F. GALVÃO DE MELO

MODELOS E MÉTODOS ESTATÍSTICOS EM PSICOLOGIA

obras avulsas 1.oa

ESCOLA NACIONAL DE SAÚDE PÚBLICA

EDIÇÕES DE SAÚDE

F. GALVÃO DE MELO

MODELOS E MÉTODOS ESTATÍSTICOS EM PSICOLOGIA

1.oa 1

F. Galvão de Melo

MODELOS E MÉTODOS ESTATÍSTICOS EM PSICOLOGIA

Lisboa 1982

Nota prévia

O presente trabalho resultou do desenvolvimento de um seminário realizado pelo autor, no âmbito do Estudo Epidemiológico da Deficiência Mental, procurando-se, numa primeira parte, fundamentar epistemologicamente os métodos quantitativos commummente usados pelos psicólogos;numa segunda e terceira partes expõem-se algumas das técnicas matemático-estatísticas, cuja importância metodológica no papel da medição em psicologia é essencial, dando conteúdo a alguns dos modelos referidos na parte inicial.

Subentende-se para a leitura do texto conhecimentos básicos(elementares) da teoria da probabilidade e da inferência estatística, assim como rudimentos de cálculo vectorial e matricial.

Em apêndice incluem-se dois tópicos importantes para a compreensão de partes fundamentais do texto.

As referências bibliográficas, em fim de volume, são acompanhadas da indicação da parte (ou partes) do texto a que respeitam, assinalando-se o número do capítulo (C1, C2, C3) ou do apêndice (A II, A III).

F.Galvão de Melo

Cascais, Agosto 1982

ÍNDICE

1. O problema da medida em psicologia. Métodos quantitativos. Formulação de modelos

1.1. Introdução

Em sentido lato pode dizer-se que a psicologia(ou qualquer disciplina pré ou proto-científica) só adquire o estatuto de ciência experimental quando se decide abertamente pelo uso dos métodos científicos de medida (métodos quantitativos),entendendo-se aqui "medir" no sentido formal de "operação" regida por uma axiomática.

Por outro lado a história do pensamento científico torna incontroverso que "o aperfeiçoamento de uma ciência consiste em efectuar a transição do sentido descritivo para o sentido operatório. Daí resulta que o ponto de vista axiomático não interessa apenas à matemática, mas também aquelas ciências tradicionalmente experimentais que atingem um certo grau de elaboração" (1), ciências que N.Mouloud(2) designa por axiomático-experimentais. Isto aponta imediatamente para a formulação de modelos cobrindo não só os aspectos de medida mas também fornecendo esquemas formais de classificação e análise.

No entanto acerca da questão do uso de modelos (abstractos)os psicólogos parece estarem ainda divididos, uns rejeitando-os porque vêem neles um perigo de resvalamento para o formalismo e para o nominalismo metafísicos; outros pelo contrário aí encontrando a possibilidade da psicologia ascender ao "novo espírito científico" que Bachelard caracterizava pela fórmula "o real demonstra-se, não se mostra"(3).

Um modelo(*) é uma teoria hipotético-dedutiva, colecção de axiomas compatíveis, a partir dos quais é deduzível um conjunto de consequências —— teoremas ou proposições da teoria.

Havendo uma regra que permita traduzir as proposições da teoria em proposições acerca dos fenómenos reais (isomorfismo entre pensamento e realidade) diremos que as consequências lógicas do modelo podem

(1) CAVEING, M - O projecto racional das ciências contemporâneas , ob. cit. na bib., p. 175.
(2) MOULOUD, N. - Les structures, la recherche et le savoir, ob. cit. na bib.
(3) ORECO, P. - ob. cit. na bib., p.79.
(*) na sua forma mais acabada.

ser comparadas às observações, constituindo então estas uma realização empírica do modelo.

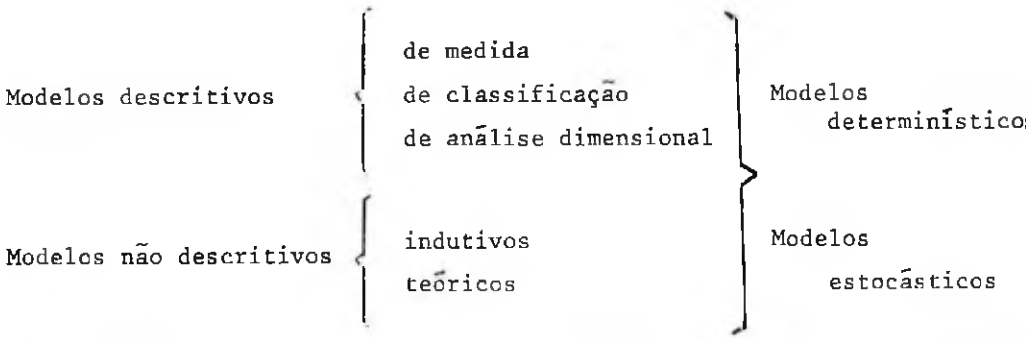
Também se diz nestas condições que o modelo é verificável.

Note-se que, do ponto de vista epistemológico, há dois níveis de problemas:

1. A adequação do modelo à realidade (verificabilidade)- nível pragmático.
2. A "veracidade" de uma teoria (consistência ou compatibilidade) - nível teorético.

Na prática científica é, em geral, mais importante saber que uma teoria esclarece a natureza de uma dada classe de fenómenos do que investigar a sua "veracidade". Assim é admissível a coexistência de modelos fornecendo explicações muito diferentes dum mesmo fenómeno:"a experiência revela indiscutivelmente que a luz, conforme o dispositivo experimental adoptado para o seu estudo, manifesta ora um carácter ondulatório(ondas luminosas), ora um caracter corpuscular (fluxo de fotões)".(4)

De um modo muito esquemático, aponte-se a seguinte tipologia dos modelos (5)



Modelos indutivos - visam explicar uma realidade observada, inferindo resultados válidos.

Modelos teóricos - estudam as consequências de um conjunto de axiomas traduzindo uma realidade dada.

Modelos descritivos - propõem-se classificar, ordenar e medir os objectos

(4) MARTINS, R. - notas aos "Elementos Metafísicos da Física" de H. Margenau , Coimbra, 1949, p. 91.
(5) Cf. R. Boudou, ob. cit. na bib.

(indivíduos) de uma dada população.

Aos modelos descritivos estão ligados os nomes de Spearman, Thurstone, Hotelling, Holzinger, Guttman, Lazarsfeld, etc.

Na maior parte dos casos estes modelos consistem em exprimir um conjunto de variáveis observáveis (notas, pontuações, êxito em provas, etc.) em função de variáveis de classificação inobserváveis, ditas variáveis latentes ou genotípicas.

Os modelos de Spearman e Thurstone são modelos estatísticos. Os de Guttman e de Coombs são algébricos ou combinatórios (modelos determinísticos). Estes modelos supõem que os dados se lhes adaptam, não de maneira estatística, mas de forma exacta.

Um modelo determinístico pode ser interpretado como exprimindo os estados mais prováveis de um processo estocástico associado, ou seja o modelo funcional que melhor se ajusta aos valores médios das variáveis do modelo estocástico.

O modelo determinístico aparece como uma "tradução" mais grosseira da situação real em estudo, do que o modelo estocástico correspondente. Neste sentido, a microfísica é fundamentalmente regida por leis estatísticas, enquanto as leis funcionais —— resultantes médias de leis estatísticas —— traduzem os fenómenos macrofísicos.

1.2. Problemas métricos em psicologia

Para se tentar definir uma teoria coerente da medida em psicologia é conveniente começar por analisar o estatuto epistemológico do conceito de medida, socorrendo-nos para isso das ciências da natureza, em particular da física, onde o seu uso, consagrado pela experiência, é incontestável.

O pensamento científico admite, hoje em dia, que "a ciência investiga os seus objectos, constrói-os e elabora-os, ela não os encontra completamente feitos ou dados na percepção ou na experiência imediata. O mundo da ciência é uma construção".(6)

Por outras palavras, o real não é dado mas construído. Ora a

(6) Jean Ullmo, ob. cit. na bib., pág.25.

própria "medição" é colaborante nesta construção, na medida em que "é a própria medição que define a grandeza a medir; esta não existe antes de ser medida como (...) o fazia crer uma intuição sumária".(7)

Na realidade medir é "interferir"; esta interferência altera a realidade que se pretende medir. O resultado da medição(*) aparece como a "medida" de um "objecto" resultante do "objecto com existência prévia ao acto de medir" e da acção do acto de medir sobre este.(**)

A oposição realismo - idealismo (dualismo matéria - espírito) dilui-se numa determinação recíproca das coisas pelo espírito - a construção dos objectos - e do espírito pelas coisas - a formação do conhecimento.(8)

Também a teoria da relatividade estabelece a não existência de duas realidades independentes - os instrumentos de medida e os fenómenos medidos - mas determinação recíproca do que mede e do que é medido."Não há um facto e um observador mas a reunião de ambos numa observação.(...) E o princípio de incerteza mostrou precisamente isto na física atómica: que o acontecimento e o observador não são separáveis".(9)

Será interessante lembrar que Planck considerou três mundos:o mundo sensível constituido pelas sensações, imagens,percepções,etc.,mundo que é alargado pelos instrumentos que o homem vai construíndo; o mundo real do qual não temos conhecimento directo, mas cuja existência é hipótese imprescindível de qualquer ciência, mundo que não é um simples aglomerado de coisas-em-si, mas um todo estruturado;e finalmente o mundo dos objectos físicos, construção intelectual do espírito científico que

---

(7) ULLMO, J. - ob. cit. na bib. p.25.

(*) Ele próprio um "objecto científico".

(**) Bohr e Heisenberg mostram que o efeito de interferência (sujeito-objecto; instrumento de observação-fenómeno observado) não pode ser eliminado, como o admitia a física clássica, representando a constante de Planck (o quantum de acção h) um limiar intransponível.

(8) Cf. A. Sérgio, prefácios e notas a "Três diálogos entre Hilas e Filonus" de J. Berkeley, Coimbra, 1958, e a "Os problemas da filosofia" de B. Russel, Coimbra, 1959.

(9) BRONOWSKI, J. - Introdução à atitude científica, Lisboa, 1972, p. 91.

se afasta progressivamente do mundo sensível para se adequar cada vez mais ao mundo real.(10)

B. Russel que no seu livro "Os problemas da filosofia" sustenta uma posição realista, admite posteriormente que "tanto o espírito como a matéria são simplesmente formas convenientes de agrupar acontecimentos" (11) e ainda que "a matéria deixou de ser uma "coisa", transformando-se simplesmente numa característica matemática das relações existentes entre estruturas lógicas complexas constituídas de acontecimentos" (12)

Recorde-se ainda que os trabalhos de Piaget, criando uma base experimental própria para a gnoseologia, estabelecem, por oposição às teses empirista e nominalista, que a leitura da experiência pressupõe no sujeito estruturas organizadoras anteriores à própria linguagem. Piaget conclui, de certo modo, uma linha que se inicia com Kant cujo esquematismo, embora fixo e inato, admitia já que os "conhecimentos à priori" têm um papel constitutivo da própria experiência, ligando-se, intimamente, a discussão da possibilidade dos juízos sintéticos transcendentais à metodologia Kanteana, que procura fundamentar a possibilidade das ciências puras e da natureza e ainda da metafísica em geral.(13)

Deste modo, privado dos "objectos absolutos" duma física ontológica primitiva põe-se ao físico o problema de saber o que mede. O "ser" como suporte de propriedades desaparece confundindo-se simplesmente com o conjunto dessas propriedades. Por outras palavras a existência do ser resulta unicamente da existência de um conjunto de propriedades objectivas, definidas por relações repetíveis.

A natureza da grandeza medida varia com o grau de aproximação da medida ou com a técnica de medição usada, podendo no primeiro caso, revelar novas qualidades capazes de modificar o objecto e no segundo apresentá-lo sob formas distintas, como no exemplo da luz citado, mostrando

---

(10) Max Planck, ob. cit. na bib.
(11) RUSSEL, B. - História da Filosofia Ocidental, 3º vol., S. Paulo, 1957, p. 408.
(12) RUSSEL, B. - Delineamentos da Filosofia, S. Paulo, 1956, p. 349.
(13) Cf. I. Kant, "Prolegómenos a toda a Metafísica futura", S. Paulo, 1959.

que as propriedades onda e corpúsculo não podem, sem contradição, figurar como atributos de um mesmo ser, de uma coisa-em-si.

O objecto de medida deixa de ser um dado no sentido próprio do termo, aparecendo como definido pela própria medida.

"É que na realidade, não há "dados" que não sejam construídos, num processo de produção em que a teoria tem um papel fundamental". (14) Mais precisamente o processo de medida e o objecto medido definem-se mutuamente, num processo dialéctico de aproximações sucessivas.

As dificuldades encontradas na medição dos objectos da física transportam-se e avolumam-se quando se passa para o campo da psicologia. Se a física, de certo modo, elimina a "quantidade"(*) que esteve na base do estabelecimento de relações, esta é, logo à partida, estranha ao objecto ou facto psicológico.

Uma sugestão para abordar o problema pode ser encontrada na atitude do físico para o qual são mais importantes as relações que as medidas permitem estabelecer do que as próprias medidas. Esta é na realidade a opção do psicólogo quando procura determinar a significação de um teste pelas suas relações com outras observações.

A via comum parece ser então a procura de "estruturas inteligíveis", por oposição a uma teoria ingénua do "realismo do objecto", a partir de isomorfismos entre as propriedades das observações experimentais e as propriedades das estruturas lógicas.

Uma teoria científica em geral (e da psicologia em particular) não pode ser interpretada como uma simples extensão dos factos experimentais, construída analiticamente, mas fundamentalmente como um processo de criação e de invenção cuja validação será feita à posteriori. Cabe assim à epistemologia actual um papel determinante no estudo da produção dos instrumentos intelectuais do conhecimento a par das operações de pro

---

(14) CASTELLS, M. - Prática epistemológica e ciências sociais, p. 25, ob. cit. na bib.

(*) É interessante notar que já em 1638 Descartes, em carta a Mersenne, escrevia: "Toute ma physique n'est autre chose que géométrie".

dução dos resultados do conhecimento.(15)

Decorre daqui que o problema da definição ou determinação prévia do objecto ou facto psicológico a medir perde a sua importância, transformando-se num falso problema, pois essa determinação — tal como na física (mais uma vez referida como ciência experimental paradigmática e não por pendor fisicalista) — não é requerida previamente ao acto de medir.

Deste modo parecem suprimidos os obstáculos com que a psicologia se defrontava podendo adquirir, por direito próprio, o estatuto de ciência experimental, embora num estádio pré-dedutivo.(*)

A diferença entre o "objecto físico" e o "objecto psicológico" esbate-se e desaparece no plano ontológico. Não há (não tem que haver) solução ontológica para o problema, mas sim solução epistemológica na medida em que os "níveis" ou "zonas de realidade" são de facto epistemológicos (e não ontológicos).

De acordo com Pierre Greco(16) o paralelismo (isomorfismo) entre o pensamento e a realidade (entre o modelo explicativo e a realidade a explicar) "não representa uma solução ontológica; é uma posição metodológica, e mais abertamente, uma hipótese epistemológica".

---

(15) Cf. Clara Dan - "Empirismo e realismo de Marx a Piaget", e M.Caveing, "O projecto racional das ciências contemporâneas", obs. cits.na bib.

(*) Admitimos uma lei do desenvolvimento das ciências que as faz passar numa ordem irreversível por quatro etapas sucessivas:descritiva,indutiva, dedutiva, axiomática, conferindo às ciências destintos graus de abstracção e racionalidade. (Cf. R. Blanché, obs. cits. na bib.)

(16) Ob. cit., pg. 108.

1.3. Níveis de medida; axiomáticas.

O uso tradicional da medida nas ciências experimentais é regido pela axiomática de Campbell(17), (conjunto das propriedades e regras operatórias dos números resultantes das medições) sendo problemática a sua verificação experimental no campo da psicologia.

Uma saída para o problema consiste em considerar o sistema dos números não com o conjunto de todas as suas propriedades, mas apenas com parte destas. O sistema repousa então sobre uma axiomática mais fraca para a qual se torna possível encontrar uma realização empírica ou modelo concreto.

Esta atitude conduz a considerar diversos níveis, cada um definido a partir das propriedades atribuídas aos números sobre os quais assentam os métodos de medida, ultrapassando-se a dicotomia de Campbell entre grandezas mensuráveis e grandezas não mensuráveis, substituída por diversos conceitos de mensurabilidade, desigualmente potentes.

Institui-se deste modo um processo dialéctico entre a evolução da psicologia versus uma disciplina experimental e o conceito de medida, regido por axiomáticas, evoluindo num sentido estrutural.

Até 1946 a medição em psicologia teve um carácter fundamentalmente empírico e intuitivo, não havendo uma reflexão crítica e epistemológica (pelo menos de carácter sistemático) sobre a validade dos métodos usados.

Isto não põe de modo algum em causa os progressos anteriores, realizados a partir do princípio do século, tal como, por exemplo, não foram postos em causa os progressos da mecânica quântica, devidos à intuição de Dirac, e anteriores à existência dum aparelho conceptual matemático apropriado. Aliás deste processo psicológico, a história da ciência apresenta inúmeros exemplos e, um caso típico, se verificou recentemente com a teoria das distribuições. Em artigo publicado na revista "Ciência" escreveu o Prof. Sebastião e Silva: "Durante cerca de meio século, electrotécnicos, físicos teóricos e alguns matemáticos usaram correntemente

---

(17) N. R. Campbell, ob. cit. na bib.

as distribuições (...) sem o saber. Até que, no momento oportuno, Laurent Schwartz num golpe de génio, soube congraçar todas essas intuições dispersas num corpo lógico e eficiente de doutrina. Mais uma vez a intuição, vaga e contraditória mas fecunda, cedeu lugar à ideia — lúcida, precisa, coerente. (18)

A solução teórica do problema inicia-se com S.S. Stevens enunciando uma teoria da medida na qual se define quatro tipos de escalas, de níveis hierarquizados: escalas nominais, ordinais, de intervalos e de razões.(19) Esta teoria é retomada e desenvolvida por Stevens no seu "Manual de Psicologia Experimental". (20)

Em rigor a maior parte das escalas usadas em psicologia são escalas ordinais , encontrando os psicólogos grandes dificuldades quando procuram ultrapassar este nível e atingir o das escalas de intervalos.

O uso da axiomática das escalas de intervalos, com dados sobre os quais apenas se pode definir empiricamente uma relação de ordem, levanta sérias dificuldades na atribuição de um conteúdo psicológico aos resultados obtidos pela elaboração aritmética dos números (dentro das operações permissíveis da axiomática).

Como os dados psicológicos apenas permitem, em geral, definir empiricamente uma relação de ordem, Coombs procura utilizar o maior número possível de modelos matemáticos fundados sobre esta relação: escalas de relação antissimétricas, transitivas, parcialmente ordenadas, etc.

Muitos destes modelos fundam-se na teoria dos reticulados. Generalizando os modelos unidimensionais utilizam-se reticulados n-dimensionais. Coombs propõe então o emprego de espaços vectoriais a) parcialmente ordenados e b) estrictamente ordenados.

Por exemplo, concebendo-se a inteligência como formada de várias aptidões mentais primárias distintas, um indivíduo A apenas pode ser considerado menos "inteligente" que um indivíduo B, num espaço vectorial

---

(18) SILVA,J.Sebastião - Como nasceu a teoria das distribuições, Ciência, nºs 15 - 16, Lisboa, 1959.

(19) On the theory of scales of measurement, ob. cit. na bib.

(20) Handbook of experimental psycology, ob cit. na bib.

(parcialmente) ordenado, quando A fôr inferior a B em cada uma das aptidões mentais primárias:

$$\vec{a} < \vec{b} \quad \text{se e só se} \quad a_i < b_i, \quad i=1, 2, \ldots, t$$

sendo $\vec{a}$ e $\vec{b}$ os vectores, cujas componentes são as t aptidões (mensuráveis) associados respectivamente a A e B. Utilizam-se ainda "modelos de distância" a partir da introdução duma escala métrica ordenada.

Debrucemo-nos então mais detalhadamente sobre os métodos permitindo a tradução numérica ou quantificação das observações, isto é, sobre os "métodos de medida".

Medir é introduzir uma aplicação φ definida no universo das observações, O, e com valores num conjunto numérico N: $O \xrightarrow{\varphi} N$

(Mais geralmente N pode ser um conjunto de natureza qualquer mas de estrutura bem determinada).

A escolha das regras definidoras de φ, isto é, de uma métrica, é em primeiro lugar um problema do psicólogo.

A escolha feita conferirá aos números certas propriedades que determinam por sua vez os métodos estatísticos cuja aplicação é válida. Como nota Reuchlin, nenhum tratamento estatístico pode conferir aos números propriedades que o psicólogo não lhes tenha atribuído na sua operação de medir.

Por este facto o estatístico não pode, muitas vezes, tirar dos números que o psicólogo lhe apresenta, conclusões, como o ilusionista tira coelhos do chapéu. (Aliás não se tiram coelhos do chapéu sem os lá ter metido previamente).

Escalas nominais ou categoriais (primeiro nível)

A escala é constituída por um conjunto de classes de equivalência. Não é um critério matemático ou estatístico que definirá a equivalência das observações, mas um critério empírico ligado à realidade em observação. Portanto um critério psicológico, definido pelo psicólogo. Este critério determinará o significado psicológico da "medida".

Um mesmo conjunto de observações pode ser repartido em classes de equivalência de diversos modos, usando critérios distintos. A escolha

do critério a usar depende inteiramente do problema que o psicólogo pretende resolver.(*)

Propriedades formais: = (igualdade, aqui com o significado de de relação de equivalência, isto é, uma relação binária, reflexiva, simétrica e transitiva).

Se uma escala nominal é constituída por n símbolos (ou nomes)(**), qualquer permutação dos n símbolos pode ser tomada para representar a escala, isto é, a escala é definida a menos das permutações do grupo simétrico de ordem n.

Para uma variável qualitativa (escala nominal) a moda (classe de maior frequência) é o único parâmetro de tendência central que se pode definir.

Relativamente à dispersão esta será tanto maior quanto maior fôr o número de classes. Para um número fixo de classes a dispersão é máxima quando a distribuição das observações pelas classes fôr uniforme. Nestas condições uma medida conveniente da dispersão é dada pela função entropia H. (21)

Escalas ordinais (segundo nível)

O psicólogo atribuirá um sentido à expressão "menor que"("maior que") utilizando para comparar duas observações, uma operação empírica satisfazendo formalmente às seguintes propriedades (definidas na matemática):

Irreflexibilidade

Antissimetria

Transitividade

definindo uma relação de ordem L

(*) Uma escala nominal é, na realidade, um esquema de classificação — — regra permitindo distribuir as observações por classes de equivalência.

(**) Cada símbolo ou nome designando uma classe de equivalência.

(21) Para a definição e propriedades da função H, V. por ex. F.Galvão de Melo, "Alguns aspectos da teoria da informação", Textos I.O.P. nº 1, Lisboa, 1974.

Estabelece-se assim um isomorfismo entre uma estrutura empírica e uma estrutura de ordem, projectando-se uma relação experimental em relação lógica.

Note-se que as propriedades a que L satisfaz (necessárias e suficientes para o matemático) são apenas necessárias para o psicólogo: a uma escala pode não ser atribuível nenhum significado psicológico.

Propriedades formais: $=, <$ $(>)$ $(<$- relação de ordem).

Uma escala ordinal é determinada a menos de uma transformação da classe das transformações monótonas (transformações preservando a ordem).

Numa escala ordinal, se a diversos elementos é atribuída a mesma pontuação, originando repetições de ordem, a ordenação não é única. Em muitos casos é desejável, do ponto de vista da análise estatística, uma única ordenação, tornando-se necessárias medições cuidadosas de maneira a minimizar o número de repetições.

No caso ordinal, o parâmetro de tendência central que "melhor" representa a distribuição observada é a mediana. A entropia H ainda é uma medida conveniente de dispersão.

Uma ideia mais precisa da distribuição pode ser obtida a partir dos quantis, permitindo situar, com o grau de aproximação desejável, uma observação relativamente à distribuição total.

A quantilagem é, no caso dos testes, um método de aferimento.

Escalas de intervalos (terceiro nível)

Introduzindo o psicólogo uma operação empírica (respeitante, portanto, aos factos e não aos números) que lhe permita definir a igualdade de intervalos — operação de equipartição (22) — poderá em colaboração com o matemático (estatístico) construir uma escala de intervalos.

A escala de intervalos envolve o conceito de unidade de distância, podendo a distância entre dois pontos da escala ser expressa em função da unidade (arbitrária) escolhida.

---

(22) V. H. Coombs, M.Dawes, A.Tversky, ob. cit. na bib.

Propriedades formais: $=, <, (>)$, $b - a = c - d$ (a, b, c, d pontos da escala)

A classe das transformações que preservam estas propriedades é a classe das transformações afins. Por outras palavras, a escala de intervalos é determinada a menos de uma transformação afim.(*)

A escala de intervalos é a primeira com significado "quantitativo". As estatísticas (paramétricas) habituais — médias, variâncias, coeficientes de correlação, etc. — podem ser calculadas, sendo igualmente legítima a aplicação dos testes paramétricos — teste t, teste F, etc.

A operação de equipartição mencionada releva de uma convenção de linguagem na medida em que a "igualdade" de dois intervalos é decidida na base de um determinado processo. As escalas assim construídas apenas são válidas no âmbito desta convenção de linguagem.

Pode-se, por exemplo, convencionar denominar "iguais" os intervalos inter-decis o que equivale a postular uma distribuição uniforme das observações sobre a escala. Noutros casos o psicólogo pode tomar a decisão de definir a igualdade de intervalos por um postulado sobre a forma da distribuição, optando por uma distribuição binomial, ou por uma distribuição gaussiana, em lugar da distribuição uniforme.(23)

As leis binomial e normal (lei de Gauss ou de Laplace-Gauss) constituem modelos privilegiados, e consagrados pelo uso, da distribuição de variáveis referidas a escalas de intervalos.

As razões deste facto — e contrariamente ao que a tradição perpetua — são de ordem heurística e pragmática.

