
2	**Célula $_{21}$**	**Célula $_{22}$**	**Célula $_{23}$**	**...**	**Célula $_{2j}$**	**Célula $_{2m}$**
3	**Célula $_{31}$**	**Célula $_{32}$**	**Célula $_{33}$**	**...**	**Célula $_{3j}$**	**Célula $_{3m}$**
...	**...**	**...**	**...**	**...**	**...**	**...**
i	**Célula $_{i1}$**	**Célula $_{i2}$**	**Célula $_{i3}$**	**...**	**Célula $_{ij}$**	**Célula $_{im}$**
m	**Célula $_{m1}$**	**Célula $_{m2}$**	**Célula $_{m3}$**	**...**	**Célula $_{mj}$**	**Célula $_{mm}$**
Total do Início	**N_1**	**N_2**	**N_3**	**...**	**N_j**	**N_m**

Na Tabela 7, a última linha mostra o número de indivíduos que iniciaram a avaliação, segundo a escala adotada. As células (***i,j***) são as observações da variação dos indivíduos decorrente da intervenção. Por exemplo, a célula C_{23} mostra os indivíduos que iniciaram na classe 3 e, no tempo 2, foram classificados na classe 2. A célula C_{i1} mostra indivíduos que estavam nas classes mais baixas no tempo 1, e passaram a ter classes máximas no tempo 2. A célula C_{22} indica que não aconteceu variação, pois os indivíduos que iniciaram no escore 2 permaneceram no mesmo escore no tempo 2. Assim, pode-se observar que as células em diagonal apresentam indivíduos que não sofreram variação, nem para melhora nem para a piora.

A fórmula do indicador proposto é:

$$\text{Indicador de Melhora (IM)} = \frac{\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} C_{ij}(j-i)}{\sum_{j=2}^{m} N_j(j-1)} \text{ X } 100$$

A expressão $(j-i)$ representa a magnitude da melhora de cada indivíduo. Por exemplo, a C_{13} representa indivíduos que possuíam no início da avaliação classe 3 e, no tempo 2, encontra-se na classe 1, portanto, a magnitude da variação para estes indivíduos foi de duas classes. O indicador proposto evidencia o percentual de variação favorável possível de ser alcançada devido à intervenção.

Se o número de indivíduos, no início da avaliação, for maior para as classes mais desfavoráveis, o indicador tenderá a valores altos. Por outro lado, se o número de indivíduos for maior nas classes mais favoráveis, no início do estudo, o indicador tenderá a valores menores. Além disso, é recomendável o seu uso quando o fenômeno de interesse produz efetivamente níveis de melhora.

Exemplo

Os dados para o exemplo foram extraídos de um estudo realizado pelo serviço de odontologia da Universidade Federal do Rio de Janeiro - UFRJ. A proposta da avaliação foi mensurar a melhora da gengivite entre a primeira e segunda consulta feitas ao dentista. Os pacientes foram mensurados segundo ***m*** = 4 classes, 1(normal) a 4(severa), (Tabela 8).

Tabela 8. Avaliação da variação da mensuração gengival nos sítios em relação à primeira consulta.

1ª Consulta / 2ª Consulta	1	2	3	4
1	**562**	**674**	**213**	**3**
2	**29**	**489**	**412**	**7**
3	**6**	**31**	**76**	**6**
4	**0**	**0**	**0**	**0**
Total do Início	**597**	**1 194**	**701**	**16**

Usando a expressão anterior, o Indicador de Melhora (IM) é calculado da seguinte maneira: [C_{12} **x** (2-1) + C_{13} **x** (3-1) + C_{14} **x** (4-1) + C_{23} **x** (3-2) + C_{24} **x** (4-2) + C_{34} **x** (4-3)] / [N_2 **x** (2-1) + N_3 **x** (3-1) + N_4 **x** (4-1)]. Substituindo na expressão anterior a notação pelas observações tem-se: **IM** = [(674**x**(1))+(213**x**(2))+(3**x**(3))+(412**x**(1))+(7**x**(2))+(6**x**(1)) / (1194**x**(1))+(701**x**(2))+(16**x**(3))] **x** 100 = **58%**. Isto é, o grupo de indivíduos alcançou 58%, do total de possibilidade de melhora possível, no início do estudo. A mesma lógica pode ser adotada para avaliar a piora, representada pelas células em vermelho.

O Indicador de Melhora torna-se útil para uma análise inicial da variação, no nível de grupo, pois demonstra o quanto de benefício foi promovido pela intervenção sem ter a preocupação de atender os pressupostos de distribuições estatísticas.