De facto, 1. O psicólogo obtém muitas vezes conjuntos de observações cujo histograma sugere um perfil binomial, sabendo-se da estatística matemática que a distribuição binomial tem, sob certas condições, como distribuição limite uma distribuição

---

(*) I. é, transformações do tipo $a' = \alpha a + \beta$ com α e β constantes.

(23) Cf. M. Reuchlin, "Précis de Statistiques", ob. cit. na bib. e "Notions de Statistique", D.E.U.G. de psychologie, Université R.Descartes (texto policopiado).

de Laplace-Gauss. É então natural adoptar um modelo Gaussiano.(*)

2. As leis binomial e normal têm propriedades formais convenientes, completamente estudadas e de uso simples.

No entanto, o psicólogo e o estatístico deverão evitar o duplo erro de crer, o primeiro, que o uso das distribuições gaussianas se impõe por razões fundadas teoricamente no campo da matemática, e o segundo, que as leis normais se impõem por resultarem experimentalmente das observações empíricas.

Escalas de razões (quarto nível)

Numa escala de razões não só a amplitude dos intervalos tem significado como igualmente o tem a razão entre duas medições. Neste caso, a origem da escala não é arbitrária, como sucedia na escala de intervalos. Isto é, o ponto zero (origem) da escala tem significado absoluto, mantendo-se a unidade de distância arbitrária.

Propriedades formais: $=, <\ (>)$, $b - a = c - d$, $a = kb$.

A classe das transformações admissíveis (isto é, que preservam as propriedades formais) é a classe das homotetias directas (multiplicação por uma constante positiva).

A escala de razões permite o uso de estatísticas implicando a existência de origem absoluta, como a média geométrica e o coeficiente de variação.

(*) Note-se que, no caso da construção de um teste, por exemplo, o psicólogo por escolha conveniente dos items e das notações pode obter histogramas com prefis fixados à priori.

2. Medidas (do grau) de associação entre duas variáveis

2.1. Introdução

A teoria clássica da regressão e correlação pressupõe variáveis referidas, pelo menos, a escalas de intervalos. Este facto parece, à partida, criar grandes limitações na aplicação dos modelos clássicos às variáveis psicológicas, na medida em que estas não ultrapassam, frequentemente, o nível ordinal.

No entanto a estatística oferece ao psicólogo uma gama de métodos, adaptados aos diferentes tipos de escalas, cobrindo a maior parte das situações postas pela prática.

Os modelos de correlação vão permitir a pesquiza de relações a partir da comparação — por meio de testes estatísticos — das observações experimentais com valores teóricos calculados sob a hipótese de independência (ausência de relação entre as variáveis).

O psicólogo concluirá pela existência de relação (variáveis correlacionadas) se admitir que as observações empíricas se afastam de modo "significativo" dos valores teóricos esperados, na hipótese de independência, o que releva, uma vez mais, de uma convenção, da fixação arbitrária de um limiar.

Resta finalmente, ao psicólogo, decidir (com a eventual ajuda do estatístico) se a relação experimental encontrada pode ser considerada forte ou fraca, importante ou desprezável, no âmbito concreto do problema em estudo. Note-se que a análise estatística é, em geral, impotente para resolver os problemas na sua essência: o estatístico colabora com o economista, o médico, o psicólogo, etc., mas é de acordo com a sua técnica e no âmbito dos seus campos de investigação que estes devem encontrar a explicação dos factos observados.

2.2. O modelo clássico da correlação

Seja (X, Y) um par aleatório com uma distribuição conjunta binormal (distribuição de Gauss a duas variáveis).

Na distribuição do par (X, Y) definem-se distribuições marginais (ou totais) e distribuições condicionais (ou parciais). A distribui

ção marginal de X é a distribuição dos valores x. Analogamente para Y. A distribuição condicional de Y, para um valor fixado x de X, é a distribuição dos valores y correspondentes a este valor de X. Igualmente se definem as distribuições condicionais de X.

Em particular a distribuição binormal é definida por uma superfície de frequência cuja equação depende de 5 parâmetros:

- os valores médios $\mu_1 = E(X)$ e $\mu_2 = E(Y)$ e as variâncias $\sigma_1^2 = E(X-\mu_1)^2$, $\sigma_2^2 = E(Y-\mu_2)^2$, das distribuições marginais de X e Y respectivamente.
- o parâmetro $\rho = (\sigma_1 \sigma_2)^{-1} \operatorname{cov}(X, Y)$, compreendido entre -1 e +1, que toma o nome de <u>coeficiente de correlação</u> do par (X,Y)

Nestas condições prova-se que

1. Fixado X = x a distribuição dos valores de Y correspondentes —— distribuição (parcial) de Y condicionada por x —— é ainda uma distribuição normal. Por outras palavras, todas as distribuições condicionais são normais.

2. Quando x varia os valores médios das distribuições condicionais de Y descrevem uma recta —— <u>recta de regressão de Y em X</u> (*).
A equação desta recta pode escrever-se na forma

$$y = E(Y \mid x) = \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x - \mu_1)$$

O coeficiente angular da recta, $\beta = \rho \frac{\sigma_2}{\sigma_1}$, é o <u>coeficiente de regressão teórico</u> de Y em X.

3. As distribuições condicionais de Y têm todas a mesma variância σ^2, independente, portanto, do valor de X fixado. σ^2 toma o nome de <u>variância condicional (ou residual)</u> e tem-se

$$\sigma^2 = \sigma_2^2 (1 - \rho^2) \quad (\sigma = \sigma_2 \sqrt{1 - \rho^2})$$

sendo σ_2^2 a variância (total) de Y.

(*) No caso normal mostra-se que esta recta coincide com a recta de regressão dos mínimos quadrados de Y e X.

4. Mutatis mutandis 1.,2. e 3. mantêm-se válidas para a distribuição condicional de X fixado Y=y.

O coeficiente de correlação ρ é uma medida do grau de linearidade da relação estocástica entre X e Y. Uma estimativa de ρ, calculada a partir de n pares de observações, $(x_1, y_1),\ldots,(x_n, y_n)$, do par (X,Y) é dada pelo coeficiente de correlação linear empírico ou coeficiente de correlação de Pearson, definido por

$$r = \frac{\sum_i (x_i-\bar{x})\ (y_i-\bar{y})}{\sqrt{\sum_i (x_i-\bar{x})^2 . \sum_i (y_i-\bar{y})^2}} \qquad (2.1)$$

Não havendo qualquer relação entre as variáveis X e Y, na população, é $\rho = 0$. Nestas condições é de esperar que um valor calculado de r não se afaste muito de zero, sendo o valor $|r| > 0$ explicado pelas variações aleatórias de amostragem.

Com $|r| >> 0$ deve-se rejeitar a hipótese de que as observações provenham de uma população com $\rho=0$, dizendo-se neste caso que o valor experimental r obtido é significativo.

Com r significativo não há, como se afirmou, nenhum critério permitindo classificar, de modo absoluto, a correlação em "forte" ou "fraca". O psicólogo decidirá se a correlação observada é suficientemente "forte" para apresentar um interesse prático ou teórico.

De um modo geral, uma correlação linear será considerada tanto mais "forte" ou "útil", quanto mais fraca fôr a variância residual(*)sendo, nestas condições, a recta de regressão um bom resumo estatístico das observações.

O coeficiente r é uma medida empírica da relação linear (estocástica) entre as variáveis X e Y —— medida do "grau de dependência" entre as duas variáveis. Nestas condições um valor de r próximo de zero si-

(*) Como se conclui da relação $\sigma^2 = \sigma_2^2(1-\rho^2)$ ou sua correspondente empírica $s^2 = s_2^2(1-r^2)$.

gnifica que é quase certa a inexistência de relação linear.

No entanto é possível a existência de uma correlação não linear ligando as variáveis.

Por outro lado r pode estar próximo de -1 ou +1 e não haver dependência (no contexto material em que se insere o significado das variáveis) entre X e Y, dando origem a correlações espúrias ou sem sentido.

Considere-se em geral o caso de um par aleatório contínuo (*) com uma distribuição arbitrária, admitindo momentos de 1.ª e 2.ª ordem. Neste caso a curva de regressão de Y em X,

$$y = E(Y|x)$$

—lugar geométrico dos valores médios condicionais de Y — não é, em geral, uma recta.

No caso binormal o coeficiente de correlação teórico ρ pode ser interpretado como uma medida da tendência da distribuição de (X, Y) se acumular em torno da recta de regressão de Y em X (ou de X em Y).

Uma medida da tendência da distribuição (não gaussiana) do par (X,Y) se concentrar em torno da curva de regressão $y=E(Y|x)$ é dada pela razão de correlação ζ_{yx}, definida por (24)

$$\zeta^2_{yx} = \frac{v\left[E(Y|X)\right]}{v(Y)} \qquad (2.2)$$

O numerador $v\left[E(Y|X)\right]$ — variância dos valores médios condicionais — representa a variância devida à regressão ou variância inter-distribuições.

Como

$$v(Y) = E\left[v(Y|X)\right] + v\left[E(Y|X)\right] \qquad (2.3)$$

(i. é, a variância de Y é igual ao valor médio das variâncias condicionais mais a variância dos valores médios condicionais)

(*) As variáveis aleatórias contínuas servem de modelo a distribuições reais caracterizadas por variáveis tomando um grande número de valores e com pequenos intervalos entre valores consecutivos da variável.

(24) Alguns autores definem a razão de correlação por $\zeta^2_{yx} = \frac{E\left[v(Y|X)\right]}{v(Y)}$
V.S.S. Wilks, "Mathematical Statistics", J.Wiley, pg.86, 1962.

(2.2) ainda se pode escrever

$$\zeta^2_{yx} = 1 - \frac{E[v(Y|X)]}{v(Y)} = 1 - \frac{\sigma_r^2}{\sigma_2^2} \qquad (2.4)$$

sendo $E\left[v(Y|X)\right] = E\left[Y-E(Y|X)\right]^2 = \sigma_r^2$ a variância residual ou variância intra-distribuições. (A variância residual é o que fica da variância (total) de Y depois de subtraída a variância devida à regressão).

Demonstra-se que

$$\zeta^2_{yx} = \rho^2 + \frac{1}{\sigma_2^2} E\left[E(Y|X) - \alpha - \beta X\right]^2 \qquad (2.5)$$

sendo $y = \alpha + \beta x = \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x-\mu_1)$

a recta de regressão dos mínimos quadrados —— regressão do tipo II (25) —— de Y em X.

De (2.5) segue-se que $\zeta^2_{yx} = 0$ se e só se E(Y|X) é independente de X. De facto, nestas condições, a curva de regressão y=E(Y|x) é uma recta paralela ao eixo dos xx e o anulamento de β implica o anulamento de ρ. Por outro lado (2.4) mostra que $\zeta^2_{yx} = 1$ se e só se todos os pontos da distribuição estão situados na curva de regressão y = E(Y|x), havendo portanto uma relação funcional perfeita entre X e Y.

A razão da correlação pode ser considerada como uma medida da tendência da distribuição (X,Y) se acumular em torno da curva de regressão. Por outras palavras ζ^2_{yx} indica em que medida a curva de regressão y =E(Y|x) pode ser usada para predizer Y conhecido X.

Quando a regressão de Y em X é linear (y= E(Y | x) é uma recta e coincide com a recta dos mínimos quadrados) (2.5) mostra que $\zeta^2_{yx} = \rho^2$. Neste caso o cálculo de ζ^2_{yx} não dá qualquer informação, desde que ρ seja conhecido.

No caso geral de regressão não linear (2.5) mostra que ζ^2_{yx} ex-

---

(25) V. F.Galvão de Melo, ob. cit. na bib., pág.126.

cede ρ^2 por uma quantidade que mede o desvio entre a curva de regressão, $y = E(Y \mid x)$, e a recta dos mínimos quadrados, $y = \alpha + \beta x$.

A razão de correlação ζ^2_{yx} de X em Y define-se trocando os papéis das variáveis nas relações indicadas.

Note-se que, as definições formais de ρ e ζ^2_{yx} ,

$$\rho = \frac{cov\ (X,Y)}{\sqrt{v(X)v(Y)}} \quad , \quad \zeta^2_{yx} = 1 - \frac{E[v(Y|X)]}{v(Y)}$$

são válidas para qualquer tipo de variáveis, desde que existam os momentos de 1.ª e 2.ª ordem; ρ e ζ_{yx} aparecem como parâmetros caracterizando a distribuição conjunta de variáveis X e Y, quer discretas (*), quer contínuas. Aliás, como veremos, ζ_{yx} vai-se revelar particularmente útil no estudo da relação entre uma variável contínua e uma variável discreta, mais concretamente entre uma escala de intervalos e uma escala categorial.

Posto isto passemos ao estudo de algumas das principais soluções particulares de que o psicólogo dispõe, soluções que se harmonizam com os diferentes tipos de escalas a que as suas medições ou observações se referem.

2.3. Estudo da correlação entre uma variável dicotómica (ou dicotomisada) e uma variável contínua

a) Correlação biserial pontual

O coeficiente de correlação empírica r é o melhor estimador do coeficiente de correlação normal ρ . No entanto, na prática, muitas vezes não se dispõem de valores experimentais em condições de efectuar o cálculo de r.

(*) Uma variável X diz-se discreta se o seu domínio ou campo de variação, D, (conjunto dos valores admissíveis de X) é finito ou numerável. Por outras palavras, se existe uma correspondência biunívoca entre D e um subconjunto dos naturais. Note-se que variáveis nominais e ordinais são sempre discretas.

Considere-se o caso em que, para cada indivíduo de uma amostra de dimensão n, o par (x,y) é determinado nas condições:

- conhecem-se os valores de X, valores observados de uma variável contínua, referida a uma escala de intervalos ou de razões, cuja distribuição se admite gaussiana.

- a variável Y, em geral discreta ou nominal é dicotômica (ou dicotomisada), isto é, cada valor observado, y, pertence a uma de duas classes complementares, designadas, por exemplo por classe inferior e classe superior.

Nestas condições define-se um coeficiente de correlação que toma o nome de correlação biserial pontual e é usado, como medida da correlação entre as pontuações de um teste (variável X contínua) e as pontuações de um item do teste (variável dicotômica: sucesso-insucesso).

A variável dicotômica pode ser o sexo, a divisão em dois grupos etários, a partição de um grupo de indivíduos a partir da existência ou não de um atributo A (doentes-não doentes, normais-neuróticos, etc.).

O coeficiente biserial pontual é definido por

$$r_{bp} = \frac{\bar{x}_p - \bar{x}_q}{s_x}\sqrt{pq} \qquad (2.6)$$

sendo X a variável contínua, s_x o desvio padrão empírico de X, p,q as proporções de indivíduos em cada uma das categorias da variável dicotômica Y (no caso do item de um teste, p é a proporção de indivíduos com "sucesso" no item e q a proporção de "insucessos"), $\bar{x}_p$, $\bar{x}_q$ as médias das pontuações (valores) da variável contínua em cada uma das categorias determinadas pela variável dicotômica.

Atendendo a que $\bar{x}_q = \frac{1}{q}(\bar{x} - p\bar{x}_p)$ ainda se pode escrever

$$r_{bq} = \frac{\bar{x}_p - \bar{x}}{s_x}\sqrt{\frac{p}{q}}$$

Exemplo 1. A seguinte tabela relaciona as pontuações de um teste aplicado a 12 indivíduos e as pontuações de um dado item.

Indivíduos	1	2	3	4	5	6	7	8	9	10	11	12	
Pontuações teste	5	7	7	10	15	24	26	30	30	36	41	45	$\Sigma x=276$
Pontuações item	0	0	0	0	1	0	0	1	0	1	1	1	$\Sigma y=5$

$$\bar{x}_p = \frac{15 + 30 + 36 + 41 + 45}{5} = 33,40 \quad ; \quad p = \frac{5}{12} = 0,42$$

$$\bar{x}_q = \frac{5 + 7 + 7 + 10 + 24 + 26 + 30}{7} = 15,57 \quad ; \quad q = \frac{7}{12} = 0,58$$

$$\bar{x} = \frac{276}{12} = 23 \qquad \text{(Note-se que } \bar{x} = p\bar{x}_p + q\bar{x}_q \text{)}$$

$$s_x = \left(\frac{1}{n} \Sigma x_i^2 - \bar{x}^2 \right)^{1/2} = 13,40$$

e finalmente $$r_{bp} = \frac{33,40 - 15,57}{13,40} \sqrt{0,42 \times 0,58} = 0,65$$

O valor calculado, 0,65, dá uma medida da capacidade discriminativa do teste relativamente aos dois grupos determinados pela dicotomia da variável Y (medida do grau em que a variável contínua diferencia ou discrimina entre as duas categorias da variável dicotômica).

b) Correlação biserial

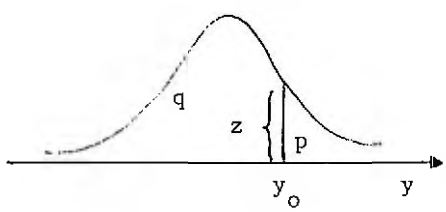
A correlação biserial é uma estimativa do coeficiente de correlação ρ entre X e Y, admitindo-se que a variável X é gaussiana e a dicotomisação foi feita a partir de uma variável Y igualmente gaussiana. Por outras palavras, dado um par binormal (X,Y) dicotomisa-se a distribuição dos yy, suposta reduzida, agrupando numa classe todos os valores de y in

feriores a um dado valor y_o, sendo a outra a classe dos valores $y > y_o$.

Assim, se numa tabela de correlação k x ℓ de um par binormal, se dicotomisa uma das variáveis passando-se para uma tabela k x 2, calcule-se em lugar do r de Pearson o coeficiente r_b , cuja expressão é

$$r_b = \frac{\bar{x}_p - \bar{x}_q}{s_x} \frac{pq}{z} = \frac{\bar{x}_p - \bar{x}}{s_x} \frac{p}{z} \qquad (2.7)$$

sendo z a ordenada da intersecção da curva normal reduzida com a vertical, passando por y_o , dividindo a área sob a curva nas proporções q e p



(2.7) representa uma simplificação algébrica do r de Pearson, quando Y apenas toma os valores 0 e 1 (26)

Para n > 50 e p ≈ 0,50 um valor aproximado do desvio padrão de r_b é dado por

$$\frac{\frac{\sqrt{pq}}{z} - r_b^2}{\sqrt{n}}$$

Nas hipóteses indicadas, r_b é uma melhor estimativa de ρ_{xy} do que r_{bp} .

Os valores $\frac{pq}{z}$ e $\frac{\sqrt{pq}}{z}$ são lidos em tabelas a partir do conhecimento de p ou q (27).

Para testar a significância de r_b (ou r_{bp}) pode-se usar a estatística

$$t = \sqrt{n-2} \frac{r_b}{\sqrt{1-r_b^2}}$$

que, na hipótese $\rho = 0$, tem aproximadamente uma distribuição de Student

(26) V Lord e Novick, ob cit. na bib. págs. 337-339.
(27) J.Faverge, ob. cit. vol.II, pág.194.

com n-2 graus de liberdade.

Note-se que $r_b > r_{bp}$, mais precisamente, comparando (2.6) e (2.7)

$$r_b = r_{bp} \frac{\sqrt{pq}}{z} \qquad (27)$$

Exemplo 2. (Faverge, ob. cit.)

Paralelamente a um exame, foi aplicado a 113 indivíduos um teste, pretendendo-se estudar a correlação do teste com os resultados do exame. O teste é pontuado de 0 a 10 tendo-se dicotomisado a escala das notas do exame, suposta normal, nas categorias aprovado, reprovado.

Os resultados obtidos resumem-se na tabela de contingência 2x10

Exame (Y) \ Teste (X)	0	1	2	3	4	5	6	7	8	9	10	Total
Aprovado	0	1	0	3	5	10	16	14	6	3	1	59
Reprov.	1	2	4	7	13	12	9	4	2	0	0	54
Total	1	3	4	10	18	22	25	18	8	3	1	113

p = proporção de aprovados $= \frac{59}{113} = 0,522$

q = proporção de reprovados $= \frac{54}{113} = 0,478$

$$\bar{x}_p = \frac{1 + 9 + 20 + 50 + 96 + 98 + 48 + 27 + 10}{59} = 6,0847$$

$$\bar{x}_q = \frac{2 + 8 + 21 + 52 + 60 + 54 + 28 + 16}{54} = 4,4630$$

$s^2_x = 3,5570$; $s_x = 1,886$; $\frac{pq}{z} = 0,6264$

vindo finalmente

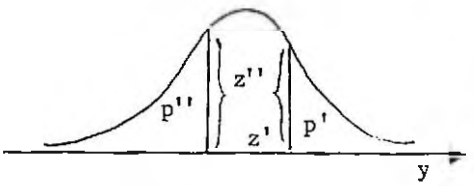
$$r_b = \frac{6,0847 - 4,4630}{1,886} \times 0,6264 = 0,54$$

O cálculo de r_{bp} para estes dados daria um valor inferior a 0,54, estimativa, em princípio, menos precisa de ρ, o que se explica por no cálculo de r_{bp} não se ter entrado com toda a informação contida na amostra dos 113 pares de pontuações.

2.4. Correlação entre uma variável tricotômica e uma variável contínua

A partir do par binormal (X,Y) divida-se o campo de variação de Y em 3 classes — superior, média e inferior.

Sejam p' e p" as proporções de indivíduos nas classes superior e inferior ($p' + p'' < 1$), $\bar{x}_{p'}$, $\bar{x}_{p''}$ as médias dos valores de X correspondentes às duas classes indicadas, z', z'' as ordenadas da curva normal reduzida correspondentes à partição em 3 classes.



Nestas condições uma estimativa de ρ é dada pelo coeficiente de correlação triserial de Burt:

$$r_{tris} = \frac{\bar{x}_{p'} - \bar{x}_{p''}}{s_x} \cdot \frac{1}{\frac{z'}{p'} + \frac{z''}{p''}} \qquad (2.8)$$

lendo-se em tabelas os valores z'/p' , z"/p" , conhecidos p' e p" (28)

Coeficiente de correlação para variáveis dicotomisadas

Supondo ainda (X,Y) binormal, uma estimativa de ρ calculada a partir do agrupamento em duas classes para cada uma das variáveis — coeficiente de correlação tetracórico — foi obtida em 1900 por K.Pearson(29) não havendo no entanto fórmula geral simples para o seu cálculo (30).

O coeficiente tetracórico, r_t é a melhor aproximação do r de Pearson, valor que se obteria se fosse possível obter medições (observações) mais precisas e completas de X e Y.

O estudo das correlações item-item pode ser feito a partir de r_t, supondo normais as distribuições subjacentes às dicotomisações.

---

(28) Faverge, ob. cit., pág. 194.
(29) "On the correlation of characters not quantitatively measurable", Royal Society Philosofical Transactions, Series A, 1900.
(30) V.Q. McNemar, ob.cit na bib. pp. 221-225.

A partir da tabela 2x2

X \ Y	β_1	β_2
α_1	a	b
α_2	c	d

(α_i, β_i categorias determinadas pela dicotomisação)

uma aproximação de r_t é dada por

$$r_t \approx \cos \frac{\pi}{1+\sqrt{\frac{ad}{bc}}}$$

estando tabelado o valor r_t em função do quociente $\frac{ad}{bc}$ (ou $\frac{bc}{ad}$) (31)

Na prática usam-se, em geral, os ábacos de Bonnardel (32) permitindo de modo expedito a determinação gráfica de r_t.

2.5. Correlação para variáveis tricotomisadas

O coeficiente eneacórico de Coumetou resulta do cálculo do r de Pearson, sobre um agrupamento grosseiro, em três classes, das observações relativas a cada variável.

O cálculo deste coeficiente apoia-se sobre várias hipóteses simplificativas o que torna o seu interesse bastante restrito. De um modo geral, para tabelas k x k ($k \geqslant 2$), obtidas por agrupamento das observações de um par (X,Y) binormal, definem-se coeficientes policóricos.

2.6. Relação entre uma escala nominal e uma escala de intervalos

Sejs S uma amostra de n indivíduos, de uma população $\mathcal{U}$, e $y_1,\ldots y_n$ os valores de uma característica mensurável Y sobre S, constituindo a distribuição empírica total de Y, de parâmetros $\bar{y}$ e s^2. Classifiquem-se os n indivíduos de S de acordo com os k valores (atributos),

(31) V.Glass e Stanley, ob. cit. na bib., tab. H, pág.535.
(32) R.Bonnardel,"Abaques pour la détermination du coeficient de corrélation tétrachorique", Éditions Scientifiques et Psychotechniques.

$\alpha_1, \alpha_2, \ldots, \alpha_k$, de uma variável qualitativa (nominal) X, obtendo-se assim uma partição de S em k categorias:

$$S = A_1 \cup A_2 \cup \ldots \cup A_k \ , \quad A_i \cap A_j = \emptyset \ , \quad i \neq j$$

sendo A_i o conjunto dos indivíduos de S gozando do atributo α_i. A cada A_i está associada uma distribuição condicional (ou parcial) — distribuição da variável condicional $Y \mid \alpha_i$, restrição de Y a A_i, de parâmetros $\bar{y}_i$, s_i^2.

Sendo n_i a dimensão de A_i, (2.3) toma o aspecto

$$s^2 = \frac{1}{n} \sum_{i=1}^{k} n_i s_i^2 + \frac{1}{n} \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2 = s_r^2 + \frac{1}{n} \sum_i n_i (\bar{y}_i - \bar{y})2$$

sendo s_r^2 a variância residual empírica.

A distribuição de Y sobre S está afectada por duas fontes de variação:

1. Diferenças sistemáticas de categoria para categoria, expressas por $\bar{y}_1 \neq \bar{y}_2 \neq \ldots \neq \bar{y}_k$.

2. Variabilidade aleatória dentro de cada categoria A_i, medida pela variância condicional s_i^2 (*)

A importância da relação entre a escala nominal $\{\alpha_1, \ldots, \alpha_k\}$ e a escala de intervalos pode ser medida usando a razão de correlação empírica η^*, definida por

$$\eta^{*2} = 1 - \frac{s_r^2}{s^2} = \frac{1}{s^2 n} \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2 \qquad (2.9)$$

(*) Admitindo a homogeneidade das variâncias parciais (homoscedasticidade) nas subpopulações de $\mathcal{U}$ determinadas por X, tem-se que s_r^2 é uma estimativa do valor comum das variâncias teóricas condicionais. A hipótese da igualdade dos valores médios das subpopulações de $\mathcal{U}$ pode ser testada pela técnica clássica da análise de variância simples. Em particular com k=2 cai-se no teste "t" de Student.

Sendo X e Y independentes, $\bar{y}_1, \ldots, \bar{y}_k$ devem ser aproximadamente iguais entre si e iguais a $\bar{y}$ (nestas condições cada A_i representa, tal como S, uma amostra aleatória de $\mathcal{U}$) o que implica $s_r^2 \simeq s^2$ e portanto $\eta^{*2} \simeq 0$.

Pelo contrário, se as categorias α_i determinarem univocamente o valor de Y vem $s_i^2 \simeq 0$, $s_r^2 \simeq 0$ e $\eta^{*2} \simeq 1$

Nos casos intermédios, em que as duas fontes de variabilidade intervêm, o valor de η^{*2} está compreendido entre 0 e 1, medindo η^* a importância da relação entre as duas escalas.

Exemplo 3.

A n indivíduos de uma população escolar aplicou-se um teste (Y) tendo-se em seguida analisado os resultados por sexo (X).

Obtém-se assim três distribuições

a) da variável Y de parâmetros empíricos $\bar{y}$ e s^2

b) da variável (condicional) Y | H de parâmetros empíricos $\bar{y}_H$, s_H^2

c) da variável (condicional) Y | M de parâmetros empíricos $\bar{y}_M$, s_M^2

Sendo $\bar{y}_H - \bar{y}_M$ significativamente diferente de zero, conclui-se que as pontuações Y estão relacionadas com o sexo (categorias H e M). A importância da relação é medida por

$$\eta^{*2} = 1 - \frac{s_r^2}{s_2^2}$$

com

$$s_r^2 = \frac{n_H s_H^2 + n_M s_M^2}{n} \qquad (n_H + n_M = n)$$

Nota: o teste "t" de Student permite testar a significância da diferença $\bar{y}_H - \bar{y}_M$, desde que as duas seguintes condições sejam satisfeitas

I) Y | H e Y|M obedecem a leis (aproximadamente) gaussianas.

II) $Y|H$ e $Y|M$ têm a mesma variância

Em geral, a condição II) não é evidente sendo então necessário testar a hipótese $v(Y|H) = v(Y|M)$.

No entanto para amostras de dimensões aproximadamente iguais a desigualdade das variâncias teóricas tem consequências relativamente pouco importantes nas conclusões derivadas da aplicação do teste "t". Assim, em caso de dúvida sobre a validade da condição II), use-se o teste "t" com amostras de dimensões sensivelmente iguais.

Com amostras de dimensões muito díspares corre-se o risco das conclusões serem seriamente afectadas.

2.7. Correlação entre duas variáveis ordinais

a) Coeficiente de correlação ordinal de Spearman

Quando as observações feitas não podem ser expressas por valores de uma escala de intervalos, admitindo unicamente uma ordenação (ou hierarquização), o coeficiente de correlação empírico de Pearson, definido por (2.1), dá origem ao coeficiente de correlação ordinal de Spearman.

$$r_s = 1 - \frac{6 \sum_i d_i^2}{n (n^2 - 1)} \qquad (2.10)$$

com $d_i = x_i - y_i$, x_i e y_i valores de uma escala ordinal, em geral a escala natural 1, 2, 3, ...

Para valores $n \geqslant 10$ a significância de r_s pode testar-se a partir da variável t de Student. Para amostras pequenas ($n<10$) veja-se S. Siegel , ob. cit. na bib., págs. 210-211 e tab. P do apêndice.

Exemplo 4. Dez estudantes, designados por A, B, ..., J, foram ordenados segundo o seu aproveitamento em trabalhos de laboratório e conhecimentos teóricos, obtendo-se o seguinte quadro,

	A	B	C	D	E	F	G	H	I	J	
Laboratório (X)	8	3	9	2	7	10	4	6	1	5	
Teoria (Y)	9	5	10	1	8	7	3	4	2	6	
d = X - Y	-1	-2	-1	1	-1	3	1	2	-1	-1	
d^2	1	4	1	1	1	9	1	4	1	1	$\Sigma d^2 = 24$

dando o cálculo de r_s o valor

$$r_s = 1 - \frac{6 \times 24}{10\,(10^2 - 1)} = 0{,}854$$

Para testar a significância de r_s, ao nível de 5%, calcule-se

$$t = \sqrt{n-2}\,.\,\frac{r_s}{\sqrt{1 - r_s^2}} = \sqrt{8}\,.\,\frac{0{,}854}{\sqrt{0{,}271}} = 4{,}64$$

Para um teste unilateral o quantil de ordem 0,95 da distribuição de Student com 8 graus de liberdade é 1,83. Como $4{,}64 > 1{,}83$, rejeita-se a hipótese nula — inexistência de ligação entre as variáveis — concluindo-se que r_s é significativamente diferente de zero.

b) Coeficiente de correlação ordinal de Kendall

Nas condições precedentes uma solução alternativa é dada pelo coeficiente τ de Kendall, medindo o grau de concordância entre as duas ordenações. A introdução deste coeficiente exige algumas definições prévias. Seja A um conjunto de n objectos, numerados de 1 a n, e π_1 e π_2 duas ordenações de A. Por outras palavras, π_1 e π_2 são duas permutações de $\pi_0 = (1, 2, \ldots, n)$.

Como se sabe, com n objectos distintos podem formar-se n! permutações de ordem n. Tome-se uma destas permutações para permutação de referên

cia ou principal, por exemplo, a permutação (ordem natural) π_0 . Em qualquer permutação π dos mesmos objectos, se dois destes se dispõem como em π_0, fazem uma permanência; se se dispõem em ordem contrária, fazem uma inversão. Contando as inversões que cada elemento de π faz com os elementos seguintes e somando os sucessivos resultados parciais, obtém-se o número de inversões da permutação. Por exemplo, $\pi = (3,4,2,5,1)$ tem, relativamente a $\pi_0 = (1,2,3,4,5)$, $2 + 2 + 1 + 1 = 6$ inversões e $2 + 1 + 1 + 0 = 4$ permanências.

Designando por P o número de permanências e por I o número de inversões, reconhece-se facilmente que $P + I = \frac{1}{2} n (n - 1) = \binom{n}{2}$.

Então, dadas duas ordenações de um conjunto de n objectos, disponham-se estes pela ordem natural relativamente a uma das ordenações. Sejam P e I, respectivamente, o número de permanências e inversões na outra ordenação. O τ de Kendall é definido por

$$\tau = 1 - \frac{2I}{\binom{n}{2}} = \frac{2P}{\binom{n}{2}} - 1 \qquad (2.11)$$

Pondo $S = P - I$ vem ainda $\tau = \frac{S}{\binom{n}{2}}$

Retomemos o exemplo anterior, pondo pela ordem natural, π_0 , a ordenação em trabalhos de laboratório

	I	D	B	G	J	H	E	A	C	F
Lab. (X)	1	2	3	4	5	6	7	8	9	10
Teor. (Y)	2	1	5	3	6	4	8	9	10	7

tem-se $I = 1 + 0 + 2 + 0 + 1 + 0 + 1 + 1 + 1 = 7$

$P = \binom{n}{2} - I = 45 - 7 = 38$; $S = P - I = 31$

e finalmente $\tau = \frac{31}{45} = 0,69$

Os coeficientes τ e r_s diferem em valor numérico, para um mes-

mo problema, sendo diferentes as escalas métricas subjacentes. Isto significa que τ e r_s não são directamente comparáveis.

Por outras palavras, se entre as variáveis X e Y conhecemos a medida de correlação r_s e entre X e Z a medida τ, não é possível decidir com qual das variáveis, Y ou Z, X está mais fortemente correlacionada.No entanto ambos os coeficientes utilizam a mesma quantidade de informação contida nos dados, concordando os testes de significância feitos sobre r_s e τ.

Para $n \geq 10$, τ tem uma distribuição aproximadamente normal de parâmetros

$$\mu_\tau = 0 \quad , \quad \sigma_\tau = \sqrt{\frac{2(2n+5)}{9n(n-1)}}$$

testando-se a significância de τ a partir da variável reduzida

$$\frac{\tau - \mu_\tau}{\sigma_\tau} = \frac{\tau}{\sigma_\tau}$$

Para valores $n \leq 10$ veja-se M.Kendall ob. cit. na bib. págs.52--53, e tab.1 do apêndice.

2.8. Correlação para variáveis discretas e nominais

A caracterização do grau de dependência entre duas variáveis discretas (nominais ou categoriais) pode ser feita por meio da contingência quadrada média (mean square contingency - K.Pearson).

Seja (X,Y) um par aleatório discreto e finito(*) com a matriz de probabilidade $\left[P_{ij}\right]$, i=1,...,n , j=1,...,m.

A contingência quadrada média é definida por

$$\varphi^2 = \sum_i \sum_j \frac{(P_{ij} - P_{i.} P_{.j})^2}{P_{i.} P_{.j}} = \sum_i \sum_j \frac{P_{ij}^2}{P_{i.} P_{.j}} - 1 \qquad (2.12)$$

com $P_{i.} = \sum_j P_{ij}$, i=1,...,n , a distribuição marginal de X e $P_{.j} = \sum_i P_{ij}$, j=1,...,m , a distribuição marginal de Y.

(*) X e Y podem representar características quantitativas ou qualitativas de uma dada população.

A partir da condição de independência —— $P_{ij} = P_{i.} P_{.j}$ para todo o par (i,j) —— reconhece-se imediatamente que $\varphi^2 = 0$ se e só se X e Y são independentes.

Como $P_{ij} \leq P_{i.}$ e $P_{ij} \leq P_{.j}$ segue-se que $\varphi^2 \leq q-1$ com $q = \min.(m,n)$.

$$\therefore \qquad 0 \leq \frac{\varphi^2}{q-1} \leq 1$$

$\varphi^2/q-1$ pode ser usado como medida, numa escala padronizada, do grau de dependência entre as variáveis.

No caso particular importante m = n = 2 vem

$$\varphi^2 = \frac{(P_{11}P_{22} - P_{12}P_{21})^2}{P_{1.}P_{2.}P_{.1}P_{.2}} \quad , \; 0 \leq \varphi^2 \leq 1 \qquad (2.13)$$

Na prática não se conhece, em geral, a matriz de probabilidade $\left[P_{ij}\right]$, sendo as probabilidades teóricas P_{ij} estimadas pelas frequências relativas $P^{*}_{ij} = f_{ij}/n$, calculadas sobre uma amostra de dimensão n.

Parte-se então de uma tabela de contingência (ou correlação) $k \times \ell$ com a matriz de frequências absolutas $\left[f_{ij}\right]$:

X \ Y	y_1	y_2	...	y_j	...	y_ℓ	
x_1	f_{11}	f_{12}	....	f_{1j}	....	$f_{1\ell}$	$f_{1.}$
x_2	f_{21}	f_{22}	....	f_{2j}	....	$f_{2\ell}$	$f_{2.}$
:	:	:	....	:	....	:	:
x_i	f_{i1}	f_{i2}	....	f_{ij}	....	$f_{i\ell}$	$f_{i.}$
:	:	:	....	:	....	:	:
x_k	f_{k1}	f_{k2}	....	f_{kj}	....	$f_{k\ell}$	$f_{k.}$
	$f_{.1}$	$f_{.2}$	....	$f_{.j}$	....	$f_{.\ell}$	n

$$f_{i.} = \sum_{j=1}^{\ell} f_{ij}$$

$$f_{.j} = \sum_{i=1}^{k} f_{ij}$$

$$n = \sum_{i=1}^{k} \sum_{j=1}^{\ell} f_{ij}$$

obtendo-se o valor experimental

$$\varphi^{*2} = \frac{\chi^{*2}}{n} = \sum_{i,j} \frac{\left(\frac{f_{ij}}{n} - \frac{f_{i.}}{n}\,\frac{f_{.j}}{n}\right)^2}{\frac{f_{i.}}{n}\,\frac{f_{.j}}{n}} = \sum_{i,j} \frac{f_{ij}^2}{f_{i.}\,f_{.j}} - 1$$

φ^{*2} é uma medida da discrepância entre a distribuição observada e a distribuição esperada na hipótese H_o de independência. A distribuição assintótica de $\chi^{*2} = n\,\varphi^{*2}$, na hipótese H_o, é a de um χ^2 com $(K-1)(\ell-1)$ graus de liberdade.

A φ^{*2} — característica empírica correspondente a φ^2 — dá-se o nome de contingência quadrada média empírica (ou experimental).

Tem-se $0 \leq \frac{\varphi^{*2}}{q-1} = \frac{\chi^{*2}}{n(q-1)} \leq 1$, $q=\min.(k,\ell)$

A quantidade $\gamma = \sqrt{\frac{\varphi^{*2}}{q-1}} = \sqrt{\frac{\chi^{*2}}{n\,(q-1)}}$ toma o nome de coeficiente de correlação entre atributos. Outra medida do grau de associação entre as variáveis X e Y é dada pelo coeficiente de contingência C definido por

$$C = \sqrt{\frac{\varphi^{*2}}{1+\varphi^{*2}}} = \sqrt{\frac{\chi^{*2}}{n+\chi^{*2}}}$$

Para tabelas de contingência Kx K é $\max.C = \sqrt{\frac{k-1}{k}} < 1$. Por exemplo, para k=3 é max.C=0,816. A comparação de C com max.C dá uma ideia do maior ou menor grau de associação.

A significância de γ e de C é testada a partir do teste χ^2 de independência.

Exemplo 5. Sujeitaram-se a provas de orientação 150 alunos de um liceu, procurando-se determinar o tipo de estudos superiores a prosseguir, obtendo-se a seguinte tabela de contingência

Sexo / Estudos	Rapazes	Raparigas	
Literários	40	50	90
Científicos	35	9	44
Técnicos	15	1	16
	90	60	150

Pretende-se saber se existe alguma relação entre o tipo de estudos e o sexo. Para isso, calcule-se o χ^{*2} da tabela:

$$\chi^{*2} = n\left(\sum_{i,j} \frac{f_{ij}^2}{f_{i.}\, f_{.j}} - 1\right) = 150\left(\frac{40^2}{90 \times 90} + \frac{50^2}{60 \times 90} + \frac{35^2}{90 \times 44} + \frac{9^2}{60 \times 44} + \right.$$

$$\left. + \frac{15^2}{90 \times 16} + \frac{1^2}{60 \times 16} - 1\right) = 23{,}64$$

χ^{*2} tem, (na hipótese H_o de independência), aproximadamente uma distribuição χ^2 com (3-1)(2-1) = 2 graus de liberdade.

Como $P(\chi^2 \geqslant 23{,}64) < 0{,}001$ rejeita-se H_o concluindo-se que os valores

$$\gamma = \sqrt{\frac{\chi^{*2}}{n(q-1)}} = \sqrt{\frac{23{,}64}{150(2-1)}} = 0{,}40$$

e

$$C = \sqrt{\frac{n+\chi^{*2}}{n+\chi^{*2}}} = \sqrt{\frac{23{,}64}{150+23{,}64}} = 0{,}37$$

são significativamente diferentes de zero.

Caso particular importante: $k = \ell = 2$

Usa-se para medida de correlação a raiz quadrada da contingência quadrada média empírica:

$$\varphi^* = \sqrt{\varphi^{*2}} = \frac{f_{11}f_{22} - f_{12}f_{21}}{\sqrt{f_{1.}f_{2.}f_{.1}f_{.2}}} \qquad (2.14)$$

que toma o nome de correlação φ^* ou correlação pontual 2x2.

Considerando as frequências relativas $p^*_{ij} = \frac{f_{ij}}{n}$, $p^*_{i.} = \frac{f_{i.}}{n}$, $p^*_{.j} = \frac{f_{.j}}{n}$ vem

$$\varphi^* = \frac{p^*_{11}\, p^*_{22} - p^*_{12}\, p^*_{21}}{\sqrt{p^*_{1.}\, p^*_{2.}\, p^*_{.1}\, p^*_{.2}}}$$

φ^* é, evidentemente, um estimador de $\varphi = \dfrac{p_{11}\, p_{22} - p_{12}p_{21}}{\sqrt{p_{1.}p_{2.}p_{.1}p_{.2}}}$

sendo $\left[p_{ij}\right]$, i,j = 1,2, a matriz de probabilidade do par dicotômico (X,Y).

Como $p_{22} = 1 - p_{11} - p_{12} - p_{21}$ o numerador de φ pode tomar o aspecto

$$p_{11}p_{22} - p_{12}\,p_{21} = p_{11} - p_{11}^2 - p_{11}p_{12} - p_{11}p_{21} - p_{12}p_{21} =$$

$$= p_{11} - p_{11}(p_{11} + p_{12}) - p_{21}(p_{11} + p_{12})$$

$$= p_{11} - (p_{11} + p_{12})(p_{11} + p_{21}) = p_{11} - p_{1\cdot}p_{\cdot 1}$$

e $\therefore$ $$\varphi = \frac{p_{11} - p_{1\cdot}p_{\cdot 1}}{\sqrt{p_{1\cdot}p_{2\cdot}p_{\cdot 1}p_{\cdot 2}}}$$

vindo expressão análoga para φ^*

Exemplo 6.

Sejam α e β dois items dicotómicos de um teste.

Sobre uma amostra de 180 indivíduos obtiveram-se os resultados.

α \ β	Suc.	Ins.	
Suc.	40	20	60
Ins.	30	90	120
	70	110	180

Cálculo da correlação $\varphi^*_{\alpha\beta}$ entre os items α e β:

- usando as notações simplificadas

$p_{\alpha\beta}$ - proporção de sucesso conjunto

p_α, p_β - proporções de sucessos em α e β

q_α, q_β - proporções de insucessos em α e β

$(p_\alpha + q_\alpha = p_\beta + q_\beta = 1)$

vem

$$\varphi^*_{\alpha\beta} = \frac{p_{\alpha\beta} - p_\alpha p_\beta}{\sqrt{p_\alpha p_\beta q_\alpha q_\beta}} = \frac{\frac{40}{180} - \frac{60}{180}\,\frac{70}{180}}{\sqrt{\frac{60}{180}\,\frac{70}{180}\,\frac{120}{180}\,\frac{110}{180}}} = 0{,}403$$

Exemplo 7.

Valores particulares de φ

α \ β	Suc.	Ins.	
Suc.	0,25	0,25	0,50
Ins.	0,25	0,25	0,50
	0,50	0,50	1

$$\varphi = \frac{p_{\alpha\beta} - p_\alpha\, p_\beta}{\sqrt{p_\alpha\, p_\beta\, q_\alpha\, q_\beta}} = \frac{0,25-0,5x0,5}{\sqrt{0,5x0,5x0,5x0,5}} = 0$$

α \ β	Suc.	Ins.	
Suc.	0,50	0	0,50
Ins.	0	0,50	0,50
	0,50	0,50	1

$$\varphi = \frac{0,50 - 0,50x0,50}{0,50x0,50} = 1$$

Em particular com $p_\alpha = q_\alpha = 0,5$

vem

$$\varphi = \frac{p_{\alpha\beta} - p_\alpha p_\beta}{p_\alpha \sqrt{p_\beta\, q_\beta}} = \frac{2\, p_{\alpha\beta} - p_\beta}{\sqrt{p_\beta\, q_\beta}}$$

com $p_\alpha = q_\alpha = p_\beta = q_\beta$ tem-se $\varphi = \frac{p_{\alpha\beta} - 0,25}{0,25} = 4\, p_{\alpha\beta} - 1$

[Nota: É frequente, na literatura, as tabelas 2x2 apresentarem-se com o aspecto (em frequências absolutas)

a	b	a+b
c	d	c+d
a+c	b+d	a+b+c+d=n

vindo

$$\varphi^* = \frac{ad - bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}$$

]

É de verificação simples que a correlação φ é um caso particular da correlação ρ_{xy} quando X e Y são variáveis aleatórias tomando apenas os valores 0 e 1. De facto, considere-se a tabela de correlação do par (X,Y)

x \ y	1	0	
1	p_{11}	p_{12}	$p_{1.}$
0	p_{21}	p_{22}	$p_{2.}$
	$p_{.1}$	$p_{.2}$	

Por definição $\rho = \dfrac{\text{cov}\ (X,Y)}{\sqrt{v(X)\ v(Y)}}$

com $E(X) = p_{1.}$, $E(Y) = p_{.1}$,

$v(X) = p_{1.}p_{2.}$, $v(Y) = p_{.1}p_{.2}$

$$\text{cov}\ (X,Y) = E(XY) - E(X)\ E(Y) = \sum_{i,j} x_i y_j p_{ij} - E(X)\ E(Y)$$

$$= p_{11} - p_{1.}p_{.1}$$

$$\therefore \quad \rho_{xy} = \frac{p_{11} - p_{1.}p_{.1}}{\sqrt{p_{1.}p_{2}p_{.1}p_{.2}}} = \varphi_{xy} \qquad \text{c.q.d.}$$

Analogamente, e sobre uma tabela experimental, tem-se $r = \varphi^*$

Os coeficientes φ^* e r_t

Em geral φ^* e r_t , calculados sobre uma mesma tabela 2x2, diferem em valor sendo $\varphi^* < r_t$, aumentando a diferença com o afastamento das dicotomias de 50% - 50% .

Usualmente, quando se fala de correlação entre pontuações dicotómicas (ou dicotomisadas) de items de testes, subentende-se o coeficiente φ^* ou o coeficiente tetracórico r_t se for admissível a hipótese de distribuições gaussianas subjacentes às dicotomias.

2.9. Análise lógica de atributos em tabelas 2x2

Sejam α e β dois atributos definidos num universo U , finito, de dimensão N, e A e B, respectivamente, os conjuntos definidos em U

por α e β (*). Às operações lógicas de conjunção, disjunção e negação sobre atributos correspondem (são isomorfas), respectivamente, as operações lógicas de intersecção, reunião e complementação sobre conjuntos. (estamos uma vez mais em face de diferenças relativas à linguagem, mas traduzindo pontos de vista equivalentes). Os atributos compostos $\alpha\beta$, $\bar{\alpha}\,\beta$, $\alpha\,\bar{\beta}$, $\bar{\alpha}\,\bar{\beta}$ induzem uma partição sobre U , definida pela classe de conjuntos $\{A \cap B\ ,\ \bar{A} \cap B,\ A \cap \bar{B},\ \bar{A} \cap \bar{B}\}$ (**).

Sejam $f(\alpha\beta), \ldots,\ f(\bar{\alpha}\,\bar{\beta})$ as frequências absolutas (***) dos atributos $\alpha\beta$, ..., $\bar{\alpha}\,\bar{\beta}$ e considere-se a tabela de contingência 2x2

	β	$\bar{\beta}$	
α	$f(\alpha\beta)$	$f(\alpha\bar{\beta})$	$f(\alpha)$
$\bar{\alpha}$	$f(\bar{\alpha}\beta)$	$f(\bar{\alpha}\,\bar{\beta})$	$f(\bar{\alpha})$
	$f(\beta)$	$f(\bar{\beta})$	N

$$N = f(\alpha) + f(\bar{\alpha}) = f(\beta) + f(\bar{\beta}) = f(\alpha\beta) + f(\alpha\bar{\beta}) + f(\bar{\alpha}\beta) + f(\bar{\alpha}\,\bar{\beta})$$

Designando por $fr(\alpha \mid \beta)$ a frequência relativa (****) de α se β, isto é,

$$fr(\alpha \mid \beta) = \frac{f(\alpha\beta)}{f(\beta)} \quad , \quad f(\beta) \neq 0 \qquad (2.15)$$

(*) Recorde-se que toda a propriedade ou atributo num dado universo U define um conjunto — o conjunto de todos os indivíduos de $\mathcal{U}$ que têm essa propriedade.

(**) Usa-se o símbolo ~ para a negação de atributos e para a complementação de conjuntos.

(***) Frequência absoluta dum atributo α numa população ou universo $\mathcal{U}$ é o número de indivíduos de $\mathcal{U}$ que possuem o rtributo α. Por outras palavras, $f(\alpha)$ = card. A é o número de elementos (cardinal) do conjunto A definido por α .

(****) Frequência relativa de α em $\mathcal{U}$ é o quociente $fr(\alpha) = \frac{f(\alpha)}{N}$, quociente entre os cardinais de A e $\mathcal{U}$.

A condição de independência dos atributos α e β em U, exprime-se por

$$\frac{f(\alpha\beta)}{f(\beta)} = \frac{f(\alpha)}{N} \qquad (2.16)$$

ou

$$f(\alpha\beta) = \frac{f(\alpha).f(\beta)}{N} \qquad (2.17)$$

(2.17) exprime a condição de independência em termos de frequências absolutas. Dividindo (2.17) por N obtém-se a condição de independência em termos de frequências relativas:

$$f_r(\alpha\beta) = f_r(\alpha)\ f_r(\beta) \qquad (2.18)$$

De (2.15) ainda se tira $f_r(\alpha\beta) = f_r(\beta)\ f_r(\alpha|\beta)$ (2.19)

isto é, a frequência relativa de $\alpha\ \beta$ é o produto da frequência relativa de β pela frequência relativa de α se β.

Note-se que (2.18) é um caso particular de (2.19).

Na prática, os resultados experimentais obtidos não satisfazem exactamente às relações indicadas, quer por a população U em causa ser um subconjunto (amostra) de uma população teórica de referência, quer por se trabalhar sobre uma amostra S de U.

Assim, mesmo no caso de independência, tem-se em geral

$$f_r(\alpha|\beta) \neq f_r(\alpha)$$

ou, equivalentemente

$$f(\alpha\beta) \neq \frac{f(\alpha).\ f(\beta)}{N}$$

resultando as desigualdades de factores fortuitos ou aleatórios.