Resumindo o método segundo as suas vantagens e desvantagens têm-se:

Vantagens

- não requer pressupostos de tipos de distribuições;
- facilidade de interpretação;
- permite avaliar a melhora ou a piora.

Desvantagens

- ainda não há cálculo para o erro-padrão do indicador.

O Quadro 3 abaixo, apresenta uma síntese comparativa dos três indicadores apresentados anteriormente.

Quadro 3. Síntese comparativa.

Itens de Comparação	Indicadores		
	Chen e Yang	Svensson	Novo Indicador
Facilidade de Interpretação			X
Trata o impacto da linha de base	X		
Considera a controvérsia ordinal		X	X
Foco no desenho *two-wave*		X	X
Estima o erro padrão		X	

6 CONSIDERAÇÕES FINAIS

O presente estudo teve o seu foco direcionado para investigações que apresentam duas ocasiões mensuradas consecutivamente, antes e depois de uma intervenção, com o objetivo de promover um melhor entendimento sobre o padrão de variação de determinadas características dos indivíduos, mensuradas por meio de variáveis ordinais.

Do ponto de vista conceitual, a mensuração da variação se configura como uma prática simples e objetiva, evidenciada pela diferença entre pelo menos duas medições. Assim, a força de um indicador (instrumento de medida), em mensurar uma variação, ao longo do tempo, quando ela de fato ocorreu é denominada sensibilidade (*responsiveness*).

Ao revisar a literatura, Terwee e colaboradores (2003) apresentaram vários tipos de definição de variação e seus respectivos indicadores sensíveis. Por esse trabalho, os autores demonstraram que há uma falta de consenso sobre a melhor definição de variação e qual a abordagem mais adequada a ser estabelecida para a sua avaliação.

Ante o exposto, ao se aplicar a metodologia de estabelecimento da sensibilidade de um indicador, deve-se estar atento as seguintes questões: a) quais são os objetivos da mensuração e que definições serão medidas? b) a validade e confiabilidade do indicador estão adequadas? c) como interpretar os resultados?

No que diz respeito à interpretação dos resultados, observa-se que a medida direta da de variação, o escore da diferença, apresenta limitações para o seu uso, com destaque para a sua confiabilidade e a alta correlação que existe entre os escores da linha de base e o escore da diferença. Ambos podem provocar interpretações equivocadas da variação.

Ainda sobre o aspecto da interpretação, a revisão mostra que para cada tipo de objetivo definido, deve-se utilizar um indicador que atenda a tal meta e, assim, não seria recomendável realizar qualquer tipo de comparação entre os indicadores de sensibilidade, pois as fórmulas seriam diferentes.

Outra questão que afeta a interpretação dos resultados são os efeitos das condições, isto é, se as condições forem diferentes no momento das mensurações, pode-se provocar um confundimento.

Não se pode esquecer que a sensibilidade pode ser influenciada pela distribuição dos valores dos escores dos indivíduos na linha de base. Por exemplo, uma pequena variação na função física de um indivíduo, que possui um valor baixo dentro de uma escala, no início do estudo, torna-se mais relevante do que indivíduos com a mesma variação, e, que possuem valores altos, também, no início (Mirjam e colaboradores, 2002).

Com respeito à apresentação dos resultados, é importante lembrar que diversas formas de sumarizar a sensibilidade são utilizadas e, frequentemente, são descritas por apenas um indicador, dificultando o entendimento da variação. Segundo Beaton (2000), os investigadores poderiam utilizar a distribuição dos valores dos escores, a média, o desvio padrão e gráficos, que mostrassem a trajetória de variação dos indivíduos dentro do grupo.

Em investigações que utilizam variáveis ordinais tem-se como ponto central de discussão a questão da subjetividade da escala. Essa subjetividade é caracterizada pelo uso arbritário de números, para representar a ordem das categorias e, consequentemente, a falta de precisão da distância entre eles. Assim, os números indicam uma ordem hierarquizada das categorias e não valores dentro de um contexto matemático (Svensson, 2001).

Como visto, há um debate sobre a maneira mais adequada de tratar os dados ordinais. Na prática, grande parte dos trabalhos mostra que o tratamento estatístico adotado considera dados ordinais, extraídos de uma escala de graduação, como numéricos. Segundo Altman (1998) a maioria das análises, utilizadas no tratamento dos dados, é realizada por pessoas que não entendem adequadamente os métodos estatísticos usados. Como conseqüência imediata desta falta de entendimento surge o excesso de erros estatísticos nos resultados. Altman destaca que há dificuldades no julgamento, por parte dos investigadores, no uso adequado de análises paramétrica e não-paramétricas. Além disso, Altman alerta, para o fato de que políticas sociais são freqüentemente influenciadas por investigações que utilizam resultados oriundos de variáveis ordinais.