Pondo $f_o(\alpha\beta) = \frac{f(\alpha).\ f(\beta)}{N}$ designe-se este valor por <u>valor de independência</u> ou <u>valor esperado</u> de $f(\alpha\beta)$ (na hipótese de independência). A diferença

$$D = f(\alpha\beta) - f_o(\alpha\beta) \qquad (2.20)$$

toma o nome de <u>desvio</u> ou <u>discrepância</u> entre o valor observado e o valor esperado.

Com D pequeno (em módulo) o seu valor pode ser atribuído ao acaso, dizendo-se que o valor $D \neq 0$ é não significativo (ou não difere de

zero significativamente). Com D grande o desvio é, em princípio, significativo, sugerindo a rejeição da hipótese de independência.

Com D significativo diz-se que os atributos estão associados na população U.

Definição 1. α diz-se completamente associado a β se β implica α e escreve-se $\beta \Rightarrow \alpha$. Isto significa que todo o indivíduo β é um indivíduo α.

A condição $\beta \Rightarrow \alpha$ é equivalente à condição $f(\alpha\beta) = f(\beta)$ e $\therefore f_r(\alpha \mid \beta) = 1$. Também se diz, neste caso, que conhecido β a predição de α é feita com probabilidade igual a 1 ou sem erro.(*)

Definição 2. α e β são equivalentes se α está completamente associado a β e β completamente associado a α e escreve-se $\alpha \Leftrightarrow \beta$ ou $\alpha \equiv \beta$.

As duas condições, $\alpha \Rightarrow \beta$ e $\beta \Rightarrow \alpha$, implicam que $f_r(\alpha \mid \beta) = 1 = f_r(\beta \mid \alpha)$.

Como $\beta \Rightarrow \alpha$ e $\alpha \Rightarrow \beta$ são, respectivamente, equivalentes a $\tilde{\alpha} \Rightarrow \tilde{\beta}$ e $\tilde{\beta} \Rightarrow \tilde{\alpha}$ ainda se tem $f_r(\tilde{\beta} \mid \tilde{\alpha}) = 1 = f_r(\tilde{\alpha} \mid \tilde{\beta})$.

Definição 3. α e β são incompatíveis se $f(\alpha\beta) = 0$.

O anulamento de $f(\alpha\beta)$ implica imediatamente que $f_r(\alpha \mid \beta) = 0$

Por outro lado, de $f(\alpha\beta) = 0$ conclui-se que

$$\frac{f(\alpha\tilde{\beta})}{f(\alpha)} = 1 = f_r(\tilde{\beta} \mid \alpha) \quad \text{isto é,} \quad \alpha \Rightarrow \tilde{\beta} \quad \text{ou,}$$

equivalentemente, $\beta \Rightarrow \tilde{\alpha}$

A condição $\alpha \Rightarrow \tilde{\beta}$ mostra que a incompatibilidade de α e β implica que $\tilde{\beta}$ está completamente associado a α ou $\tilde{\alpha}$ está completamente associado a β.

(*) Esta noção é generalizada no último §.

Exemplo 8.

a)

	β	$\tilde{\beta}$	
α	30	20	50
$\tilde{\alpha}$	0	50	50
	30	70	100

α completamente associado a β:

$\beta \Longrightarrow \alpha$ ou $f_r(\alpha|\beta) = 1$

b)

	β	$\tilde{\beta}$	
α	45	0	45
$\tilde{\alpha}$	20	35	55
	65	35	100

β completamente associado a α:

$\alpha \Longrightarrow \beta$ ou $f_r(\beta|\alpha) = 1$

c)

	β	$\tilde{\beta}$	
α	40	0	40
$\tilde{\alpha}$	0	60	60
	40	60	100

α e β equivalentes:

$f_r(\alpha|\beta) = 1 = f_r(\beta|\alpha)$

d)

	β	$\tilde{\beta}$	
α	18	42	60
$\tilde{\alpha}$	12	28	40
	30	70	100

α e β independentes:

$f_r(\alpha|\beta) = f_r(\alpha|\tilde{\beta}) = f_r(\alpha)$

e)

	β	$\tilde{\beta}$	
α	0	30	30
$\tilde{\alpha}$	20	50	70
	20	80	100

α e β incompatíveis: $f(\alpha\beta)=0$

f)

	β	$\tilde{\beta}$	
α	20	30	50
$\tilde{\alpha}$	10	40	50
	30	70	100

α e β correlacionados (associação imperfeita)

Coeficiente de associação de Yule

Retome-se o desvio ou discrepância

$$D = f(\alpha\beta) - \frac{f(\alpha)\, f(\beta)}{N}$$

entre o valor observado e o valor de independência.

A partir de D construa-se o desvio relativo

$$Q = \frac{ND}{f(\alpha\beta)\, f(\bar{\alpha}\,\bar{\beta}) + f(\alpha\bar{\beta})\, f(\bar{\alpha}\beta)} = \frac{N\, f(\alpha\beta) - f(\alpha)\, f(\beta)}{f(\alpha\beta)\, f(\bar{\alpha}\,\bar{\beta}) + f(\alpha\bar{\beta})\, f(\bar{\alpha}\beta)}$$

dando uma melhor ideia do grau de associação entre α e β.

Q toma o nome de coeficiente de associação de Yule[33], sendo sempre $-1 \leq Q \leq 1$. Atendendo a que

$$f(\alpha\beta) + f(\bar{\alpha}\beta) + f(\alpha\bar{\beta}) + f(\bar{\alpha}\,\bar{\beta}) = N,$$

$$f(\alpha) = f(\alpha\beta) + f(\alpha\bar{\beta}) \quad , \quad f(\beta) = f(\alpha\beta) + f(\bar{\alpha}\beta)$$

ainda se pode escrever

$$Q = \frac{f(\alpha\beta)\, f(\bar{\alpha}\,\bar{\beta}) - f(\alpha\bar{\beta})\, f(\bar{\alpha}\beta)}{f(\alpha\beta)\, f(\bar{\alpha}\,\bar{\beta}) + f(\alpha\bar{\beta})\, f(\bar{\alpha}\beta)} \qquad (2.21)$$

É um exercício simples verificar que:

— $Q = 0$ se e só se α e β são independentes;

— $Q = -1$ se e só se $f(\alpha\beta)=0$ ou $f(\bar{\alpha}\,\bar{\beta})=0$;

— $Q = 1$ se e só se $\alpha \Longrightarrow \beta$ ou $\beta \Longrightarrow \alpha$.

Registe-se finalmente que se (α,β) é um par de atributos independentes, também $(\bar{\alpha},\beta)$, $(\alpha,\bar{\beta})$, $(\bar{\alpha},\bar{\beta})$ são pares de atributos independentes.

Exemplo 9. Cálculo de Q para cada uma das alíneas do exemplo 8.

a) $Q = \frac{30\times50-0\times20}{30\times50+0\times20} = 1$; b) $Q = \frac{45\times35-0\times20}{45\times35+0\times20} = 1$

(33) Sobre outros coeficientes de associação, V. H.R. Alker, ob. cit. na bib., pags 62-68.

c) $Q = \frac{40x60-0x0}{40x60-0x0} = 1$; d) $Q = \frac{18x28-42x12}{18x28+42x12} = 0$

e) $Q = \frac{0x50-20x30}{0x50+20x30} = -1$; f) $Q = \frac{20x40-10x30}{20x40+10x30} = 0,45$

2.10 Índice de predição em tabelas de contingência rxs

N indivíduos de uma população U são classificados segundo duas variáveis X e Y (quantitativas ou qualitativas). Sejam $C_\alpha = \{\alpha_1, \ldots, \alpha_r\}$ e $C_\beta = \{\beta_1, \ldots, \beta_s\}$ as classe dos atributos associados às duas classificações.

Tome-se como valor predito de C_β, associado à escolha aleatória de um indivíduo $u \in U$, o atributo β_j tal que $P(\beta_j)$ é máximo. Analogamente o valor predito de C_α, associado a uma escolha ao acaso, é o valor α_i que maximiza a probabilidade $P(\alpha_i)$. Então, a probabilidade de predição correcta, relativamente à classe C_β, é justamente a probabilidade

$$\max_j P(\beta_j) = \max_j P(B_j) = \max_j p_{\cdot j}$$

sendo B_j o subconjunto de U definido por β_j.

A probabilidade de erro (predição incorrecta) é

$$P(\varepsilon) = 1 - \max_j P(\beta_j) = 1 - \max_j p_{\cdot j}$$

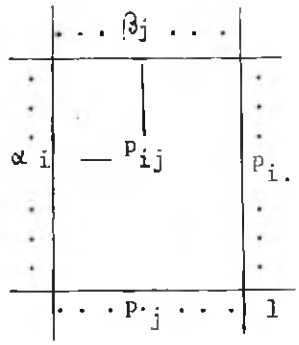


Tabela de contingência (em termos de probabilidades) de C_α e C_β.

Predição condicional de C_β dado C_α

Escolhido $u \in U$ sabe-se que $u \in A_i$, sendo $A_i \subset U$ o subconjunto de U determinado por α_i. Pretende-se predizer C_β, isto é, o subconjunto B_j tal que $u \in B_j$.

Definição. O valor predito de C_β conhecido A_i é o atributo β_j (ou subconjunto B_j) tal que

$$P(\beta_j | \alpha_i) = P(B_j | A_i) \text{ é máximo.}$$

$$\therefore$$

Probabilidade de predição correcta = $\max_j P(B_j|A_i)$

Probabilidade de erro conhecido $A_i = 1 - \max_j P(B_j|A_i) = P(\varepsilon|A_i)$

A probabilidade média de erro conhecida a classe C_α é

$$P(\varepsilon|C_\alpha) = \sum_i P(\varepsilon|A_i)P(A_i) = \sum_i \left[1-\max_j P(B_j|A_i)\right] P(A_i) =$$

$$= \sum_i P(A_i) - \sum_i \max_j P(B_j|A_i)\, P(A_i) = 1 - \sum_i \max_j P(A_i \cap B_j) =$$

$$= 1 - \sum_i \max_j p_{ij}$$

$P(\varepsilon) - P(\varepsilon | C_\alpha)$ é a redução média do erro de predição de C_β conhecido C_α

O índice $$\lambda_\beta = \frac{P(\varepsilon)-P(\varepsilon|C_\alpha)}{P(\varepsilon)} \qquad (2.22)$$

proposto por Goodman e Kruskal representa a redução média relativa (ou proporcional) do erro de predição de C_β conhecido C_α.

De acordo com as definições dadas tem-se

$$\lambda_\beta = \frac{1 - \max_j P(B_j) - 1 + \sum_i \max_j P(A_i B_j)}{1 - \max_j P(B_j)} = \frac{\sum_i \max_j P(A_i B_j) - \max_j P(B_j)}{1 - \max_j P(B_j)} =$$

$$= \frac{\sum_i \max_j p_{ij} - \max_j p_{.j}}{1 - \max_j p_{.j}} \qquad (2.23)$$

Analogamente se define $$\lambda_\alpha = \frac{\sum_j \max_i p_{ij} - \max_i p_{i.}}{1 - \max_i p_{i.}} \qquad (2.24)$$

Sobre uma tabela de contingência experimental, em função das frequências absolutas observadas, $f_{ij} = f(\alpha_i\ \beta_j)$, calculam-se as estimativas

$$\lambda^*_\beta = \frac{\sum_i \max_j f_{ij} - \max_j f_{.j}}{n - \max_j f_{.j}} \quad , \quad \lambda^*_\alpha = \frac{\sum_j \max_i f_{ij} - \max_i f_{i.}}{n - \max_i f_{i.}}$$

de λ_β e λ_α respectivamente.

Exemplo 10.

	β_1	β_2	β_3	
α_1	0	7	3	10
α_2	2	6	2	10
α_3	3	1	6	10
α_4	5	1	4	10
	10	15	15	40

$$\lambda^*_\beta = \frac{7+6+6+5-15}{40-15} = \frac{9}{25} = 0{,}36$$

significando este valor que na predição de C_β o conhecimento de C_α reduz, em média, a probabilidade de erro de 36%.

O cálculo de λ^*_α dá $\lambda^*_\alpha = \frac{5+7+6-10}{40-10} = \frac{8}{30} = 0{,}27$

Se a informação que se tem acerca de C_α não reduz a probabilidade de erro, isto é, $P(\mathcal{E}) = P(\mathcal{E} \mid C_\alpha)$, então $\lambda_\beta = 0$ e diz-se que não há associação preditiva entre C_α e C_β.
Por outros lado se $\lambda_\beta = 1$, isto é, $P(\mathcal{E} \mid C_\alpha) = 0$, diz-se que há associação preditiva completa de C_α para C_β ou que C_α implica C_β e escreve-se $C_\alpha \Longrightarrow C_\beta$

Exemplo 11.

	β_1	β_2	β_3	β_4	
α_1	0,2	0	0	0	0,2
α_2	0	0,3	0	0	0,3
α_3	0	0	0,4	0,1	0,5
	0,2	0,3	0,4	0,1	1

$$\lambda_\beta = \frac{0,2+0,3+0,4-0,4}{1-0,4} = \frac{0,5}{0,6} = 0,83$$

$$\lambda_\alpha = \frac{0,2+0,3+0,6+0,1-0,5}{1-0,5} = \frac{0,5}{0,5} = 1$$

mostrando estes valores que o conhecimento de C_β determina univocamente C_α, não sendo a inversa verdadeira. Por exemplo, conhecido α_3 há dois valores possíveis para β: β_3 e β_4.

Com $\lambda_\beta = \lambda_\alpha = 1$ tem-se $C_\alpha \Longrightarrow C_\beta$ e $C_\beta \Longrightarrow C_\alpha$, dizendo-se que C_α e C_β são equivalentes: $C_\alpha \Longleftrightarrow C_\beta$ ou $C_\alpha \equiv C_\beta$.

O conhecimento de uma das classes permite predizer, sem erro, a outra.

Em geral, os índices λ_α e λ_β têm valores diferentes, sendo portanto, índices não simétricos. Uma medida simétrica da capacidade preditiva de C_α para C_β e de C_β para C_α é dada por

$$\lambda_{\alpha\beta} = \frac{\sum_i \max_j p_{ij} + \sum_j \max_i p_{ij} - \max_j p_{.j} - \max_i p_{i.}}{2 - \max_j p_{.j} - \max_i p_{i.}}$$

O valor $\lambda_{\alpha\beta}$ está sempre compreendido entre λ_α e λ_β.

Do exemplo 10 vem

$$\lambda^*_{\alpha\beta} = \frac{7 + 6 + 6 + 5 + 5 + 7 + 6 - 15 - 10}{2 \times 40 - 15 - 10} = 0,31$$

valor compreendido entre λ^*_α e λ^*_β.

O valor calculado $\lambda^*_{\alpha\beta} = 0,31$ permite afirmar que o conhecimento de C_α ou C_β reduz, em média, a probabilidade de erro da predição de C_β ou C_α de 31%.

3. INTRODUÇÃO À ANÁLISE FACTORIAL

3.1. Introdução

A análise factorial é um método matemático que visa fundamentalmente substituir, sem perda de informação, n variáveis primitivas por q<n outras variáveis, não observáveis directamente, que tomam o nome de categorias ou factores, de tal modo que, a partir destes q factores, é possível avaliar aproximadamente os valores tomados pelas variáveis primitivas ou, mais simplesmente, reconstruir as suas intercorrelações. O método baseia-se no facto de que certas variáveis (testes) são a consequência ou o resultado de qualidades ou características não mensuráveis (variáveis latentes ou genotípicas).

As relações entre as variáveis observáveis e as variáveis latentes exprimem-se analiticamente por um sistema de equações.

A informação que estava repartida pelas variáveis primitivas é condensada nos factores, sendo estes, em geral, independentes na população teórica a que respeitam, ou de correlação muito fraca. Os factores representam, deste modo, classes ou grupos de variáveis.

A existência de um coeficiente de correlação não nulo entre dois testes sugere a existência de uma relação provável entre os processos psicológicos implicados nos dois testes.

O método factorial permite, a partir da análise de uma tabela (matriz) de correlações entre testes, substituir estes por um conjunto menos numeroso de variáveis-factores. Os factores assim definidos são de natureza formal, não tendo necessariamente existência "real". Constituem essencialmente um sistema de referência permitindo a representação simples de um grande número de factos experimentais, sendo neste sentido, conceitos operacionais.

No entanto, a tendência natural, é a procura de factores psicologicamente significativos.

Há muitos métodos de análise factorial, matematicamente equivalentes, isto é, igualmente correctos do ponto de vista lógico.

A escolha entre os diversos métodos baseia-se em critérios, dos quais o principal é o "princípio de economia" completado, em geral, por um "princípio de verosimilhança psicológica".

Procura-se assim descrever um conjunto de variáveis (testes) por meio de um número mínimo de factores comuns. Este mínimo é determinado por duas condições independentes e de níveis conceptuais diferentes. Uma condição dizendo respeito às interpretações, teóricas ou práticas, possíveis, dos factores comuns relativamente à situação em análise — domínio psicológico. A outra referindo-se às relações matemáticas a que os dados estão necessariamente submetidos — domínio lógico-matemático.

Os modelos factoriais são modelos probabilísticos. Quando se afirma que um teste pertence a um grupo (factor) toma-se uma decisão probabilística, sendo o grau de "pertença" do teste ao grupo expresso por um coeficiente dito de saturação.

O método mais antigo de análise factorial deve-se ao psicólogo inglês Charles Spearman. Na análise factorial clássica de Spearman os indivíduos são classificados ao longo de uma dimensão psicológica única, definida por um factor geral f=g. A "posição" de um indivíduo i é determinada pelo valor g_i que g toma para i — modelos unidimensionais. Mais precisamente, a pontuação x_{ji} de um indivíduo i numa prova j é uma função linear de g e de um factor único u_j ligado à prova j:

$$x_{ji} = a_j g_i + b_j u_{ji} \qquad (3.1)$$

A análise de estrutura latente de Lazarsfeld pode ser considerada como uma adaptação do modelo de Spearman para variáveis qualitativas, caso frequente em psicologia social e sociologia.

A análise multidimensional ou multifactorial de Thurstone resulta dos insucessos verificados na aplicação do método de Spearman, tendo-se constatado que certos resultados psicométricos não podiam ser explicados por um modelo da forma (3.1).

Nos modelos multidimensionais de Thurstone o número de factores (dimensões) não é limitado à priori, sugerindo (3.1) a seguinte generalização

$$x_j = c_{j1} f_1 + c_{j2} f_2 + \ldots + c_{jq} f_q = \sum_{l=1}^{q} c_{jl} f_l \qquad (3.2)$$

sendo para o indivíduo i

$$x_{ji} = \sum_{\ell=1}^{q} c_{j\ell} f_{\ell i} \qquad (3.3)$$

(3.3) exprime que o "êxito" ou rendimento (pontuação) do indivíduo i na prova j é uma função linear de q aptidões distintas (factores), $f_1,\ldots,f_q$, (com $f_{1i},\ldots f_{qi}$ as notas de i nos q factores e $c_{j1},\ldots,c_{jq}$ coeficientes traduzindo a importância da ligação entre o teste j e cada um dos factores). Thurstone aceita um número de factores indefinido, determinado à posteriori, pelos resultados experimentais.

Supondo os factores independentes, estes podem ser traduzidos por um sistema de q eixos ortogonais, sendo cada indivíduo representado por um ponto do espaço definido por estes eixos.

Seguindo Torrens-Ibern[34], no desenvolvimento das técnicas de análise factorial distinguem-se sucessivamente três direcções principais:

1. Reprodução das correlações. Partindo do estudo das intercorrelações de um conjunto de testes procura-se reproduzi-las, supondo-as resultantes de factores em número reduzido (Spearman, Holzinger, Burt, Thurstone, etc.).

2. Explicação da variabilidade. A análise factorial procura soluções conduzindo ao número mínimo de factores susceptíveis de explicar a variabilidade própria dos testes ou variáveis medidas. (Pearson, Hotelling, Kelley, etc.).

3. Estimação de parâmetros. Admitindo à priori a forma matemática das ligações entre as variáveis-testes e as variáveis-factores, estimam-se os parâmetros das ligações ou do modelo escolhido, por meio de critérios estatísticos da máxima verosimilhança, menores quadrados, etc. (Lawley, Rao, Harman, etc.)

3.2. A análise de grupos (Cluster analysis)

Um dos métodos de análise dum conjunto de variáveis (testes) é a procura de uma partição deste conjunto de tal modo que as variáveis de cada subconjunto da partição estejam relativamente "próximas" umas das ou

(34) Ob. cit. na bib., prefácio.

tras. Esta "proximidade" é definida a partir de uma estrutura particular de correlação que constitui o critério de existência dum tal agrupamento de testes em subconjuntos (clusters), cada subconjunto definindo um factor.

As correlações entre testes dum mesmo subconjunto serão, em média, mais elevadas que as correlações entre testes de subconjuntos diferentes. A análise de grupos — que se pode considerar como uma aproximação à análise factorial — tem como vantagem principal a comodidade do seu emprego e a relativa simplicidade das técnicas que utiliza. Em particular a "proximidade" entre duas variáveis pode ser definida à custa de modelos muito gerais.

Num grande número de situações emprega-se a análise de grupos pelo facto de que as medidas de que se dispõe não satisfazem aos critérios mais rigorosos da análise factorial. Por outro lado, o trabalho de cálculo, exigido pela análise factorial, para um número elevado de variáveis é considerável, sendo as operações ligadas à análise de grupos mais fáceis de realizar. No entanto os métodos da análise factorial são mais precisos que os da análise de grupos.

Considere-se,como exemplo, a seguinte matriz de correlações entre 6 testes, A, B, C, D, E, F.

Testes	A	B	C	D	E	F
A	.	0,01	0,64	0,06	0,72	-0,13
B	0,01	.	-0,07	0,42	0,08	0,56
C	0,64	-0,07	.	-0,02	0,55	-0,19
D	0,06	0,42	-0,02	.	0,11	0,47
E	0,72	0,08	0,55	0,11	.	-0,04
F	-0,13	0,56	-0,19	0,47	-0,04	.

Um grupo (cluster) é definido como um subconjunto de testes "altamente" correlacionados entre si e "fracamente" correlacionados com os testes que não pertencem ao grupo. Neste caso a constituição dos grupos poderá ser feita por simples inspecção da matriz de correlações. Uma maneira de iniciar a análise é considerar em primeiro lugar a maior correlação (em valor absoluto) da matriz.

No caso presente r(A,E) = 0,72. Os teste A e E podem então formar o núcleo do primeiro grupo. Qualquer outro teste fortemente correlacionado com A e E pode ser colocado no mesmo grupo. Como r(C,A) = 0,64 e r(C,E) = 0,55 junte-se C ao par (A,E). Nenhum dos outros três testes é fortemente correlacionado com A, C ou E, portanto nenhum deles pertence a este grupo.

Havendo correlações altas entre os testes não figurando no primeiro grupo, um ou mais grupos adicionais se podem formar.

Como r(B,D) = 0,42, r(B,F) = 0,56, r(D,F) = 0,47 segue-se que estes três testes constituem outro grupo.

No conjunto dos seis testes há portanto dois grupos, cada um definindo ou representando um factor. Reordenando as colunas (e as linhas) da matriz de correlação põem-se os grupos em evidência:

	A	E	C	B	F	D
A	.	0,72	0,64	0,01	-0,13	0,06
E	0,72	.	0,55	0,08	-0,04	0,11
C	0,64	0,55	.	-0,07	-0,19	-0,02
B	0,01	0,08	-0,07	.	0,56	0,42
F	-0,13	-0,04	-0,19	0,56	.	0,47
D	0,06	0,11	-0,02	0,42	0,47	.

3.3. O modelo teórico geral da análise factorial

Seja x_{ji} a medida do comportamento do indivíduo i na situação j, por exemplo, a pontuação obtida por i no teste j (neste sentido falaremos indistintamente nas variáveis x_j, x_k,... ou nos testes j, k,...)

A hipótese fundamental da análise factorial afirma que o "comportamento observado" x_{ji} (do indivíduo i na situação j) pode ser previsto conhecendo-se o valor de m características do indivíduo, não directamente observáveis, na situação j. É-se assim conduzido ao modelo factorial geral

$$x_{ji} = \varphi_j (f_{1i}, f_{2i}, \ldots, f_{mi}) \qquad (3.4)$$

sendo x_{ji} o valor observado de x_j para i e $f_{1i},\ldots,f_{mi}$ os valores (não directamente observáveis) das variáveis latentes ou genotípicas, $f_1,\ldots f_m$, para o indivíduo i. As variáveis latentes tomam o nome de variáveis-factores ou simplesmente factores.

A variável x_j é, portanto, uma função dos m factores considerados —— aplicação dum "espaço factorial" a m dimensões, F^m, em $\mathbb{R}$ (conjunto dos reais):

$$F^m \xrightarrow{\varphi_j} \mathbb{R} \qquad \left[x_j = \varphi_j\,(f_1, f_2, \ldots, f_m)\right]$$

A utilização prática do modelo (3.4) implica a adopção de hipóteses restritivas sobre a forma das funções φ_j e sobre as variáveis x_j e f_ℓ. O modelo clássico restrito satisfaz às hipóteses:

a) Hipótese de linearidade - fixação de uma subclasse da classe das aplicações φ_j, mais precisamente fixação da classe das aplicações lineares de F^m em $\mathbb{R}$:

$$x_j = \sum_{\ell=1}^{m} c_{j\ell}\, f_\ell$$

b) Hipótese de normalidade - as distribuições de x_j e f_ℓ são supostas normais reduzidas (note-se que basta considerar os factores f_ℓ gaussianos para que as variáveis x_j o sejam e vice-versa, facto que decorre das relações lineares que as ligam; por outro lado basta exigir a normalidade, efectuando-se em seguida a redução, a partir da mudança de variável $x'=(x-\mu)\sigma^{-1}$. Na realidade há que efectuar esta transformação apenas nas variáveis x_j, podendo os factores ser considerados, por definição, reduzidos).

c) Hipótese de independência - os factores f_ℓ são independentes (dados dois factores, f_k e f_ℓ por exemplo, a sua correlação é nula).

Estas hipóteses, em primeira análise arbitrárias e limitativas são, na realidade, pouco restritivas. De facto, qualquer função matemática pode ser, em primeira aproximação, expressa por uma função linear. Por outro lado, parece natural postular a normalidade das distribuições em jogo, na medida em que as distribuições normais se prestam de forma privilegiada ao tratamento matemático-estatístico (cf. "princípio de economia").

Acresce ainda que a métrica das variáveis psicológicas é, em

geral, arbitrária permitindo, por construção, escolher a forma mais conveniente para as respectivas distribuições (*).

Finalmente a hipótese de trabalho da independência dos factores apresenta-se como a mais discutível, justificando-se ainda pela simplicidade do tratamento matemático ulterior. Esta hipótese não é, no entanto, fundamental podendo usar-se factores não independentes, em correlação (embora representando capacidades ou aptidões separáveis), havendo então lugar à realização de uma análise factorial, de tipo 2, sobre os factores em correlação.

3.4. Os modelos de decomposição em factores comuns e factores únicos.

Nestes modelos postula-se a existência de factores comuns, em número q a determinar, intervindo em parte ou na totalidade das variáveis x_j observadas, e factores únicos ligados a uma única variável, em número n, havendo em jogo um conjunto ou bateria de n testes (variáveis).

A equação linear do modelo toma então a forma

$$x_j = \sum_{\ell=1}^{q} a_{j\ell} f_\ell + b_j u_j \qquad (3.5)$$

sendo $a_{j\ell}$ o coeficiente de saturação da variável x_j (ou do teste j) em factor f_ℓ ($a_{j\ell} = 0$ se f_ℓ não contribui para x_j) e b_j o coeficiente de saturação de x_j no factor u_j, factor que não intervém em nenhuma outra das variáveis observadas.

Atendendo às hipóteses do modelo linear restrito, procure-se a correlação ρ_{jk} entre os testes

$$x_j = \sum_{\ell=1}^{q} a_{j\ell} f_\ell + b_j u_j \quad \text{e} \quad x_k = \sum_{\ell=1}^{q} a_{k\ell} f_\ell + b_k u_k$$

(*) De facto, no caso da construção de testes, o psicólogo por escolha conveniente dos items e das notações pode obter histogramas com perfis aproximando distribuições teóricas fixadas previamente.

Por definição $\rho_{jk} = \rho(x_j, x_k) = \dfrac{cov(j,k)}{\sigma_j \sigma_k}$, mas

$cov(j,k) = E(jk) - E(j)E(k) = E(jk)$ e $\sigma_j = \sigma_k = 1$, pela hipótese de normalidade reduzida. Portanto,

$$\rho_{jk} = E(jk) = E(x_j x_k) = E\left[\left(\sum a_{j\ell} f_\ell + b_j u_j\right)\left(\sum a_{kt} f_t + b_k u_k\right)\right] =$$

$$= E\left(\sum_\ell \sum_t a_{j\ell} a_{kt} f_\ell f_t + \sum_\ell a_{j\ell} b_k f_\ell u_k + \sum_t a_{kt} b_j f_t u_j + b_j b_k u_j u_k\right) =$$

$$= \sum_\ell \sum_t a_{j\ell} a_{kt} E(f_\ell f_t) + \sum_\ell a_{j\ell} b_k E(f_\ell u_k) + \sum_t a_{kt} b_j E(f_t u_j) +$$

$$+ b_j b_k E(u_j u_k)$$

Atendendo a que os factores são independentes e com distribuições reduzidas tem-se

$$E(f_\ell f_t) = cov(f_\ell, f_t) = \begin{cases} 0 \text{ se } \ell \neq t \\ v(f_\ell) = 1 \text{ se } \ell = t \end{cases}$$

$E(f_\ell u_k) = cov(f_\ell, u_k) = 0$, $E(f_t u_j) = cov(f_f, u_j) = 0$

$E(u_j u_k) = cov(u_j, u_k) = 0$, vindo finalmente

$$\rho_{jk} = \rho(j,k) = \sum_{\ell=1}^{q} a_{j\ell} a_{k\ell} \qquad (3.6)$$

Tem-se, ainda, como facilmente se verifica, $a_{j\ell} = \rho(x_j, f_\ell)$, isto é, as saturações $a_{j\ell}$ são os coeficientes de correlação entre as variáveis x_j e os factores comuns f_ℓ. Analogamente, $b_j = \rho(x_j, u_j)$. (Note-se que afirmar a normalidade reduzida das pontuações x_j é postular a existência de uma população teórica (infinita) sobre a qual

$$E(x_j) = \int_{-\infty}^{\infty} x_j \Psi(x_j)\, dx_j = 0 \quad \text{e} \quad v(x_j) = \int_{-\infty}^{\infty} x_j^2 \Psi(x_j)\, dx_j = 1 ,$$

sendo $\Psi(x_j)$ a densidade de probabilidade normal reduzida).

Aos testes j e k associem-se os vectores das saturações em fac

tores comuns(*)

$$\vec{a}_j = (a_{j1},\ldots,a_{jq}) \quad , \quad \vec{a}_k = (a_{k1},\ldots,a_{kq})$$

Com esta notação é imediato que a correlação entre j e k é o produto interno dos vectores saturação, numa base ortonormada (**):

$$\rho_{jk} = \vec{a}_j.\vec{a}_k = \sum_{\ell=1}^{q} a_{j\ell}\, a_{k\ell} = \|\vec{a}_j\|.\|\vec{a}_k\| \cos(\vec{a}_j, \vec{a}_k) \qquad (3.7)$$

designando $\|\vec{a}_j\|$ a norma (comprimento ou módulo) do vector $\vec{a}_j$:

$$\|\vec{a}_j\|^2 = \vec{a}_j \,.\, \vec{a}_j = \sum_{\ell=1}^{q} a_{j\ell}^2 \qquad (3.8)$$

Retomemos as correlações $\rho_{jk} = \sum_{\ell} a_{j\ell}\, a_{k\ell}$, formalmente interpretadas como produtos escalares, num espaço factorial F^q a q dimensões , sendo $a_{j\ell}$, $a_{k\ell}$ as coordenadas dos vectores num referencial ortonormado.

A variável $x_j = \sum_{\ell} a_{j\ell} f_\ell + b_j u_j$ pode representar-se pelo vector $\vec{a}_j = (a_{j1},\ldots,a_{jq})$, havendo uma correspondência biunívoca entre as variáveis x_j e os vectores $\vec{a}_j$.

Neste modelo vectorial, duas variáveis independentes (não correlacionadas) x_j e x_k são traduzidas por vectores ortogonais. De facto,

$$\rho_{jk} = 0 = \vec{a}_j.\vec{a}_k \quad \text{implica} \quad \vec{a}_j \perp \vec{a}_k$$

Como por hipótese as variáveis-factores são independentes, serão traduzidas no modelo por vectores ortogonais. Note-se que as variáveis f_ℓ (tal como as variáveis u_j) estão incluídas nas variáveis x_j : por exemplo, fazendo em (3.5) $a_{j1}=1$, $a_{j\ell}= 0$, $\ell \neq 1$, $b_j = 0$ vem $x_j = f_1$.

Então f_1 é representado pelo vector $\vec{F}_1 = (1,0,\ldots,0)$, f_2 pelo vector $\vec{F}_2 = (0,1,0,\ldots,0),\ldots,f_q$ pelo vector $\vec{F}_q = (0,\ldots,0,1)$, constituindo estes vectores uma base ortonormada para o espaço F^q. Tem-se evidente-

---

(*) Analogamente se podem considerar os vectores $\vec{c}$ das saturações nos m = q+n factores (comuns e únicos): $\vec{c}_j = (c_{j1},\ldots,c_{jm})$. Considerando o espaço factorial completo F^m (espaço dos factores comuns e únicos) tem-se que F^q é um subespaço de F^m sendo $\vec{a}_j$ a projecção ortogonal de $\vec{c}_j$ no subespaço F^q.

(**) V. apêndice III.

mente, para todo o j

$$\vec{a}_j = a_{j1}\vec{F}_1 + a_{j2}\vec{F}_2 + \ldots + a_{jq}\vec{F}_q \qquad (3.9)$$

De uma maneira geral (do ponto de vista da simplicidade e economia) é desejável que o número de factores comuns seja tão pequeno quanto possível. Spearman tinha dado a esta condição a forma mais radical quando encarava a possibilidade de exprimir as correlações a partir de um único factor comum ao conjunto das variáveis.

Os factores comuns subdividem-se em factores gerais, intervindo em todas as variáveis e factores de grupo, intervindo em grupos de variáveis (factores comuns a determinados subconjuntos de variáveis). Por outro lado, os factores únicos decompõem-se ortogonalmente em factores específicos, cada um relativo a uma dada variável, e factores de erro, associados a cada observação particular da variável e variando de observação para observação (isto é, o factor de erro é característico da observação e não da variável).

Os factores com maior interesse experimental são os factores de grupo. Consideremos, por exemplo, o seguinte modelo factorial relativo a 6 testes

	f_1	f_2	f_3
x_1	*	*	
x_2	*	*	*
x_3	*	*	
x_4	*	*	*
x_5	*		*
x_6	*		*

sendo f_1 um factor comum geral e f_2, f_3 factores comuns parciais — — factores de grupo.

A expressão analítica do modelo é

$$x_1 = a_{11}\, f_1 + a_{12}\, f_2$$

$$x_2 = a_{21}\, f_1 + a_{22}\, f_2 + a_{23}\, f_3$$

$$x_3 = a_{31}\, f_1 + a_{32}\, f_2$$

$$x_4 = a_{41}\, f_1 + a_{42}\, f_2 + a_{43}\, f_3$$

$$x_5 = a_{51}\, f_1 + \quad + a_{53}\, f_3$$

$$x_6 = a_{61}\, f_1 + \quad + a_{63}\, f_3$$

sendo $\vec{a}_1 = (a_{11}, a_{12}, 0)$, $\vec{a}_2 = (a_{21}, a_{22}, a_{23})$, $\vec{a}_3 = (a_{31}, a_{32}, 0)$,

$\vec{a}_4 = (a_{41}, a_{42}, a_{43})$, $\vec{a}_5 = (a_{51}, 0, a_{53})$, $\vec{a}_6 = (a_{61}, 0, a_{63})$

os vectores saturação.

A representação algébrica das ligações entre as variáveis x_j e os factores comuns (gerais e de grupo) f_ℓ, constitui o modelo factorial, cuja complexidade é definida pelo número q de factores comuns necessários para se ter uma reconstituição suficientemente precisa das variáveis x_j, em particular das intercorrelações, a partir das equações de ligação do modelo.

Decomposição das variâncias

De acordo com as hipóteses a variância teórica de x_j é unitária.

Com $x_j = \sum_\ell a_{j\ell}\, f_\ell + b_j\, u_j$ vem

$$v(x_j) = \sum_{\ell=1}^{q} a_{j\ell}^2 + b_j^2 = h_j^2 + b_j^2 = 1 \qquad (3.10)$$

A $h_j^2 = \sum_\ell a_{j\ell}^2 = a_{j1}^2 + \ldots + a_{jq}^2$ (3.11)

—— soma dos quadrados das saturações em factores comuns —— dá-se o nome de comunalidade de x_j, tomando b_j o nome de unicidade.

Introduzindo os vectores saturação $\vec{a}_j = (a_{j1}, \ldots, a_{jq})$ é

$$h_j^2 = \sum_{\ell} a_{j\ell}^2 = \vec{a}_j . \vec{a}_j = \| a_j \|^2 \qquad (3.12)$$

isto é, a comunalidade é o quadrado da norma do vector $\vec{a}_j$.

De (3.7) tem-se ainda

$$\rho_{jk} = h_j h_k \cos(\vec{a}_j , \vec{a}_k) \qquad (3.13)$$

o que mostra que a correlação entre dois testes depende das normas dos vectores que os representam e do respectivo ângulo.

Decompondo os factores únicos em factores específicos(*) e factores de erro, obtém-se

$$x_j = \sum_{\ell=1}^{q} a_{j\ell} f_\ell + d_j v_j + e_j \varepsilon_j \quad , \quad \text{cov}(v_j , \varepsilon_j) = 0$$

com v_j, factor específico; d_j, saturação de x_j em v_j ; ε_j factor de erro; e_j, saturação de x_j em ε_j ; a que corresponde a decomposição da variância

$$v(x_j) = \sum_{\ell=1}^{q} a_{j\ell}^2 + d_j^2 + e_j^2 = 1 = h_j^2 + d_j^2 + e_j^2 \qquad (3.14)$$

Comparando (3.10) e (3.14) vem $b_j^2 = d_j^2 + e_j^2$, sendo d_j^2 a especificidade do teste e e_j^2 a variância de erro.

A relação (3.10) mostra imediatamente que a comunalidade h_j^2 , dum teste j representa a proporção da variância total do teste atribuível aos factores comuns - proporção da variância "explicada".

Analogamente o quadrado da saturação do teste j em factor f_ℓ , $a_{j\ell}^2$, representa a proporção da variância do teste "explicada" pelo fac-

---

(*) Os factores específicos, válidos unicamente para os testes a que dizem respeito, não influem nos outros testes; por definição, não estão em correlação com nenhum outro factor (quer comum, quer específico) e ∴ os seus eixos representativos são perpendiculares a todos os outros eixos factoriais, no espaço factorial completo $F^m = F^{q+n}$, a q+n dimensões. Interpretação equivalente consiste em tomar os factores j como factores específicos, admitindo que na unicidade b_j^2 há uma parte não atribuível aos factores, mas imputável aos erros de medida. Assim, a unicidade decompõe-se em duas partes: a especificidade d_j^2 (dizendo respeito à variabilidade própria do teste) e a variância de erro e_j^2 : — $b_j^2 = d_j^2 + e_j^2$

tor f_ℓ .

Seja r_{jj} o coeficiente de fiabilidade do teste j, isto é, a correlação entre duas séries de medições x'_j e x''_j de x_j sobre a população teórica, com

$$x'_j = \sum_{\ell=1}^{q} a_{j\ell} f_\ell + d_j v_j + e_j \varepsilon'_j$$

$$x''_j = \sum_{\ell=1}^{q} a_{j\ell} f_\ell + d_j v_j + e_j \varepsilon''_j$$

sendo ε'_j e ε''_j variáveis normais reduzidas e independentes.

Então,

$$r_{jj} = \rho(x'_j , x''_j) = \frac{\mathrm{cov}(x'_j , x''_j)}{\sigma_{x'_j} \sigma_{x''_j}} = E(x'_j x''_j) =$$

$$= E\left[\left(\sum_\ell a_{j\ell} f_\ell + d_j v_j + e_j \varepsilon'_j \right) \left(\sum_t a_{jt} f_t + d_j v_j + e_j \varepsilon''_j \right) \right] =$$

$$= \sum_{\ell=1}^{q} a_{j\ell}^2 + d_{j\ell}^2 = h_j^2 + d_j^2 < 1 \qquad (3.15)$$

(3.15) mostra que r_{jj} representa a proporção da variância calculada que corresponde à variância da característica mensurável x_j. A proporção complementar é atribuível aos erros de medição, cujo valor é e_j^2 . As relações entre as diferentes partes da variância podem resumir-se nas igualdades

variância = comunalidade + unicidade : $v(x_j) = 1 = h_j^2 + b_j^2$

unicidade = especificidade + variância de erro: $b_j^2 = d_j^2 + e_j^2$

fiabilidade = comunalidade + especificidade: $r_{jj} = h_j^2 + d_j^2 = 1 - e_j^2$

Note-se que, na prática, apenas se podem conhecer valores experimentais para r_{jj}, isto é, avaliações ou estimativas da fiabilidade, obtidas sobre certas amostras da população teórica e a partir de certas técnicas ou métodos (*)

(*) V. apendice II.

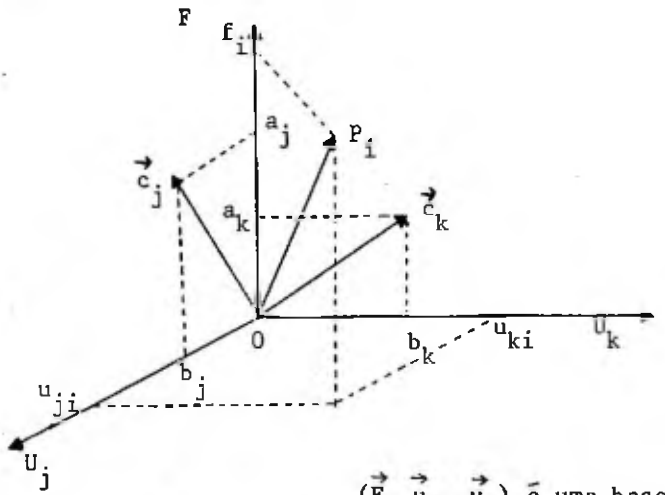
Exemplo 1.

Sejam j e k dois testes admitindo a seguinte decomposição:

$$x_j = a_j f + b_j u_j \quad ; \quad x_k = a_k f + b_k u_k$$

com f factor comum aos dois testes, u_j e u_k factores únicos.

Os factores f, u_j e u_k admitem-se independentes. No espaço tridimensional, F^3, definido pelo factor comum e pelos factores únicos tem-se



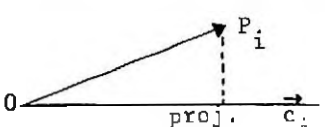
$$\vec{F} = (1,0,0)$$
$$\vec{u}_j = (0,1,0)$$
$$\vec{u}_k = (0,0,1)$$
$$\vec{c}_j = (a_j, b_j, 0)$$
$$\vec{c}_k = (a_k, 0, b_k)$$

$(\vec{F}, \vec{u}_j, \vec{u}_k)$ é uma base ortonormada para o espaço factorial completo F^3.

Seja $\mathcal{U}$ uma população teórica infinita sobre a qual x_j, x_k, f, u_j e u_k são normais reduzidas. No espaço dos factores os indivíduos de $\mathcal{U}$ serão representados por pontos cujas projecções ortogonais sobre os eixos factoriais são as pontuações em factores e cujas projecções ortogonais sobre os testes (vectores $\vec{c}$) dão as pontuações nos testes. De facto, seja $(P_i - 0)$ o vector de F^3 representando o indivíduo i: $P_i - 0 = (f_i, u_{ji}, u_{ki})$ e calcule-se, por exemplo, a projecção ortogonal de $(P_i - 0)$ sobre $\vec{c}_j$.

Como $\| \vec{c}_j \| = 1$ vem

$$\text{Proj.}(P_i - 0) = (P_i - 0).\vec{c}_j = (f_i, u_{ji}, u_{ki}).(a_j, b_j, 0) =$$



$$= a_j f_i + b_j u_{ji} = x_{ji} \qquad \text{c.q.d.}$$

Relativamente ao referencial $\Sigma (0, F, u_j, u_k)$ a nuvem de pontos

representando os indivíduos de $\mathcal{U}$ é uma esfera de centro O. (Porquê?)

Em geral, a representação gráfica dos factores únicos é desprovida de interesse, limitando-se a representação ao espaço dos factores comuns, considerando-se apenas a projecção dos vectores teste neste espaço, isto é, os vectores saturação em factores comuns.

Note-se que as correlações ρ_{jk} entre os testes apenas dependem dos vectores e dos respectivos ângulos; por outras palavras, as correlações são determinadas pela configuração e não pela estrutura (V.§3.5), podendo os eixos factoriais rodar em torno da origem sem alterar as correlações entre os testes.

Exemplo 2. Seja $\vec{a}_j$ o vector saturação em factores comuns do teste j. Se $\|\vec{a}_j\| \approx 1$ o teste j tem fraca unicidade e fiabilidade elevada. Se $\|\vec{a}_j\| \ll 1$ (norma fraca) a unicidade de j é importante e a sua fiabilidade ou é pequena ou j tem uma forte especificidade.

3.5. O ponto de vista matemático

Na prática, dado um conjunto de n variáveis (testes), uma análise factorial pretende "explicar" as intercorrelações (entre as n variáveis) a partir de um número mínimo de variáveis de decomposição ou factores comuns. A estes factores procurar-se-á atribuir, numa segunda fase, significado psicológico.

Considere-se a tabela (matriz) das correlações empíricas ou experimentais, obtidas sobre uma amostra de dimensão N

variáveis (testes)	x_1	x_2 ..... x_k ..... x_n
x_1	-	r_{12} .... r_{1k} .... r_{1n}
x_2	r_{21}	- .... r_{2k} .... r_{2n}
:	:	: : : : :
x_j	r_{j1}	r_{j2} .... r_{jk} .... r_{jn}
:	:	: : : : :
x_n	r_{n1}	r_{n2} .... r_{nk} .... -

Como $r_{jk} = r_{kj}$ para todo o j e k há apenas

$$\frac{n^2-n}{2} = \frac{n(n-1)}{2} = \binom{n}{2}$$

correlações distintas.

Seja q o número de factores comuns e $\vec{a}_1,\ldots,\vec{a}_n$ os vectores saturação. Pondo

$$\vec{a}_j,\ \vec{a}_k = r_{jk}\ ,\quad j<k\ ,\quad j,k = 1,2,\ldots,n \qquad (3.16)$$

obtem-se um sistema de $\binom{n}{2}$ equações em nq incógnitas, ligando as correlações empíricas conhecidas às saturações desconhecidas .

(3.16) escreve-se explicitamente na forma

$$\begin{aligned}
&\vec{a}_1.\vec{a}_2 = \sum_{\ell=1}^{q} a_{1\ell}a_{2\ell} = a_{11}\,a_{21} + a_{12}\,a_{22} + \ldots + a_{1q}\,a_{2q} = r_{12}\\
&\vec{a}_1.\vec{a}_3 = \sum_{\ell=1}^{q} a_{1\ell}a_{3\ell} = a_{11}\,a_{31} + a_{12}\,a_{32} + \ldots + a_{1q}\,a_{3q} = r_{13}\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \qquad (3.17)\\
&\vec{a}_j.\vec{a}_k = \sum_{\ell=1}^{q} a_{j\ell}a_{k\ell} = a_{j1}\,a_{k1} + a_{j2}\,a_{k2} + \ldots + a_{jq}\,a_{kq} = r_{jk}\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned}$$

Do ponto de vista matemático o método factorial consiste na resolução do sistema (3.17) de $\binom{n}{2}$ equações a nq incógnitas (*)

Discussão do sistema (3.17)

Em primeiro lugar vamos ver que se existe uma solução para(3.17) então há uma infinidade de soluções.

De facto, seja $\vec{y}_j = (y_{j1},\ldots,y_{jq})$, j=1,2....,n , uma solução de (3.17), isto é,

$$\vec{y}_j.\vec{y}_k = \sum_{\ell=1}^{q} y_{j\ell}y_{k\ell} \equiv r_{jk}\ ,\quad j<k\ ,\quad j,k = 1,\ 2,\ldots,\ n$$

(*) Na prática, o sistema factorial fundamental a resolver obtém-se juntando ao sistema (3.17) as n relações $\vec{a}_j.\vec{a}_j=h^2$ (V. §3.6 e §3.9) obtendo-se um novo sistema de $\binom{n}{2}+n = \frac{n(n+1)}{2}$ equações a nq+n=n(q+1) incógnitas, não se alterando a diferença entre o número de incógnitas e o número de equações.

Os valores $y_{j1},\ldots,y_{jq}$ são as projecções ortogonais do vector $\vec{y}_j$ sobre os eixos factoriais. Rodando os eixos obtém-se novas projecções $y'_{j1},\ldots, y'_{jq}$ sobre o novo sistema ortogonal de referência, mantendo-se invariantes as normas e os produtos internos, isto é,

$$y'^{2}_{j1} + \ldots + y'^{2}_{jq} = y^{2}_{j1} + \ldots + y^{2}_{jq} \quad (\text{ou } \sum_{\ell} y'^{2}_{j\ell} = \sum_{\ell} y^{2}_{j\ell}) \qquad (3.18)$$

$$\sum_{\ell=1}^{q} y'_{j\ell}\, y'_{k\ell} = \sum_{\ell=1}^{q} y_{j\ell}\, y_{k\ell} \equiv r_{jk} \qquad (3.19)$$

As relações (3.19) mostram que $(y'_{j1}, \ldots, y'_{jq})$ é uma nova solução de (3.17), exprimindo (3.18) a invariância das comunalidades.

Seguindo Thurstone designe-se por <u>configuração</u> o conjunto dos vectores $\vec{a}_1,\ldots,\vec{a}_n$, supostos aplicados em O — origem do referencial de eixos factoriais.

Uma configuração fica univocamente determinada por um conjunto de valores constantes para medidas das normas e ângulos dos vectores saturação. Uma configuração é, portanto, invariante para as transformações ortogonais de eixos. Por outras palavras, a configuração não muda se o conjunto de vectores rodar como um sólido invariável.

Então, obtida uma solução do sistema (3.17) a cada rotação dos eixos corresponde uma nova solução, todas correspondendo a uma mesma configuração.

Põe-se agora o problema de determinar o número de soluções não redutíveis umas às outras por rotação dos eixos, ou seja, o número de configurações diferentes satisfazendo ao sistema (3.17).

Este sistema, como vimos, é um sistema de $\binom{n}{2}$ equações a nq incógnitas. Este número de incógnitas pode ser reduzido por escolha conveniente do sistema de referência.

Fixado q considere-se uma configuração $(\vec{a}_1, \vec{a}_2,\ldots,\vec{a}_n)$, $n > q$.

Destes n vectores há q linearmente independentes e qualquer subconjunto de q+1 vectores é linearmente dependente. (Note-se que se apenas $p < q$ dos vectores $\vec{a}_j$, $j=1,2,\ldots,n$, fossem independentes, estes vectores podiam ser "mergulhados" num espaço factorial a $p<q$ dimensões,

contrariamente à hipótese feita).

Designe-se por $\vec{a}'_1, \ldots, \vec{a}'_q$ q vectores linearmente independentes da configuração $(\vec{a}_j,\ j=1,\ldots n)$, constituindo uma base(em geral não ortogonal) para o espaço F^q. Escolha-se agora uma solução particular desta configuração por fixação de uma posição particular do sistema ortogonal de referência, constituído pelos q eixos factoriais.