A mensuração da variação de variáveis qualitativas do tipo ordinal é, tradicionalmente, aplicada em estudos pareados (*two-wave*) nos quais cada indivíduo é o seu próprio par. Neste sentido, a revisão procurou identificar indicadores apropriados para avaliar a variação entre dois valores consecutivos dentro de um grupo e, também, propor um novo indicador.

Inicialmente, para realizar uma boa avaliação é necessário verificar a distribuição dos valores no início do estudo, pois o grupo pode apresentar grandes concentrações na base, no topo ou em ambas as partes da escala. Em tais casos, o uso do escore de diferença seria problemático, em razão da alta correlação entre a linha de base e o escore de diferença.

Como visto, o uso de médias e desvios padrões não é recomendável, pois não se tem a exata distância entre os valores das categorias. Ainda, como estabelecer uma hipótese de diferença igual a zero, entre mensurações consecutivas, se os valores não são números e sim “rótulos numéricos”? Sobre esse aspecto, diversos autores (Gaito, 1980; Knapp. 1990; Velleman e Wilkinson, 1993; Kampen e Swyngedouw, 2000;

Svensson, 2000) discutem sobre a validade e a confiabilidade dos resultados obtidos por meio de estratégias de análises, tipicamente desenvolvidas para variáveis numéricas, sendo utilizadas para variáveis ordinais.

Sendo assim, o Indicador de Melhora (IM) proposto é uma alternativa para se conhecer a variação para um grupo, pois, além de ser fácil de calcular, tem uma interpretação simples. Para uma avaliação mais detalhada da variação do grupo a abordagem de Svensson seria apropriada, pois verifica as variações individuais que contribuíram para a variação geral.

De qualquer forma deve-se ressaltar que ao se deparar com uma técnica estatística usada na avaliação da mensuração com variáveis ordinais, os resultados desta análise devem ser interpretados com cautela.

REFERÊNCIAS

Altman DG. Statistical Reviewing for Medical Journals. Statistics in Medicine 1998;17:2661-2674.

Baker BO, Hardyck, CD, Petrinovich LF. Weak measurements vs. strong statistics: an empirical critique of S.S. Steven`s proscriptions on statistics. Education and Psychological Measurement 1966;26:291-309.

Beaton DE. Understanding the relevance of measured change through studies of responsiveness. Spine 2000;24:3192-3199.

Bergh DD, Fairbank JF. Measuring and testing change in strategic management research. Strategic Management Journal 2002;23:359-366.

Clarke PS. Analysing change based on two measures taken under different conditions. Statistics in Medicine 2005;25(22):3401-3415.

Chen MK, Yang GL. A quantitative index for evaluating patient care with longitudinal data. International Journal of Epidemiology 1979;8(3):265-271.

Cohen, J. Statistical power analysis for the behavioral sciences. 2nd ed. Lawrence Erlbaum, Hillsdale, NJ, 1988.

Cronbach LJ, Furby L. How we should measure "change" - or should we? Psychologichal Bulletin 1970;74:68-80.

Davidson M, Keating JL. A comparison of five low back disability questionnaires: reliability and responsiveness. Physical Therapy 2002; 82:8-24.

De Vet HCW, Terwee CB, Bouter LM. Current challenges in clinimetrics. J Clin Epidemiol 2003;56:1137-41.

Gaito J. Measurement scales and statistics: resurgence of an old misconception. Psychological Bulletin. 1980;87(3):564-567.

Granados AJ, Peterson PJ. Hazardous waste indicators for national decision makers. Journal of Enviroment Management 1999;55:249-263.

Guyatt GH, Kirshner B, Jaeschke R. Measuring health-status: what are the necessary measurement properties? Journal of Clinical Epidemiology 1992;45:1341-1345.

Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. Journal of Clinical Epidemiology 2000,53:459-468.

Kampen J, Swyngedouw M. The ordinal controversy revisited. Quality & Quantity 2000;34:87-102.

Kazandjian, VA, Vallejo, P. Local evaluation of quality using generic measurement tools. Gaceta Sanitária, 2004;18:225-234.

Kirshner B, Guyatt GH. A methodological framework for assessing health indices. Journal of Chronic Disease 1985;38:27-36.

Knapp TR. Treating ordinal scales as interval scales: an attempt to resolve the controversy. Nursing Research 1990;39:121-123.