Assim, tome-se para F_1 a linha de acção de $\vec{a}'_1$,para F_2 um eixo perpendicular a $\vec{a}'_1$ e situado no plano $(\vec{a}'_1, \vec{a}'_2)$, (por exemplo, tome-se para F_2 a linha de acção do vector

$$\vec{v} = \vec{a}'_2 - \frac{\vec{a}'_2 \cdot \vec{a}'_1}{\|\vec{a}'_1\|^2}\ \vec{a}'_1$$

sendo de verificação simples que $\vec{v} \cdot \vec{a}'_1 = 0$)

Para F_3 tome-se um eixo do subespaço linear $(\vec{a}'_1, \vec{a}'_2, \vec{a}'_3)$ e prependicular ao plano $(\vec{a}'_1, \vec{a}'_2)$,..., finalmente tome-se para F_q um eixo do espaço F^q perpendicular ao hiperplano $(\vec{a}'_1, \ldots, \vec{a}'_{q-1})$.

Relativamente ao referencial assim definido, $\vec{a}'_1$ apenas tem componente não nula segundo F_1 , $\vec{a}'_2$ apenas tem componentes não nulas segundo F_1 e F_2 , $\vec{a}'_3$ apenas tem componentes segundo F_1 F_2 e F_3,..., e finalmente $\vec{a}'_{q-1}$ tem componentes não nulas segundo $F_1, \ldots, F_{q-1}$.

Deste modo anularam-se $(q-1) + (q-2) + (q-3) + \ldots + 1 = \frac{q(q-1)}{2}$ saturações . O número de incógnitas é então $nq - \frac{q(q-1)}{2}$

Sendo $\binom{n}{2} = \frac{n(n-1)}{2}$ o número de equações, se

$$nq - \frac{q(q-1)}{2} > \frac{n(n-1)}{2} \quad \text{ou} \quad q > \frac{1}{2}\,(1+2n - \sqrt{1+8n})$$

então o sistema é <u>indeterminado</u> admitindo uma infinidade de soluções,cada uma definindo uma configuração diferente(*).

(*) De um modo geral, a escolha do sistema de referência (fixação de um referencial em F^q) implica a imposição de $\binom{q}{2} = \frac{q(q-1)}{2}$ condições aos a_{jl} obtendo-se um sistema de $\binom{n}{2} + \binom{q}{2}$ equações a nq incógnitas, sendo o sistema indeterminado se

$$nq > \frac{n(n-1)}{2} + \frac{q\,(q-1)}{2} \quad \text{ou} \quad nq - \frac{q(q-1)}{2} > \frac{n(n-1)}{2}$$

relação obtida acima.

Para que haja solução única é necessário que

$$q \leq \frac{1}{2}\left(1 + 2n - \sqrt{1 + 8n}\right)$$

Com $q < \frac{1}{2}\left(1 + 2n - \sqrt{1 + 8n}\right)$ o sistema é, em geral, incompatível.

Por exemplo com n=4 e q=1 há $\binom{4}{2} = 6$ equações a 4 incógnitas (as 4 saturações no factor comum único) em geral incompatíveis, excepto se pelo menos duas das equações são consequências das restantes. Com q=2 há 6 equações a 8 incógnitas, sistema indeterminado, admitindo uma infinidade de configurações.

Na prática e em geral a estrutura das variáveis analisadas permite formular a hipótese da existência de $q \ll \frac{1}{2}\left(1+2n - \sqrt{1 + 8n}\right)$ para o qual o sistema (3.17) é possível e determinado, isto é, admitindo uma configuração única. Nestas condições um certo número de equações são consequências das restantes.

Note-se que o sistema (3.17) é um sistema experimental estando os valores conhecidos r_{jk} sujeitos a erros de amostragem.

Isto implica que só aproximadamente se pode verificar a dependência de algumas equações das restantes, estando esta dependência muitas vezes encoberta pelos citados erros de amostragem.

Este facto implica uma indeterminação quanto à escolha do número q — número mínimo de factores suficientes para "explicar" as intercorrelações.

Em geral q considera-se suficiente quando a junção de mais um factor dá uma contribuição desprezável para as intercorrelações.

Isto não significa que o número de factores independentes, subjacentes às variáveis em estudo, seja exactamente q, mas apenas que há q factores com importância na análise, sendo os restantes (que eventualmente existam) desprezáveis.

As saturações obtidas a partir do sistema experimental são estimativas das saturações do modelo teórico considerado.

Em resumo, na resolução de (3.17) começa-se por procurar uma solução particular da configuração única, definindo o espaço dos factores comuns, a partir de eixos factoriais arbitrários.

Em seguida, por rotação dos eixos procura-se uma nova solução particular a que correspondam eixos factoriais com significado psicológico.

O conjunto formado pelos vectores (configuração) e pelos eixos factoriais toma o nome de estrutura (Thurstone) .

(Nota: Nas considerações até aqui feitas apareceram dois conceitos de independência: a independência linear e a independência estocástica (ou probabilística).

No modelo vectorial definido, duas variáveis x_j e x_k são independentes em probabilidade, se os vectores saturação associados (vectores dum espaço a q dimensões) são ortogonais, isto é, $\vec{a}_j \cdot \vec{a}_k = 0$.

Se $\vec{a}_j \cdot \vec{a}_k \neq 0$ as variáveis são correlacionadas e

$$r_{jk} = h_j h_k \cos(\vec{a}_j, \vec{a}_k)$$

Por outro lado, $\vec{a}_j$ e $\vec{a}_k$ são linearmente independentes se uma relação do tipo

$$\lambda \vec{a}_j + \mu \vec{a}_k = \vec{0} \quad \text{implica} \quad \lambda = \mu = 0$$

É de fácil verificação que a independência estocástica implica a independência linear, mas a inversa não é verdadeira.

De facto, uma relação do tipo $\lambda \vec{a}_j + \mu \vec{a}_k = \vec{0}$, válida apenas para $\lambda = \mu = 0$ não implica que $\vec{a}_j \perp \vec{a}_k$.

Por exemplo, seja $\vec{a}_j = (2,0)$, $\vec{a}_k = (2,3)$. A relação vectorial $\lambda \vec{a}_j + \mu \vec{a}_k = 0$ implica que $(2\lambda, 0) + (2\mu, 3\mu) = (0,0)$

ou, $\begin{cases} 2\lambda + 2\mu = 0 \\ 3\mu = 0 \end{cases}$ sistema cuja solução é $\lambda = \mu = 0$

e, portanto, $\vec{a}_j$ e $\vec{a}_k$ são linearmente independentes.

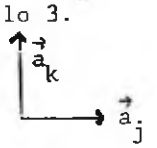
No entanto, $\vec{a}_j \cdot \vec{a}_k = 2 \times 2 + 0 \times 3 = 4 \neq 0$ e $\vec{a}_j$, $\vec{a}_k$ não são ortogonais e, equivalentemente, x_j e x_k são variáveis correlacionadas.

Por outro lado, se $\vec{a}_j \cdot \vec{a}_k = 0$ a combinação linear $\lambda \vec{a}_j + \mu \vec{a}_k$ só se anula com $\lambda = \mu = 0$. De facto, internando $\lambda \vec{a}_j + \mu \vec{a}_k = 0$ por $\vec{a}_j$ obtém-se $\lambda \|\vec{a}_j\|^2 = 0$

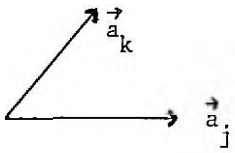
como por hipótese $\|\vec{a}_j\|^2 \neq 0$ vem $\lambda = 0$. Analogamente se conclui que $\mu = 0$.

Exemplo 3.

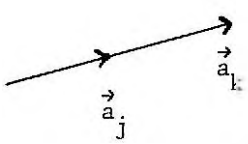
a)



x_j e x_k variáveis não correlacionadas, $\vec{a}_j$, $\vec{a}_k$ vectores linearmente independentes e ortogonais.

b) 

x_j e x_k variáveis correlacionadas, $\vec{a}_j$, $\vec{a}_k$ vectores linearmente independentes, não ortogonais.

c) 

x_j e x_k variáveis correlacionadas, $\vec{a}_j$, $\vec{a}_k$ vectores linearmente dependentes.)

3.6. A representação matricial

O ponto de partida de toda a análise factorial é a matriz das observações $[\, x_{ji} \,]$ a partir da qual se calcula a matriz das correlações experimentais $R^* = [\, r_{jk} \,]$. Substituindo em R^* a diagonal principal, por definição unida e unitária, pelas comunalidades desconhecidas, h^2_j, obtém-se uma nova matriz R^*_h — matriz transformada de $[\, r_{jk} \,]$(*) — que permitirá estimar, em primeiro lugar, as comunalidades e em seguida as saturações, ou seja, o modelo factorial completo. O problema assim posto admite duas vias distintas de abordagem:

a) Por considerações teóricas, fixa-se previamente o número q de factores comuns, ao qual se iguala a característica da matriz R^*_h. Nestas condições, todos os menores de ordem superior a q devem ser nulos, obtendo-se assim equações para a estimação das comunalidades. Com $q = 1$, temos a abordagem clássica de Spearman.

b) O número q de factores comuns não é limitado por considerações teóricas prévias, sendo fixado de modo a minimizar a unicidade e dar uma representação suficientemente precisa das intercorrelações. É o caso da análise multifactorial de Thurstone.

Retome-se a equação $x_{ji} = \sum_{\ell=1}^{m} c_{j\ell} f_{\ell i}$ (3.20)

(*) Designando por H a matriz diagonal das comunalidades e por I a matriz identidade, tem-se $R^*_h = R^* + (H-I)$.

dando a decomposição da pontuação do teste j para o indivíduo i, em m factores, não se explicitando aqui a distinção entre factores comuns e factores únicos (ou específicos).

O valor x_{ji} pode considerar-se o elemento genérico de uma matriz X em que a linha j é constituída pelos valores observados de uma variável (teste) j, sobre N indivíduos e a coluna i é formada pelas pontuações do indivíduo i em n testes.

Testes \ Indivíduos	1	2	...	N
1	x_{11}	x_{12}	...	x_{1N}
2	x_{21}	x_{22}	...	x_{2N}
:	:	:	:::	:
n	x_{n1}	x_{n2}	...	x_{nN}

$$X = \underset{(n \times N)}{\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nN} \end{bmatrix}}$$

Matriz das pontuações de N indivíduos em n testes.

Analogamente, considere-se o quadro das saturações de n testes em m factores

testes \ factores	1	2	...	m
1	c_{11}	c_{12}	...	c_{1m}
2	c_{21}	c_{22}	...	c_{2m}
:	:	:	:::	:
n	c_{n1}	c_{n2}	...	c_{nm}

$$F_o = \underset{(n \times m)}{\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ c_{n1} & c_{n2} & \cdots & c_{nm} \end{bmatrix}}$$

Matriz factorial - matriz dos coeficientes de saturação relativos a m factores e n testes.

A linha i da matriz F_o é o vector saturação da variável i no espaço factorial completo F^m.

Considere-se ainda o quadro e a correspondente matriz:

Indivíduos / Factores	1	2	...	N
1	f_{11}	f_{12}	...	f_{1N}
2	f_{21}	f_{22}	...	f_{2N}
⋮	⋮	⋮	⋮	⋮
m	f_{m1}	f_{m2}	...	f_{mN}

$$P_{(m \times N)} = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1N} \\ f_{21} & f_{22} & \cdots & f_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ f_{m1} & f_{m2} & \cdots & f_{mN} \end{bmatrix}$$

Matriz das pontuações de N indivíduos em m factores.

Atendendo a (3.20) e à definição das matrizes X, F_o e P é imediato que

$$X_{(nxN)} = F_{o\,(nxm)} \times P_{(mxN)}$$

(por exemplo, $x_{11} = c_{11} f_{11} + c_{12} f_{21} + c_{13} f_{31} + \ldots + c_{1m} f_{m1} = \sum_{\ell=1}^{m} c_{1\ell} f_{\ell 1}$)

Sendo $q < m$ o número de factores comuns e designando, como habitualmente, por $a_{j\ell}$ as saturações em factores comuns, tem-se

$$\rho_{jk} = \sum_{\ell=1}^{q} a_{j\ell} a_{k\ell}$$

ρ_{jk} — correlação teórica entre os testes j e k — pode considerar-se como o elemento genérico de uma matriz quadrada (nxn), simétrica, R, com elementos diagonais unitários.

Pondo $F_{(nxq)} = [a_{i\ell}]$, matriz teórica das saturações de n testes em q factores comuns e sendo $F^T = [a_{\ell j}]$ a matriz transposta de F, considere-se o produto (igualdade factorial fundamental)[35]

$$F F^T = R_h$$

Os elementos da diagonal principal de R_h são da forma $\sum_{\ell} a_{j\ell}^2 = h_j^2$, isto é, as comunalidades dos testes. Os restantes elementos de R_h coincidem com os da matriz R.

(35) L.L. Thurstone (1935), ob. cit. na bib., pag.70.

Como a matriz F tem característica q resulta que a característica de $R_h = F F^T$ ainda é q. É então válido o[36]

Teorema fundamental: Substituindo numa matriz de correlação a diagonal principal (diagonal unida de elementos unitários) pelas comunalidades obtém-se uma matriz R_h de característica igual ao número de factores comuns.

Em particular seja uma matriz respeitante a n testes com um único factor comum.

$$R_h = \begin{bmatrix} h_1^2 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & h_2^2 & \cdots & \rho_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{n1} & \rho_{n2} & \cdots & h_n^2 \end{bmatrix}$$

De acordo com o teorema fundamental, todos os menores de R_h de ordem 2 se anulam. Há a considerar três tipos de menores:

a) menores incluindo apenas correlações, por exemplo:

$$\begin{vmatrix} \rho_{31} & \rho_{32} \\ \rho_{n1} & \rho_{n2} \end{vmatrix} \quad ; \quad \begin{vmatrix} \rho_{1j} & \rho_{1n} \\ \rho_{kj} & \rho_{kn} \end{vmatrix}$$

b) menores incluindo uma comunalidade, por exemplo:

$$\begin{vmatrix} h_1^2 & \rho_{13} \\ \rho_{21} & \rho_{23} \end{vmatrix} \quad ; \quad \begin{vmatrix} \rho_{23} & \rho_{2j} \\ \rho_{j3} & h_j^2 \end{vmatrix}, \quad j \geq 3$$

c) menores principais, incluindo portanto duas comunalidades,:

$$\begin{vmatrix} h_1^2 & \rho_{13} \\ \rho_{31} & h_3^2 \end{vmatrix} \quad ; \quad \begin{vmatrix} h_j^2 & \rho_{jk} \\ \rho_{kj} & h_k^2 \end{vmatrix}, \quad k > j$$

Os determinantes do tipo a) dão origem às igualdades tetrádicas:

$$\rho_{31}\,\rho_{n2} - \rho_{32}\,\rho_{n1} = 0; \quad \rho_{1j}\,\rho_{kn} - \rho_{1n}\,\rho_{kj} = 0$$

As diferenças dos primeiros membros são designadas por tétrades, diferenças tetrádicas ou de 4.ª ordem.

---

(36) L.L. Thurstone (1938), ob. cit. na bib., pag.6.

Os determinantes do tipo b) permitem exprimir as comunalidades em função das correlações:

$$h_1^2 = \frac{\rho_{21}\,\rho_{13}}{\rho_{23}} \; ; \quad h_j^2 = \frac{\rho_{2j}\,\rho_{j3}}{\rho_{23}}$$

Os segundos membros destas igualdades tomam o nome de <u>tríades</u>.

No caso geral de q factores comuns a matriz R_h tem característica q e, por consequência, todos os menores de ordem q+1 se anulam.

O anulamento dos menores de ordem q+1, não incluindo comunalidades, permite estabelecer relações de compatibilidade entre as correlações — generalização das igualdades tetrádicas.

Os menores incluindo uma comunalidade permitem exprimir estas em função das correlações.

Na prática, parte-se de uma matriz R_h^* — matriz transformada de $R^* = [r_{jk}]$ — sendo r_{jk} uma estimativa da correlação teórica ρ_{jk}. As comunalidades serão estimadas a partir das correlações empíricas r_{jk}. Como é desejável trabalhar com um número de factores comuns tão pequeno quanto possível, procede-se a estas estimações de modo a minimizar a característica de R_h^*.

Se, por considerações teóricas, se admite a hipótese H de que o número de factores comuns é q, impor-se-á que a característica da matriz R_h^* seja q, ou, equivalentemente, que todos os menores de R_h^* de ordem superior a q se anulem.

Obtém-se deste modo um certo número de condições ligando as correlações empíricas, as saturações e as comunalidades. As equações resultantes da anulação dos menores de ordem q+1, incluindo uma comunalidade, permitem estimar estas. Menores distintos incluindo uma mesma comunalidade dão origem a diferentes estimativas.

Provando-se a equivalência estatística destas estimativas, aceita-se a hipótese H, continuando-se a análise. De contrário, dever-se-á reformular H, procurando novas soluções.

Na hipótese q=1 as tétrades empíricas não são nulas devido às flutuações aleatórias de amostragem, devendo, no entanto, os seus valores distribuir-se em torno de zero.

A partir das tríades experimentais estimam-se as comunalidades desconhecidas, havendo, como vimos, várias estimativas, teoricamente equivalentes, de cada comunalidade. (Por exemplo, $h_1'^2 = r_{21}\, r_{13}/r_{23}$ e $h_1''^2 =$

$= r_{41}\ r_{15}/r_{45}$ são duas estimativas de h_1^2). As diferenças entre as tríades relativas a uma mesma comunalidade devem ser não significativas do ponto de vista estatístico, para que a hipótese $q=1$ seja admissível.

Com $q>1$ os menores de ordem $q+1$ de R_h^* não devem ser significativamente diferentes de zero, estimando-se ainda as comunalidades a partir do anulamento dos menores de ordem $q+1$, incluindo uma comunalidade.

Inversamente, quando se verifica experimentalmente que todos os menores de ordem $q+1$ de R_h^* não são significativamente diferentes de zero, é admissível a hipótese de que os factores comuns são em número q.

Distinguindo na matriz F_o os factores comuns e os factores únicos, tem-se:

$$F_o = \begin{bmatrix} a_{11} & \cdots & a_{1q} & b_1 & \cdots & o \\ a_{21} & \cdots & a_{2q} & o\ b_2 & \cdots & o \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{nq} & o & \cdots & b_n \end{bmatrix}$$

Decompondo os factores únicos em factores específicos e factores de erro obtém-se a matriz F_1 de Thurstone:

$$F_1 = \begin{bmatrix} a_{11} & \cdots & a_{1q} & d_1 & \cdots & o & e_1 & \cdots & o \\ a_{21} & \cdots & a_{2q} & o\ d_2 & \cdots & o & o\ e_2 & \cdots & o \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{nq} & o & \cdots & d_n & o & \cdots & e_n \end{bmatrix}$$

$R_1 = F_1\ F_1^T$ é uma matriz cujos elementos não diagonais são as correlações ρ_{jk}. Os elementos diagonais são

$$\sum_{l=1}^{q} a_{jl}^2 + d_j^2 + e_j^2 = h_j^2 + d_j^2 + e_j^2 = v(x_j) = 1$$

isto é, $R_1 = R$. Assim, numa análise pretendendo atingir o estudo das próprias pontuações, usa-se a matriz empírica R^*, de diagonal unitária.

Considerando apenas os factores comuns e os factores específicos tem-se a matriz

$$F_2 = \begin{bmatrix} a_{11} & \cdots & a_{1q} & d_1 & \cdots & o \\ a_{21} & \cdots & a_{2q} & o\ d_2 & \cdots & o \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{nq} & o & \cdots & d_n \end{bmatrix}$$

Reconhecendo-se facilmente que no produto $F_2 F_2^T$

a) os elementos não diagonais coincidem com os da matriz FF^T;

b) o elemento genérico da diagonal principal é

$$r_{jj} = \sum_{\ell=1}^{q} a_{j\ell}^2 + d_j^2 = h_j^2 + d_j^2$$

Assim, quando numa análise se pretende atingir os factores comuns e os factores específicos, transforma-se a matriz R^* substituindo a diagonal principal pelas fiabilidades.

3.7. A hipótese de um único factor comum (Spearman)

O modelo factorial mais simples que se pode considerar admite que as pontuações obtidas numa bateria de n testes podem ser explicadas por um único factor comum e por n factores específicos (ou únicos) — — teoria dos dois factores.

O número total de factores em jogo é então n+1. As correlações entre os testes são explicadas pelo factor comum geral.

As equações fundamentais do modelo são:

$x_1 = a_1 f + b_1 u_1$

$x_2 = a_2 f + b_2 u_2$

.....

$x_j = a_j f + b_j u_j$

.....

$x_n = a_n f + b_n u_n$

f — factor comum (factor geral, g, de Spearman)

u_j — factor específico (único) relativo ao teste j.

Colocando-nos no espaço dos factores comuns — espaço factorial F_1 a uma dimensão — a matriz R_h (matriz R das correlações teóricas com as comunalidades na diagonal) tem característica unitária, anulando-se todos os menores de ordem igual ou superior a dois. Por outro lado, sendo

$$F = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

a matriz das saturações em factor comum, obtém-se a igualdade teórica factorial fundamental (Thurstone):

$$R_h = F\,F^T = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} = \begin{bmatrix} a^2_1 & a_1a_2 & a_1a_3 & \cdots & a_1a_n \\ a_2a_1 & a_2^2 & a_2a_3 & \cdots & a_2a_n \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_na_1 & a_na_2 & a_na_3 & \cdots & a_n^2 \end{bmatrix} =$$

$$= \begin{bmatrix} h_1^2 & \rho_{12} & \rho_{13} & \cdots & \rho_{1n} \\ \rho_{21} & h_2^2 & \rho_{23} & \cdots & \rho_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{n1} & \rho_{n2} & \rho_{n3} & \cdots & h_n^2 \end{bmatrix}$$

verificando-se imediatamente que quaisquer duas linhas (ou colunas) de R_h são proporcionais. Em particular, para as correlações das linhas j e k tem-se:

$$\frac{\rho_{j1}}{\rho_{k1}} = \frac{\rho_{j2}}{\rho_{k2}} = \cdots = \frac{\rho_{jn}}{\rho_{kn}} = \frac{a_j}{a_k}$$

Uma matriz nestas condições diz-se que apresenta a ordem hierárquica de Spearman. Neste sentido, uma matriz hierárquica é uma matriz de característica igual a um.

No espaço dos testes pode ser introduzida uma relação de ordem pondo $j > k$ se e só se $a_j > a_k$.

Analisemos mais em pormenor a relação $R_h = F\,F^T$. Em primeiro lugar, é imediato que as comunalidades são os quadrados das saturações no factor comum f.

A soma dos elementos da coluna j de R_h é :

$$\sum_{t=1}^{n} \rho_{tj} = \sum_{t=1}^{n} a_t a_j = a_j \sum_{t=1}^{n} a_t \qquad (3.21)$$

Somando (3.21) em j obtém-se a soma dos elementos de R_h :

$$\sum_{j=1}^{n} a_j \sum_{t=1}^{n} a_t = \left(\sum_{t=1}^{n} a_t \right)^2 \qquad (3.22)$$

De (3.21) e (3.22) deduz-se que

$$\ddot{a}_j = \frac{a_j \sum_t a_t}{\sqrt{\left(\sum_t a_t\right)^2}} = \frac{\text{total coluna j}}{\sqrt{\text{total geral}}} \qquad (3.23)$$

Na prática, uma matriz experimental R^* nunca apresenta exactamente a ordem hierárquica, pondo-se o problema de decidir se o desvio ou discrepância relativamente à hipótese de Spearman é não significativo.

Seja $R_h^* = \begin{bmatrix} h_1^2 & r_{12} & \dots & r_{1n} \\ r_{21} & h_2^2 & \dots & r_{2n} \\ \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & \dots & h_n^2 \end{bmatrix}$

e aceite-se a hipótese de um factor comum. Dever-se-á ter, por exemplo, as igualdades aproximadas entre as tríades.

$$\frac{r_{12}r_{13}}{r_{23}} \approx \frac{r_{12}r_{14}}{r_{24}} \approx \frac{r_{12}r_{15}}{r_{25}} \approx \frac{r_{13}r_{14}}{r_{34}} \approx \dots \qquad (3.24)$$

obtendo-se assim diversas estimativas de $a_1^2 = h_1^2$, cuja média se pode tomar para estimativa final de a_1^2:

$$\hat{a}_1^2 = \hat{h}_1^2 = \frac{1}{\binom{n-1}{2}} \sum_{\substack{j<k \\ j,k \neq 1}} \frac{r_{1j}\,r_{1k}}{r_{jk}} = \frac{2}{(n-1)(n-2)} \sum_{j<k} \frac{r_{1j}\,r_{1k}}{r_{jk}}$$

e, em geral, $\hat{a}_\ell^2 = \hat{h}_\ell^2 = \frac{2}{(n-1)(n-2)} \sum_{\substack{j<k \\ j,k \neq \ell}} \frac{r_{\ell j}\,r_{\ell k}}{r_{jk}}$ (3.25)

Uma estimativa equivalente, proposta por Spearman, obtém-se pelo quociente

$$\hat{h}_1^2 = \frac{r_{12}r_{13} + r_{12}r_{14} + r_{12}r_{15} + r_{13}r_{14} + \dots}{r_{23} + r_{24} + r_{25} + r_{34} + \dots} = \frac{\Sigma\, r_{1j}r_{1k}}{\Sigma\, r_{jk}}, \quad j \neq k \neq 1 \qquad (3.26)$$

Seja $T_1 = r_{21} + r_{31} + \dots + r_{n1} = \Sigma_j\, r_{j1}$ a soma das correlações da primeira coluna de R_h^*, $T'_1 = r^2_{21} + r^2_{31} + \dots + r^2_{n1} = \Sigma_j\, r^2_{j1}$ a soma dos quadrados das parcelas de T_1 e $T = \Sigma_i \Sigma_j\, r_{ij} = 2 \sum_{i<j} r_{ij}$ a soma das $n^2-n = n(n-1)$ correlações da matriz R_h^*.

Como $T_1^2 - T_1' = \left(\sum_j r_{j1}\right)^2 - \sum_j r_{j1}^2 = 2(r_{12}r_{13} + r_{12}r_{14} + r_{12}r_{15} + \ldots)$

$$T - 2T_1 = 2\sum_{i<j} r_{ij} - 2\sum_j r_{j1} = r_{23} + r_{24} + r_{25} + r_{34} + \ldots$$

(3.26) pode escrever-se na forma

$$\hat{h}_1^2 = \frac{\left(\sum_j r_{j1}\right)^2 - \sum_j r_{j1}^2}{\sum_i \sum_j r_{ij} - 2\sum_j r_{j1}} = \frac{T_1^2 - T_1'}{T - 2T_1}$$

Um cálculo análogo dá as estimativas $\hat{h}_\ell^2$, $\ell=2,\ldots,n$:

$$\hat{h}_\ell^2 = \frac{T_\ell^2 - T_\ell'}{T - 2T_\ell} \qquad (3.27)$$

Estimadas as comunalidades obtêm-se em seguida as saturações dividindo o total de cada coluna de R_h^* , depois de inseridos na diagonal os valores $\hat{h}_\ell^2$, pela raiz quadrada do total geral ou, mais simplesmente , extraindo a raiz a cada $\hat{h}_\ell^2$.

Exemplo 4.

Considere-se a tabela das intercorrelações de 5 testes

	1	2	3	4	5
1	-	0,74	0,67	0,61	0,55
2		-	0,61	0,56	0,50
3			-	0,50	0,45
4				-	0,41
5					-

cuja matriz R é

$$R = \begin{bmatrix} - & 0,74 & 0,67 & 0,61 & 0,55 \\ 0,74 & - & 0,61 & 0,56 & 0,50 \\ 0,67 & 0,61 & - & 0,50 & 0,45 \\ 0,61 & 0,56 & 0,50 & - & 0,41 \\ 0,55 & 0,50 & 0,45 & 0,41 & - \end{bmatrix}$$

Quaisquer duas linhas (ou colunas) de R são aproximadamente proporcionais, o que prova a existência de um factor comum aos 5 testes.

(Por exemplo, para as linhas 2 e 5 tem-se $\frac{0,74}{0,55} \simeq \frac{0,61}{0,45} \simeq \frac{0,56}{0,41} \simeq 1,36$)

Igualando a zero os menores do tipo b)

$$\begin{bmatrix} h_1^2 & 0,74 \\ 0,67 & 0,61 \end{bmatrix} \qquad \text{e} \qquad \begin{bmatrix} h_1^2 & 0,67 \\ 0,55 & 0,45 \end{bmatrix}$$

obtém-se duas estimativas de h_1^2 :

$$h_1'^2 = \frac{0,67 \times 0,74}{0,61} = 0,81 \text{ e } h_2''^2 = \frac{0,55 \times 0,67}{0,45} = 0,82$$

Uma estimativa final de h_1^2 obtém-se a partir de

$$\tilde{h}_1^2 = \frac{T_1^2 - T_1'}{T - 2T_1} = \frac{(0,74+0,67+0,61+0,55)^2-(0,74^2+0,67^2+0,61^2+0,55^2)}{2(0,61+0,56+0,50+0,50+0,45+0,41)} = 0,81$$

De modo análogo se obtinham estimativas de $h_2^2,\ldots,h_5^2$.

3.8 A hipótese de um factor geral e de factores de grupo (Holzinger)

A hipótese dum único factor comum é, em muitos casos, substituída pela necessidade de se admitir a existência de factores comuns a certos subconjuntos de testes, isto é, factores de grupo. Esta extensão natural da teoria dos dois factores de Spearman conduz ao chamado método bi-factorial.

Na realidade constata-se, frequentemente, que numa bateria de testes existem grupos de testes com afinidades muito fortes, isto é, altamente correlacionados. Admite-se então, em cada um destes grupos,um factor(nem geral, nem específico) tendo um papel preponderante. Introduzem-se assim os factores de grupo.

A hipótese agora em jogo pressupõe então que cada variável (teste)depende dum factor comum geral e de factores de grupo (além dos

factores únicos ou específicos), aceitando-se ainda a independência mútua de todos os factores.

Considere-se o seguinte modelo relativo a 5 testes com um factor comum geral, f_1, e dois factores de grupo f_2,f_3:

$$
\begin{aligned}
x_1 &= a_{11}f_1 + a_{12}f_2 + \qquad + b_1u_1 \\
x_2 &= a_{21}f_1 + a_{22}f_2 + \qquad + b_2u_2 \\
x_3 &= a_{31}f_1 + \qquad + a_{33}f_3 + b_3u_3 \\
x_4 &= a_{41}f_1 + \qquad + a_{43}f_3 + b_4u_4 \\
x_5 &= a_{51}f_1 + \qquad + a_{53}f_3 + b_5u_5
\end{aligned}
$$

a que corresponde a matriz F das saturações em factores comuns:

$$
F = \begin{bmatrix}
a_{11} & a_{12} & 0 \\
a_{21} & a_{22} & 0 \\
a_{31} & 0 & a_{33} \\
a_{41} & 0 & a_{43} \\
a_{51} & 0 & a_{53}
\end{bmatrix}
\begin{matrix}
1 \\ 2 \\ 3 \\ 4 \\ 5
\end{matrix}
\begin{matrix}
\} \text{ grupo 1} \\ \\ \} \text{ grupo 2}
\end{matrix}
$$

O cálculo de $R_h = F\,F^T$ dá

$$
\begin{bmatrix}
a_{11} & a_{12} & 0 \\
a_{21} & a_{22} & 0 \\
a_{31} & 0 & a_{33} \\
a_{41} & 0 & a_{43} \\
a_{51} & 0 & a_{53}
\end{bmatrix}
\times
\begin{bmatrix}
a_{11} & a_{21} & a_{31} & a_{41} & a_{51} \\
a_{12} & a_{22} & 0 & 0 & 0 \\
0 & 0 & a_{33} & a_{43} & a_{53}
\end{bmatrix}
=
\begin{bmatrix}
h_1^2 & \rho_{12} & \rho_{13} & \rho_{14} & \rho_{15} \\
\rho_{21} & h_2^2 & \rho_{23} & \rho_{24} & \rho_{25} \\
\rho_{31} & \rho_{32} & h_3^2 & \rho_{34} & \rho_{35} \\
\rho_{41} & \rho_{42} & \rho_{43} & h_4^2 & \rho_{45} \\
\rho_{51} & \rho_{52} & \rho_{53} & \rho_{54} & h_5^2
\end{bmatrix}
=
$$

$$
\left[\begin{array}{cc|c}
a_{11}^2 + a_{12}^2 & a_{11}\,a_{21} + a_{12}\,a_{22} & \begin{matrix} a_{11}\,a_{31} \\ a_{11}\,a_{41} \\ a_{11}\,a_{51} \end{matrix} \\
a_{11}\,a_{21} + a_{12}\,a_{22} & a_{21}^2 + a_{22}^2 & \begin{matrix} a_{21}\,a_{31} \\ a_{21}\,a_{41} \\ a_{21}\,a_{51} \end{matrix} \\
\hline
\begin{matrix} a_{11}\,a_{31} & a_{21}\,a_{31} \end{matrix} & & \\
\begin{matrix} a_{11}\,a_{41} & a_{21}\,a_{41} \end{matrix} & & \\
\begin{matrix} a_{11}\,a_{51} & a_{21}\,a_{51} \end{matrix} & &
\end{array}\right]
$$

$$
\left[\begin{array}{ccc}
\begin{matrix} a_{11}\,a_{31} \\ a_{21}\,a_{31} \end{matrix} & \begin{matrix} a_{11}\,a_{41} \\ a_{21}\,a_{41} \end{matrix} & \begin{matrix} a_{11}\,a_{51} \\ a_{21}\,a_{51} \end{matrix} \\
\hline
a_{31}^2 + a_{33}^2 & a_{31}\,a_{41} + a_{33}\,a_{43} & a_{31}\,a_{51} + a_{33}\,a_{53} \\
a_{31}\,a_{41} + a_{33}\,a_{43} & a_{41}^2 + a_{42}^2 & a_{41}\,a_{51} + a_{43}\,a_{53} \\
a_{31}\,a_{51} + a_{33}\,a_{53} & a_{41}\,a_{51} + a_{43}\,a_{53} & a_{51}^2 + a_{53}^2
\end{array}\right]
$$

A matriz R_h pode-se fraccionar em $2x2=4$ submatrizes, aparecendo as saturações em factores de grupo, apenas nas submatrizes diagonais.

Na prática, e admitida a hipótese da existência de factores comuns parciais, há que constituir os grupos de variáveis de tal modo que as correlações intra grupos sejam mais fortes que as correlações inter grupos. No interior dos grupos, os factores de grupo adicionam a sua contribuição à do factor geral (V. submatrizes diagonais de R_h). Feito o agrupamento dos testes, passa-se à estimação dos parâmetros do modelo.

Para o agrupamento das variáveis, Holzinger propõe o uso de um coeficiente de pertença, B, permitindo estimar em que medida um grupo de variáveis se distingue do conjunto das outras. B é definido pelo quociente, em percentagem, da média das intercorrelações das variáveis do grupo

pela média das correlações destas com as restantes variáveis.

Seja ν o número de variáveis do grupo e n o número total de variáveis. O número de intercorrelações do grupo é $\frac{\nu(\nu-1)}{2}$ e o número de correlações com as restantes variáveis é $\nu(n-\nu)$. B é, então, dado pela expressão

$$B\ (1,\ldots,\nu) = 100 \cdot \frac{2\,\frac{\sum_{j=1}^{\nu}\sum_{k>j}^{\nu} r_{jk}}{\nu(\nu-1)}}{\frac{\sum_{j=1}^{\nu}\sum_{\ell=\nu+1}^{n} r_{j\ell}}{\nu(n-\nu)}} = 200\,\frac{n-\nu}{\nu-1}\,\frac{\sum_{j=1}^{\nu}\sum_{k>j}^{\nu} r_{jk}}{\sum_{j=1}^{\nu}\sum_{\ell=\nu+1}^{n} r_{j\ell}} \qquad (3.28)$$

A repartição das variáveis em grupos inicia-se pelas duas variáveis de mais forte correlação, dando lugar a um coeficiente de pertença alto. Junta-se, em seguida, uma nova variável (teste) calculando-se o novo B. À medida que se vão juntando coeficientes de correlação ao numerador de (3.28), os valores obtidos oscilarão em torno do primeiro valor calculado até que a junção de uma variável provoque uma queda brusca no valor de B. Isto significa que a variável em questão não pertence ao grupo considerado. O processo continua, não se integrando no grupo nenhum teste provocando uma queda no valor de B. Recomeça-se a análise para um segundo grupo a partir do mais forte coeficiente de correlação restante.

Estimação das saturações

a) Em factor geral

Determinados os grupos, reordenem-se as variáveis de modo que na matriz empírica $R^{*}=\left[r_{jk}\right]$ (matriz das correlações experimentais) as matrizes fraccionadas diagonais sejam as submatrizes das correlações intra grupos. Quando os grupos são em número de três ou mais, estima-se facilmente a saturação dum teste em factor geral, a partir duma tríade de correlações não pertencentes às submatrizes da diagonal principal, isto é, correlações entre variáveis de grupos diferentes(*).

(*) A necessidade de utilizar variáveis pertencentes a três grupos distintos torna este método não aplicável aos modelos com dois factores de grupo. Pode-se usar neste caso um método derivado do método baricêntrico.

Da equação $R_h^* = FF^T$ tira-se, se j e k pertencem a grupos distintos,

$$r_{jk} = a_{j1}a_{k1}$$

multiplicando ambos os membros desta igualdade pelo quadrado da saturação de x_ℓ em f_1 vem

$$a^2_{\ell 1}\, r_{jk} = a_{\ell 1}\, a_{j1} \cdot a_{\ell 1} a_{k1} = r_{\ell j} \cdot r_{\ell k}$$

e ∴

$$a^2_{\ell 1} = \frac{r_{\ell j}\, r_{\ell k}}{r_{jk}} \qquad (3.29)$$

Para que este cálculo seja possível é necessário e suficiente que os três coeficientes de correlação digam respeito a variáveis não tendo um mesmo factor de grupo. Sobre dados empíricos, tríades diferentes dão estimativas diferentes da mesma saturação. Tal como no caso de Spearman, obtém-se uma estimação final sobre várias tríades, seja calculando a sua média aritmética, seja pelo quociente das somas relativas às expressões do numerador e do denominador de (3.29)

Assim, e em geral,

$$a^2_{\ell 1} = \frac{\sum_{j,k} r_{\ell j}\, r_{\ell k}}{\sum_{j,k} r_{jk}} \quad , \quad \ell \neq j \neq k \qquad (3.30)$$

pertencendo os testes ℓ , j e k a grupos diferentes.

Sendo ν o número de triades possíveis para a avaliação de $a^2_{\ell 1}$, tem-se a estimação pela média

$$a^2_{\ell 1} = \frac{1}{\nu} \sum_{j,k} \frac{r_{\ell j}\, r_{\ell k}}{r_{jk}} \quad , \quad \ell \neq j \neq k \qquad (3.31)$$

Holzinger e Harman (1941) e Thompson (1950) estabeleceram algoritmos para o cálculo das saturações em factor geral, com controle de cálculos[37].

(37) V. uma exposição simplificada em Reuchlin, 1964, ob. cit. na bib.pags. 204-208.

b) Em factores de grupo

Calculadas as saturações em factor geral, seja F_1 a matriz coluna formada por estes valores e calcule-se a matriz quadrada $F_1 F_1^T = R_1^*$, matriz cujos elementos representam a contribuição do factor geral para R_h^*.

A extracção do factor geral dá então lugar à matriz dos resíduos $R_h^* - R_1^*$, reconhecendo-se facilmente que, se o modelo factorial de que se partiu é compatível com as correlações observadas, os resíduos não devem ser significativamente diferentes de zero, nas submatrizes não diagonais (matrizes das correlações residuais entre variáveis de grupos diferentes para as quais apenas contribui o factor geral).

Pelo contrário, os elementos das submatrizes diagonais (para as quais contribuem os factores de grupo) devem ser significativamente diferentes de zero. Então, sendo o modelo factorial de que se partiu compatível com as observações, as submatrizes residuais diagonais devem ter característica um, caindo-se no caso de Spearman. Havendo várias tríades associadas a uma saturação $a_{\ell k}$ têm-se as estimações

$$\hat{a}^2_{\ell m} = \frac{\sum_{j,k} r_{\ell j}\, r_{\ell k}}{\sum_{j,k} r_{jk}} \qquad (3.32)$$

$$\hat{a}^2_{\ell m} = \frac{1}{\nu} \sum_{j,k} \frac{r_{\ell j}\, r_{\ell k}}{r_{jk}} \quad , \qquad m > 1 \ , \ \ell \neq j \neq k \qquad (3.33)$$

sendo ν o número de tríades possíveis para $a^2_{\ell m}$. Os testes ℓ, j, k pertencem agora a um mesmo grupo.

3.9. Uma solução factorial geral (Hotelling)

Retome-se o sistema factorial fundamental $FF^T = R_h^*$, isto é, o sistema (3.17) completado pelas n equações $\vec{a}_j . \vec{a}_j = h_j^2$, $j=1, 2, \ldots, n$.

Explicitamente:

$$\vec{a}_1 . \vec{a}_1 = \sum_{\ell=1}^{q} a_{1\ell}^2 = a_{11}^2 + a_{12}^2 + \ldots + a_{1q}^2 = h_1^2$$

$$\vec{a}_1 . \vec{a}_2 = \sum_{\ell=1}^{q} a_{1\ell} a_{2\ell} = a_{11} a_{21} + a_{12} a_{22} + \ldots + a_{1q}\ {}^{2}{}_{2q} = r_{12}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \qquad (3.34)$$

$$\vec{a}_j . \vec{a}_k = \sum_{\ell=1}^{q} a_{j\ell} a_{k\ell} = a_{j1}\, a_{k1} + a_{j2}\, a_{k2} + \ldots + a_{jq}\, a_{kq} = r_{jk}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

(3.34) é um sistema de $\frac{n(n-1)}{2} + n = \frac{n(n+1)}{2}$ equações a $nq + n = n(q+1)$ incógnitas ($n.q$ incógnitas $a_{j\ell}$ e n incógnitas h_j).

Sendo o sistema determinado é-o, como se viu, a menos de uma rotação no espaço F^q dos factores comuns. Imponham-se aos $a_{j\ell}$ as $\frac{q(q-1)}{2}$ condições

$$\sum_{t=1}^{n} a_{tj}\, a_{tk} = 0 \quad , \quad j < k \quad , \quad j, k = 1, 2, \ldots, q \qquad (3.35)$$

traduzindo a ortogonalidade das colunas da matriz F. Na forma matricial (3.35) é equivalente a

$$F^T F = \Lambda_q = \Lambda \qquad (3.36)$$

sendo $\Lambda_q = \Lambda = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & \lambda_q \end{bmatrix}$ uma matriz diagonal q x q

Multiplicando $F F^T = R_h^*$ à direita por F e atendendo a (3.36)

vem

$$F F^T . F = R_h^* F = F \Lambda \qquad (3.37)$$

Pondo $F = \left[a_{j\ell} \right] = \left[\vec{c}_1 \quad \vec{c}_2 \quad \ldots \quad \vec{c}_q \right]$

com $\vec{c}_\ell = \begin{bmatrix} a_{1\ell} \\ a_{2\ell} \\ \vdots \\ a_{n\ell} \end{bmatrix}$ (vector das saturações das n variáveis em factor f_ℓ ($\ell = 1, \ldots, q$))

(3.37) toma o aspecto

$$R_h^* \cdot \left[\vec{c}_1 \ldots \vec{c}_q\right] = \left[\vec{c}_1 \ldots \vec{c}_q\right] \cdot \Lambda$$

ou $$\left[R_h^* \vec{c}_1 \ldots R_h^* \vec{c}_q\right] = \left[\lambda_1 \vec{c}_1 \ldots \lambda_q \vec{c}_q\right]$$

desdobrando-se esta igualdade matricial nos q sistemas

$$R_h^* \vec{c}_\ell = \lambda_\ell \vec{c}_\ell \quad , \qquad \ell = 1, 2, \ldots, q \qquad (3.38)$$

Supostas as comunalidades conhecidas cada sistema (3.38) é um sistema de n equações a n incógnitas. Para $\ell = 1$ vem explicitamente

$$\begin{bmatrix} h_1^2 & r_{12} & r_{13} & \ldots & r_{1n} \\ r_{21} & h_2^2 & r_{23} & \ldots & r_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{n1} & r_{n2} & r_{n3} & \ldots & h_n^2 \end{bmatrix} \cdot \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix} = \lambda_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix}$$

ou

$$\begin{aligned} &(h_1^2 - \lambda_1)a_{11} + r_{12}\, a_{21} + r_{13}\, a_{31} + \ldots + r_{1n}\, a_{n1} = 0 \\ &r_{21}\, a_{11} + (h_2^2 - \lambda_1)\, a_{21} + r_{23}\, a_{31} + \ldots + r_{2n}\, a_{n1} = 0 \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &r_{n1}\, a_{11} + r_{n2}\, a_{21} + r_{n3}\, a_{31} + \ldots + (h_n^2 - \lambda_1)\, a_{n1} = 0 \end{aligned}$$

(3.38) mostra que $\lambda_1, \ldots, \lambda_q$ são q valores próprios da matriz R_h^* (e $\vec{c}_1, \ldots, \vec{c}_q$ os respectivos vectores próprios). Os λ_ℓ são portanto as raízes do polinómio característico $\left|R_h^* - \lambda I\right| = 0$.

Em geral, das n raízes desta equação escolhem-se as q maiores que se tomam para valores de $\lambda_1, \ldots \lambda_q$.

Como $F^T F = \Lambda$ tem-se $\sum_{t=1}^{n} a^2_{t\ell} = \lambda_\ell$, $\ell=1,\ldots, q$, o que mostra que esta solução factorial maximiza a soma dos quadrados das saturações relativas a cada factor ($a^2_{t\ell}$ — quadrado da saturação de x_t em f_ℓ).

Tendo as comunalidades sido correctamente fixadas há q raízes λ não negativas. Caso contrário, dever-se-á proceder a uma re-estimação das comunalidades, ou reduzir o número q de factores comuns.

3.10. Análise baricêntrica em factores comuns (Thurstone)

O método baricêntrico (ou centróide) de Thurstone e o método de Hotelling são técnicas especialmente construídas para extraír sucessivamente os factores, por ordem de importância, terminando a análise, quando apenas restarem factores negligenciáveis.

Qualquer dos métodos se propõe a efectivação da condensação estatística, sob a forma de factores comuns, duma tabela de correlações calculada.

No método de Hotelling, a importância de um factor é medida pela soma dos quadrados das saturações relativas a este factor.

Por outras palavras, um factor é tanto mais importante quanto maior fôr a proporção da variância por ele "explicada" (isto é, a proporção da variância imputável ao factor).

No método baricêntrico de Thurstone, a importância de um factor é medida pela soma das saturações relativas ao factor. Procura-se, então, maximizar, para cada factor, a soma dos coeficientes de saturação, obtendo-se, assim, de forma relativamente simples, uma solução factorial arbitrária, que se pode tornar psicologicamente significativa por rotação.

O método de Hotelling exige, em geral, o recurso a computador. Por este facto, e sendo método baricêntrico susceptível de tratamento manual, passaremos a expôr resumidamente este método.

Na prática, o ponto de partida é o conhecimento da matriz $R^* = [r_{jk}]$ das correlações experimentais, iniciando-se a análise por uma esti

mação das comunalidades, h^2_j, que se introduzem na diagonal principal de de R^*, obtendo-se a matriz experimental R^*_h. Uma estimação eficiente das comunalidades é dispensável, se o número de variáveis (testes) é grande (20 ou mais).

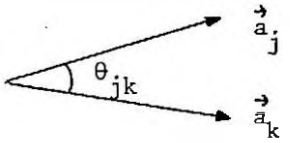
De facto, como veremos, a comunalidade de uma variável apenas intervém, em cálculos ulteriores, como uma parcela na soma das correlações desta variável com cada uma das outras.

Sendo as variáveis em número elevado, um erro sobre a comunalidade tem, pois, um pequeno efeito sobre a soma total.

Pelo contrário, sendo o número n de variáveis pequeno (principalmente quando $n < 10$), torna-se imperioso encontrar estimativas eficientes das comunalidades (isto é, estimativas centradas ou assintoticamente centradas e de variância mínima).

Em primeira aproximação, e como estimativa grosseira, tome-se para h_j o valor $\hat{h}_j = \max_k r_{jk}$

De facto, fixado j, para um teste k é



$$r_{jk} = h_j h_k \cos(\vec{a}_j, \vec{a}_k) = h_j h_k \cos \theta_{jk}$$

Fazendo a hipótese provisória da igualdade de todas as comunalidades, vem $r_{jk} = h^2_j \cos \theta_{jk}$

Seja k o teste mais fortemente correlacionado com j, ou seja, o teste para o qual $\theta_{jk} \simeq 0$ e $\therefore \cos \theta_{jk} \simeq 1$

Nestas condições, vem imediatamente $h^2_j \simeq r_{jk}$

c.q.d.

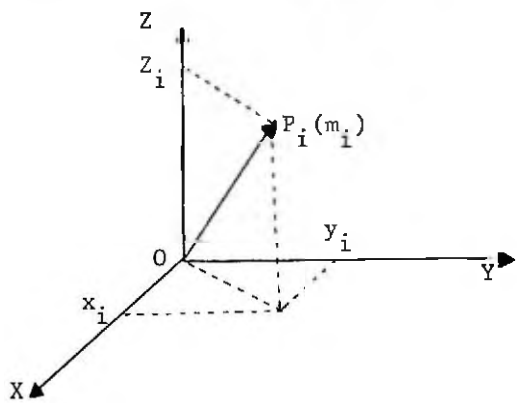
Este processo apenas serve, em geral, como ponto de partida para se obter iterativamente novas aproximações.

Antes de voltar ao problema da estimação das comunalidades, passemos à descrição do método baricêntrico.

É necessário para tal introduzir previamente algumas definições.

Baricentro ou centro de gravidade dum conjunto de pontos, num

dado referencial, é o ponto cujas coordenadas, relativas a cada um dos eixos do referencial, são as médias (simples ou ponderadas) das coordenadas correspondentes de cada ponto do conjunto. Por exemplo, em $\mathbb{R}^3$ tomemos k pontos $P_1, P_2, \ldots, P_k$, sendo o ponto genérico P_i caracterizado pelo "peso" m_i, i=1, ...,k.



Dado um referencial cartesiano Σ (O,X,Y,Z) cada ponto P_i é determinado em Σ pelo seu vector posição $(P_i - O)$ de componentes (x_i, y_i, z_i).

O baricentro do conjunto $\{P_1(m_1), \ldots, P_k(m_k)\}$ é o ponto G tal que

$$G - O = \frac{\sum_i m_i (P_i - O)}{\sum_i m_i} \quad , \quad \frac{1}{m} \sum_i m_i \, (P_i - O)$$

com $m = \sum_i m_i$. As coordenadas de G são, obviamente,

$$x_G = \frac{\sum_i m_i \, x_i}{\sum_i m_i} \quad , \qquad y_G = \frac{\sum_i m_i \, y_i}{\sum_i m_i} \quad , \qquad z_G = \frac{\sum_i m_i \, z_i}{\sum_i m_i}$$

Em particular, com $m_1 = \ldots = m_k = p$ vem $\sum_i m_i = kp$ e

$$x_G = \frac{\sum x_i}{k} \quad , \qquad y_G = \frac{\sum y_i}{k} \quad , \qquad z_G = \frac{\sum z_i}{k}$$

(médias aritméticas simples)

Como facilmente se reconhece

$$k \left\| G - O \right\| = \left[(\sum x_i)^2 + (\sum y_i)^2 + (\sum z_i)^2 \right]^{1/2}$$

Rodando o referencial Σ em torno de O de modo a fazer coincidir o eixo OX com a linha de acção do vector (G-O) obtém-se um novo referencial $\Sigma'(O, X', Y', Z')$ no qual o ponto G tem por coordenadas $(x'_G, 0, 0)$ e, portanto,

$$\sum x'_i = k \, x'_G = k \left\| G - O \right\| , \qquad \sum y'_i = \sum z'_i = 0$$

sendo (x_i', y_i', z_i') as coordenadas de P_i em Σ'

Então, em Σ' , $\Sigma x_i'$ é máximo anulando-se $\Sigma y_i'$ e $\Sigma z_i'$

Retome-se a matriz factorial F (matriz das saturações em factores comuns)

$$F = \begin{bmatrix} \vec{a}_1 \\ \vec{a}_2 \\ \vdots \\ \vec{a}_n \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1q} \\ a_{21} & a_{22} & \cdots & a_{2q} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nq} \end{bmatrix}$$

cujas linhas são os vectores saturação (em factores comuns).