Labovitz S. Some observations on measurement and statistics. Social Forces 1967;46:151-160.

Lavalley MP, Felson DT. Statistical presentation and analysis of ordered categorical outcome data in rheumatology journals. Arthrits & Rheumatism 2002;47(3):255-259.

Leffondré K, Abrahamowichz M, Regeasse A, Hawker A, Badley EM, McCusker J, Belzile E. Statistical measures were proposed for identifying longitudinal patterns of change in quantitative health indicators. Journal of Clinical Epidemiology 2004;57:1049-1062.

Liang MH. Longitudinal construct validity. Establishment of clinical meaning in patient evaluative instruments. Med Care 2000;38(9) Suplement II: II-84 - II-90.

Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum. 1985;28:542-547.

Mainz, J. Defining and classifying clinical indicators for quality improvement. International Journal for Quality in Health Care. 2003;15(6):523-530.

Marcus-Roberts HM, Roberts FS. Meaningless statistics. Journal of Educational Statistics 1987;12:383-394.

Mant, J. Process versus outcome indicators in the assessment of quality of health care. International Journal for Quality in Health Care. 2001;13(6):475-480.

Maassen GH. Kelley's formula as a basis for the assessment of reliable change. Psychometrika. 2000;65(2);187-197.

McCullagh P. Regression – Models for Ordinal Data. Journal of the Royal Statistical Society Series B – Methodological. 1980;42(2):109-142.

Merbitz C, Morris J, Grip JC. Ordinal Scales and foundations of misinference. Arch Phys Med Rehabil 1989;70:308-312.

Mirjam AGS, Carol MM, Timothy JM, Donald LP, Dennis AR. The Clinical Significance Consensus Meeting Group. Assessing meaningful change in quality of life over time: a user's guide for clinicians. Mayo Clin Proc. 2002;77:561-571.

Norman GR, Sridar FG, Guyatt GH, Walter SD. Relation of distribution- and anchor-based approaches in interpretation of changes in health-related quality of life. Medical Care 2001;39:1039-1047.

OECD. Towards more sustainable household consumption patterns indicators to measure progress. Organization for Economic Co-operation and Development, Paris 1999.

Pereira JCR. Análise de dados qualitativos: estratégias metodológicas para as ciências da saúde, humanas e sociais. 3ª ed. 1ª reimpressão. - São Paulo: Editora da Universidade de São Paulo, 2004.

Plewis I. Analyzing Change. Chichester: Wiley 1985.

Rogosa D, Brandt David, Zimowski M. A growth curve approach to the measurement of change. Psychological Bulletin 1982;92(3):726-748.

Senders VL. Measurement and statistics. New York: Oxford University Press 1958.

Siegel S. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill, 1956.

Sonn U, Svensson E. Measures of individual and group changes in ordered categorical data: application to the ADL staircase. Scand J Rehab Med 1997;29:233-242.

Stevens, SS. On the theory of scales of measurement. Science 1946;103:677-680.

Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 3rd ed. Reprinted 2004 e 2005. New York: Oxford University Press, 2005;196-212.

Stucki G, Daltroy L, Katz N, Johanneson M, Liang MH. Interpretation of change scores in ordinal clinial scales and health status measures: the whole may not equal the sum of the parts. J Clin Epidemiol 1996;49:711-717.

Svensson E. Comparison of the quality of assessments using continuous and discrete ordinal rating scales. Biometrical Journal 2000;42(4):417-434.

Svensson E. Guidelines to statistical evaluation of data from rating scales and questionnaries. J Rehab Med 2001;33:47-48.

Svensson E. Ordinal invariant measures for individual and group changes in ordered categorical data. Statistics in Medicine 1998;17:2923-2936.

Terwee CB, Dekker FW, Wiersinga WM, Prummel FM, Bossuyt PMM. On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. Quality of Life Research 2003;12:349-362.

Tu Y, Maddick IH, Griffiths GS, Gilthorpe MS. Mathematical coupling can undermine the statistical assessment of clinical research: illustration from the treatment of guided tissue regeneration. Journal of Dentistry 2004;32:133-142.

Twisk J, Proper K. Evaluation of the results of a randomized controlled trial: how to define change between baseline and follow-up. Journal of Clinical Epidemiology 2004;57:223-228.

Velleman PF, Wilkinson L. Nominal, ordinal, interval, and ratio typologies are misleading. The American Statistician. 1993;47(1):65-72.

Wright JG, Young NL. A comparison of different indices of responsiveness. J Clin Epidemiol 1997;50(3):239-246.