O nosso fim é a resolução do sistema factorial fundamental

$$F\,F^T = R_h^*$$

sendo $F\,F^T$ a matriz das incógnitas e R_h^* a matriz das correlações empíricas com as comunalidades na diagonal principal.

Procure-se uma solução particular para a qual

$S_1 = a_{11} + a_{21} + \ldots + a_{n1}$ seja máximo

e

$S_2 = \ldots = S_q = 0$

(S_ℓ = soma dos elementos da coluna ℓ da matriz F)

Seja $(0, F_1, \ldots, F_q)$ um referencial cartesiano (ortogonal) e $P_1, P_2, \ldots P_n$ os pontos cujos vectores posição são

$$(P_1 - 0) = \vec{a}_1, \ldots, (P_n - 0) = \vec{a}_n$$

A solução particular procurada que maximize S_1 e anule $S_2, \ldots, S_q$ corresponde a rodar o referencial $(0, F_1, \ldots F_q)$ em torno de 0 de modo a que o eixo OF_1 coincida com a linha de acção do vector (G-0), sendo G o baricentro dos pontos $P_1, \ldots P_n$. Obtém-se assim um novo referencial $(0, F_1', \ldots, F_q')$ que toma o nome de referencial baricêntrico, do qual se fixou a posição do primeiro eixo factorial OF_1.

Seja então F' a matriz factorial referida ao referencial baricêntrico e calcule-se $F'\,F'^{T} = R_h^*$. (Note-se desde já que a matriz R_h^* é independente da escolha do referencial ao qual se refere a matriz factorial; por outras palavras, R_h^* é invariante para as transformações ortogonais de eixos. De facto, seja B uma transformação ortogonal e FB a transformada de F, tem-se

$$(FB)(FB)^T = (FB)\,B^T F^T = F(BB^T)\,F^T = FF^T = R_h^* \quad \text{c.q.d.)}$$

$$F'F'^T = \begin{bmatrix} \vec{a}_1 \\ \vdots \\ \vec{a}_n \end{bmatrix} \begin{bmatrix} \vec{a}_1^T \cdots \vec{a}_n^T \end{bmatrix} = \begin{bmatrix} \vec{a}_1.\vec{a}_1 & \vec{a}_1.\vec{a}_2 \cdots & \vec{a}_1.\vec{a}_n \\ \vec{a}_2.\vec{a}_1 & \vec{a}_2.\vec{a}_2 \cdots & \vec{a}_2.\vec{a}_n \\ \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \vec{a}_n.\vec{a}_1 & \vec{a}_n.\vec{a}_2 \cdots & \vec{a}_n.\vec{a}_n \end{bmatrix} = \begin{bmatrix} h_1^2 & r_{12} \cdots & r_{1n} \\ r_{21} & h_2^2 \cdots & r_{2n} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots \\ r_{n1} & r_{n2} \cdots & h_n^2 \end{bmatrix}$$

O cálculo da soma dos elementos da primeira coluna de $F'\,F'^T$ dá

$$\vec{a}_1.\vec{a}_1+\vec{a}_2.\vec{a}_1+\ldots+\vec{a}_n.\vec{a}_1=(\vec{a}_1+\ldots+\vec{a}_n).\vec{a}_1=(S'_1,S'_2,\ldots,S'_q).(a'_{11},a'_{12},\ldots,a'_{1q}) =$$

$$= (S'_1,0,\ldots,0).(a'_{11},\ldots,a'_{1q}) = a_{11}S'_1$$

Para a segunda coluna tem-se

$$(\vec{a}_1 + \ldots + \vec{a}_n).\vec{a}_2 = (S'_1,0,\ldots,0).(a'_{21},a'_{22},\ldots,a'_{2q}) = a'_{21}S'_1$$

e, de um modo geral, para a coluna genérica número j é

$$(\vec{a}_1 + \ldots + \vec{a}_n).\vec{a}_j = (S'_1,0,\ldots,0).(a'_{j1},\ldots,a'_{jq}) = a'_{j1}\,S'_1$$

A soma de todos os elementos da matriz vale então

$$a'_{11}S'_1 + a'_{21}\,S'_1 + \ldots + a'_{n1}S'_1 = (a'_{11}+a'_{21}+\ldots+a'_{n1})\,S'_1 = S'^2_1$$

Como

$$\frac{a'_{j1}S'_1}{\sqrt{S'^2_1}} = a'_{j1}$$ conclui-se que

— as saturação das variáveis no primeiro factor baricêntrico

obtêm-se dividindo o total de cada coluna de R^*_h pela raiz quadrada do total geral de R^*_h.

Determinam-se assim os valores $a'_{11}, a'_{21}, \ldots, a'_{n1}$, isto é, a primeira componente de cada um dos vectores $\vec{a}_1, \ldots, \vec{a}_n$. Por outras palavras, é conhecida neste momento a primeira coluna da matriz factorial baricêntrica F':

$$F'_1 = \begin{bmatrix} a'_{11} \\ a'_{21} \\ \vdots \\ a'_{n1} \end{bmatrix}$$

O produto $R^*_{(1)} = F'_1 F'^{T}_1$ é a matriz das contribuições do primeiro factor baricêntrico para a matriz R^*_h. De facto, da correlação

$$r_{jk} = \vec{a}_j \cdot \vec{a}_k = \sum_{\ell=1}^{q} a_{j\ell}\, a_{k\ell} = a_{j1}\, a_{k1} + \sum_{\ell=2}^{q} a_{j\ell}\, a_{k\ell}$$

é conhecida a parcela $a_{j1}\, a_{k1} = a'_{j1}\, a'_{k1}$

A diferença $R^*_h - R^*_{(1)}$ é a matriz das correlações residuais(*), atribuíveis aos factores $f_2, \ldots f_q$ e toma o nome de <u>primeira matriz residual</u>.

Repare-se que $R^*_h - R^*_{(1)}$ é a matriz das intercorrelações das primeiras variáveis residuais $y_j = x_j - a'_{j1} f_1$, $j = 1, 2, \ldots, n$.

As colunas da matriz $R^*_h - R^*_{(1)}$ têm soma nula. Considere-se a coluna genérica número j (j=1,...,n)

$$\begin{bmatrix} \vec{a}_1 \cdot \vec{a}_j \\ \vec{a}_2 \cdot \vec{a}_j \\ \vdots \\ \vec{a}_n \cdot \vec{a}_j \end{bmatrix} - \begin{bmatrix} a'_{11}\, a'_{j1} \\ a'_{21}\, a'_{j1} \\ \vdots \\ a'_{n1}\, a'_{j1} \end{bmatrix} = \begin{bmatrix} \vec{a}_1 \cdot \vec{a}_j - a'_{11}\, a'_{j1} \\ \vec{a}_2 \cdot \vec{a}_j - a'_{21} a'_{j1} \\ \cdots\cdots\cdots \\ \vec{a}_n \cdot \vec{a}_j - a'_{n1}\, a'_{j1} \end{bmatrix}$$

(*) Exceptuando os elementos da diagonal principal. Note-se que o anulamento de $R^*_h - R^*_{(1)}$ significaria que há apenas um factor comum, caindo-se na hipótese de Spearman.

cuja soma é

$$(\vec{a}_1 + \ldots + \vec{a}_n) \cdot \vec{a}_j - a'_{j1} (a'_{11} + \ldots + a'_{n1}) = a_{j1} S'_1 - a_{j1} S'_1 = 0 \text{ c.q.d.}$$

As primeiras variáveis residuais, y_j, podem ser representadas por vectores saturação do subespaço factorial $F'_1 = 0$ a q-1 dimensões(hiperplano coordenado do referencial $(0, F'_1, \ldots, F'_q)$). De facto,

$$y_j = x_j - a'_{j1} f_1 = a'_{j2} f_2 + \ldots + a'_{jq} f_q + b_j u_j$$

é representada pelo vector saturação com q-1 componentes

$$\vec{\alpha}_j = (a'_{j2}, \ldots, a'_{jq})$$

sendo, evidentemente,

$$R^*_h - R^*_{(1)} = \begin{bmatrix} \vec{\alpha}_1 \\ \vdots \\ \vec{\alpha}_n \end{bmatrix} \begin{bmatrix} \vec{\alpha}_1^T \ldots \vec{\alpha}_n^T \end{bmatrix}$$

(De outro modo, $\vec{\alpha}_j$ é a projecção de $\vec{a}_j$ no hiperplano $F'_1=0$)

O anulamento dos totais de coluna da matriz $R^*_h - R^*_{(1)}$ mostra que esta não pode ser tratada directamente como a matriz R^*_h.

A determinação da posição do segundo eixo factorial baricêntrico (extracção do segundo factor) não pode ser feita pela maximização da soma da primeira coluna da matriz $\begin{bmatrix} \vec{\alpha}_1 \\ \vdots \\ \vec{\alpha}_n \end{bmatrix}$ pois esta soma tem por valor $S'_2 = 0$

(note-se que a 1.ª coluna de $\begin{bmatrix} \vec{\alpha}_1 \\ \vdots \\ \vec{\alpha}_n \end{bmatrix}$ coincide com a 2.ª coluna de $\begin{bmatrix} \vec{a}_1 \\ \vdots \\ \vec{a}_n \end{bmatrix}$)

Reconhe-se ainda facilmente que o baricentro dos pontos $Q_1, \ldots, Q_n$ de vectores de posição $(Q_i - O) = \vec{\alpha}_i$, i=1,...,n é o ponto O (origem do referencial baricêntrico, ficando, portanto, indeterminada a posição de O F_2.

Para ultrapassar esta dificuldade, procede-se a "reflexões", isto é, à mudança de sinal em certas variáveis de modo a tornar os totais de coluna da matriz $R^*_h - R^*_{(1)}$ diferentes de zero.

Mais precisamente, procure-se tornar todos os totais de coluna do mesmo sinal de modo a dar o valor máximo ao somatório geral das intercorrelações e, portanto, ao somatório dos valores absolutos das saturações. Aumenta-se assim, num certo sentido, a "importância" do factor a extrair, implicando uma solução "económica" quanto ao número de factores.

Note-se que é legítimo inverter o sinal de todas as correlações implicando uma variável sob a condição de inverter em seguida o sinal da saturação (correlação) desta variável no factor a extrair.

Refira-se ainda que esta técnica de mudança de sinal pode ser necessária desde o início da análise, quando as intercorrelações não são todas positivas, como sucede, por exemplo, com variáveis ligadas a testes de personalidade(*).

Sobre o quadro dos valores reflectidos da matriz $R_h^* - R_{(1)}^*$ extrai-se então o segundo factor baricêntrico, ficando-se a conhecer a segunda coluna de F' : $F_2' = \begin{bmatrix} a'_{12} \\ a'_{22} \\ \vdots \\ a'_{n2} \end{bmatrix}$

A matriz $R_{(2)}^* = F_2' F_2'^T$ é a matriz das contribuições do segundo factor baricêntrico para a matriz R_h^*. Calcule-se em seguida a segunda matriz residual $R_h^* - R_{(1)}^* - R_{(2)}^*$. Se os elementos desta matriz não são significativamente diferentes de zero, termina-se a análise. Caso contrário continue-se sobre $R_h^* - R_{(1)}^* - R_{(2)}^*$ o processo indicado de modo a extrair o terceiro factor baricêntrico, e assim sucessivamente, até se atingir uma matriz residual não significativa.

(*) Como se sabe, a escala métrica de uma variável psicológica é arbitrária. Um teste de inteligência, por exemplo, permite unicamente uma ordenação das crianças a que foi aplicado, sendo indiferente a atribuição do número de ordem 1 ao mais ou ao menos "inteligente". Em qualquer caso, uma inversão no sentido da medida (inversão da escala) é acompanhada por uma inversão de sinal nas correlações desta variável com outras. Em particular, isto é verdade para a correlação (saturação) desta variável com uma variável hipotética , o factor.

Exemplos numéricos

Exemplo 5 (Reuchlin, 1964). Suponha-se dada uma matriz F das saturações de 4 variáveis em 2 factores independentes, o que permitirá construir artificialmente a tabela (matriz) das intercorrelações. A partir da matriz $R_h = F\,F^T$ vamos calcular as saturações das variáveis nos factores baricêntricos.

$$F = \begin{array}{c} \begin{matrix} f_1 & f_2 \end{matrix} \\ \begin{bmatrix} 0{,}40 & 0{,}70 \\ 0{,}60 & 0{,}40 \\ 0{,}10 & 0{,}80 \\ 0{,}50 & -\,0{,}50 \end{bmatrix} \end{array} \begin{matrix} \longrightarrow \text{ factores} \\ 1 \\ 2 \\ \downarrow \text{ variáveis} \\ 3 \\ 4 \end{matrix}$$

No referencial ortogonal $(0, F_1, F_2)$ é

$\vec{a}_1 = (0{,}40 \ ; \ 0{,}70)$

$\vec{a}_2 = (0{,}60 \ ; \ 0{,}40)$

$\vec{a}_3 = (0{,}10 \ ; \ 0{,}80)$

$\vec{a}_4 = (0{,}50 \ ; \ 0{,}50)$

Calculando $R_h = F\,F^T$ obtém-se a seguinte tabela

	1	2	3	4	
1	(0,65)	0,52	0,60	0,55	
2	0,52	(0,52)	0,38	0,50	
3	0,60	0,38	(0,65)	0,45	
4	0,55	0,50	0,45	(0,50)	
T (totais de coluna)	2,32	1,92	2,08	0,69	Σ T=8,32
$T/\sqrt{\Sigma T}$	0,80 (a'_{11})	0,67 (a'_{21})	0,72 (a'_{31})	0,69 (a'_{41})	

Os elementos da diagonal principal são as comunalidades das variáveis. Quando se trabalha sobre correlações experimentais não se conhece a diagonal principal, iniciando-se a análise por uma estimativa das comunalidades.

A última linha da tabela dá as saturações das variáveis no primeiro factor baricêntrico. A primeira coluna da matriz factorial baricêntrica, F', é então

$$F'_1 = \begin{bmatrix} a'_{11} \\ a'_{21} \\ a'_{31} \\ a'_{41} \end{bmatrix} = \begin{bmatrix} 0,80 \\ 0,67 \\ 0,72 \\ 0,69 \end{bmatrix}$$

$R_{(1)} = F'_1 F'^{T}_1$ dá as contribuições do primeiro factor baricêntrico para a matriz R_h ; $R_h - F'_1 F'^{T}_1$ é a primeira matriz residual

$$R_{(1)} = F'_1 F'^{T}_1 = \begin{bmatrix} 0,80 \\ 0,67 \\ 0,72 \\ 0,69 \end{bmatrix} \begin{bmatrix} 0,80 & 0,67 & 0,72 & 0,69 \end{bmatrix} = \begin{bmatrix} 0,64 & 0,54 & 0,58 & 0,55 \\ 0,54 & 0,45 & 0,48 & 0,46 \\ 0,58 & 0,48 & 0,52 & 0,50 \\ 0,55 & 0,46 & 0,50 & 0,48 \end{bmatrix}$$

$$R_h - R_{(1)} = \begin{bmatrix} 0,01 & -0,02 & 0,02 & 0,00 \\ -0,02 & 0,07 & -0,10 & 0,04 \\ 0,02 & -0,10 & 0,13 & -0,05 \\ 0,00 & 0,04 & -0,05 & 0,02 \end{bmatrix}$$

$$T = \begin{bmatrix} 0,01 & -0,01 & 0,00 & 0,01 \end{bmatrix}$$

A matriz linha T dá os totais de coluna de $R_h - R_{(1)}$. Esta linha, a menos de erros de arredondamento, é constituído por zeros, como se provou. O facto de aparecerem totais não nulos resulta de se terem usado arredondamentos apenas com duas casas decimais.

Para prosseguir a análise e extrair o segundo factor baricêntrico "reflictam-se" as variáveis 1 e 3 obtendo-se:

	1	2	3	4	
1	0,01	0,02	0,02	0,00	
2	0,02	0,07	0,10	0,04	
3	0,02	0,10	0,13	0,05	
4	0,00	0,04	0,05	0,02	
T	0,05	0,23	0,30	0,11	$\Sigma T=0,69$
$T/\sqrt{\Sigma T}$	0,06	0,28	0,36	0,13	
Saturações	-0,06 (a'_{12})	0,28 (a'_{22})	-0,36 (a'_{32})	0,13 (a'_{42})	

Repetindo o processo utilizado com R_h obtêm-se as duas últimas linhas da tabela.
A penúltima linha dando as saturações com sinal provisório (de acordo com as reflexões), a última linha dando as saturações com o sinal correspondente ao sentido da medição inicial das variáveis (sinal que deve prevalecer).

A segunda coluna de F' é então formada pelos elementos da última linha da tabela e

$$F' = \begin{matrix} \\ 0,80 \\ 0,67 \\ 0,72 \\ 0,69 \end{matrix} \begin{matrix} f'_1 \\ \left[\begin{matrix} \\ \\ \\ \\ \end{matrix}\right. \end{matrix} \begin{matrix} f'_2 \\ \begin{matrix} -0,06 \\ 0,28 \\ -0,36 \\ 0,13 \end{matrix} \left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right] \end{matrix} \quad \rightarrow \text{factores baricêntricos}$$

Calculemos a matriz das contribuições para R_h do segundo factor baricêntrico

$$R_{(2)} = F'_2 F'^{T}_2 = \begin{bmatrix} -0,06 \\ 0,28 \\ -0,36 \\ 0,13 \end{bmatrix} \begin{bmatrix} -0,06 & 0,28 & -0,36 & 0,13 \end{bmatrix} = \begin{bmatrix} 0,00 & -0,02 & 0,02 & -0,01 \\ -0,02 & 0,08 & -0,10 & 0,04 \\ 0,02 & -0,10 & 0,13 & -0,05 \\ -0,01 & 0,04 & -0,05 & 0,02 \end{bmatrix}$$

O cálculo da segunda matriz residual dá

$$R_h - R_{(1)} - R_{(2)} = \begin{bmatrix} 0,01 & 0 & 0 & 0,01 \\ 0 & -0,01 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0,01 & 0 & 0 & 0 \end{bmatrix}$$

verificando-se que esta matriz é nula, a menos de erros de arredondamento, o que está de acordo com o exemplo artificial de que se partiu.

Na prática, trabalhando com valores experimentais, não se obtêm resultados tão nítidos, havendo sempre incerteza quanto ao número de factores a extraír.

Uma verificação global dos cálculos feitos consiste em efectuar o produto $F' F'^{T}$ e compará-lo com a matriz R_h.

No caso presente é

$$F' F'^{T} = \begin{bmatrix} (0,64) & 0,52 & 0,60 & 0,54 \\ 0,52 & (0,53) & 0,38 & 0,50 \\ 0,60 & 0,38 & (0,65) & 0,45 \\ 0,54 & 0,50 & 0,45 & (0,49) \end{bmatrix} \simeq R_h$$

sendo a comparação evidentemente satisfatória.

Como se parte de uma matriz factorial dada, F, podem-se comparar as matrizes F e F', reconhecendo-se que são completamente diferentes. Na ausência de um critério, não é possível decidir qual é a melhor. Pode no entanto verificar-se gráfica e analiticamente a sua compatibilidade.

No referencial ortogonal $(0, F_1, F_2)$ considerem-se os pontos de coordenadas

$P_1 \equiv (0,4;0,7)$, $P_2 \equiv (0,6;0,4)$, $P_3 \equiv (0,1;0,8)$, $P_4 \equiv (0,5;0,5)$ —

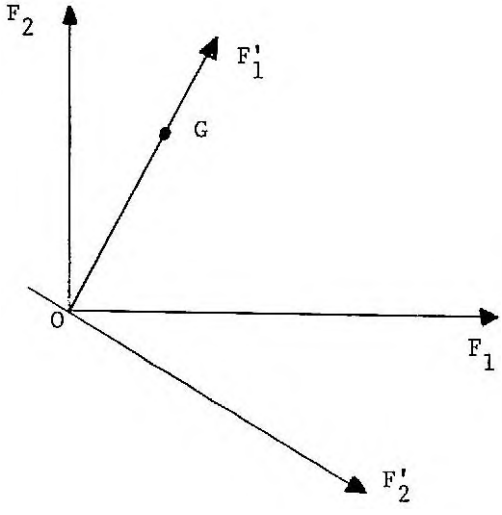
— extremidades dos vectores saturação supostos aplicados em 0.

O baricentro G de P_1, P_2, P_3, P_4 é o ponto de coordenadas

(em $(0, F_1, F_2)$)

$$\frac{0,4 + 0,6 + 0,1 + 0,5}{4} = 0,4$$

$$\frac{0,7 + 0,4 + 0,8 + 0,5}{4} = 0,6$$



verificando-se facilmente que as coordenadas de $P_1, \ldots, P_4$ no referencial $(0, F'_1, F'_2)$ são os elementos das linhas (vectores saturação) da matriz F', o que mostra a compatibilidade das soluções.

Por outras palavras, as soluções F e F' obtêm-se uma da outra por rotação. Existe então uma matriz ortogonal B tal que

$$F' = B\,F \qquad \text{ou} \qquad F = B^{-1}\,F' \qquad (B^{-1} = B^{T})$$

(Note-se que no eixo $O\,F_2'$ houve inversão no sentido das medições, o que implica que o referencial (O, F_1', F_2') é um referencial horário)

Exemplo 6. (Faverge 1972). Considere-se a seguinte tabela de correlações na qual se inseriram as comunalidades, supostas conhecidas,

	1	2	3	4	5	
1	(0,36)	0,42	0,30	0,00	0,12	
2	0,42	(0,49)	0,35	0,00	0,14	
3	0,30	0,35	(0,25)	0,00	0,10	(R_h^*)
4	0,00	0,00	0,00	(0,49)	0,56	
5	0,12	0,14	0,10	0,56	(0,68)	
T	1,20	1,40	1,00	1,05	1,60	$\sum T = 6{,}25$
$T/\sqrt{\sum T}$	0,48	0,56	0,40	0,42	0,64	

$$F_1' = \begin{bmatrix} 0{,}48 \\ 0{,}56 \\ 0{,}40 \\ 0{,}42 \\ 0{,}64 \end{bmatrix}$$

e $F_1' F_1'^{T} = R^*_{(1)} =$

$$\begin{bmatrix} 0{,}2304 & 0{,}2688 & 0{,}1920 & 0{,}2016 & 0{,}3072 \\ 0{,}2688 & 0{,}3136 & 0{,}2240 & 0{,}2352 & 0{,}3584 \\ 0{,}1920 & 0{,}2240 & 0{,}1600 & 0{,}1680 & 0{,}2560 \\ 0{,}2016 & 0{,}2352 & 0{,}1680 & 0{,}1764 & 0{,}2688 \\ 0{,}3072 & 0{,}3584 & 0{,}2560 & 0{,}2688 & 0{,}4096 \end{bmatrix}$$

$$R_h^* - R^*_{(1)} = \begin{bmatrix} 0{,}1296 & 0{,}1512 & 0{,}1080 & -0{,}2016 & -0{,}1872 \\ 0{,}1512 & 0{,}1764 & 0{,}1260 & -0{,}2352 & -0{,}2184 \\ 0{,}1080 & 0{,}1260 & 0{,}0900 & -0{,}1680 & -0{,}1560 \\ -0{,}2016 & -0{,}2352 & -0{,}1680 & 0{,}3136 & 0{,}2912 \\ -0{,}1872 & -0{,}2184 & -0{,}1560 & 0{,}2912 & 0{,}2704 \end{bmatrix}$$

$$T = \begin{bmatrix} 0{,}000 & 0{,}0000 & 0{,}0000 & 0{,}0000 & 0{,}0000 \end{bmatrix}$$

O facto de não se terem introduzido arredondamentos leva a que todos os elementos da matriz linha T sejam nulos.

Mudando o sinal às variáveis 4 e 5 obtém-se a nova matriz T (dos totais de coluna)

$$T = [0{,}7776 \quad 0{,}9072 \quad 0{,}6480 \quad 1{,}2096 \quad 1{,}1232]$$

com $\Sigma T_i = 4{,}6656$ e $T/\sqrt{\Sigma T_i} = T/2{,}16 = [0{,}36 \quad 0{,}42 \quad 0{,}30 \quad 0{,}56 \quad 0{,}52]$

Recordando que se "reflectiram" as variáveis 4 e 5, tem-se finalmente

$$F_2' = \begin{bmatrix} 0{,}36 \\ 0{,}42 \\ 0{,}30 \\ -0{,}56 \\ -0{,}52 \end{bmatrix}$$

verificando-se facilmente que a segunda matriz residual, $R_h^* - R_{(1)}^* - R_{(2)}^*$ ($R_{(2)}^* = F_2' F_2'^T$), é nula, o que implica o termo da análise. A matriz factorial baricêntrica é

$$F' = \begin{bmatrix} 0{,}48 & 0{,}36 \\ 0{,}56 & 0{,}42 \\ 0{,}40 & 0{,}30 \\ 0{,}42 & -0{,}56 \\ 0{,}64 & -0{,}52 \end{bmatrix}$$

obtendo-se as seguintes equações factoriais baricêntricas

$$x_1 = 0{,}48\, f_1' + 0{,}36\, f_2' + b_1 u_1$$

$$x_2 = 0{,}56\, f_1' + 0{,}42\, f_2' + b_2 u_2$$

$$x_3 = 0{,}40\, f_1' + 0{,}30\, f_2' + b_3 u_3$$

$$x_4 = 0{,}42\, f_1' - 0{,}56\, f_2' + b_4 u_4$$

$$x_5 = 0{,}64\, f_1' - 0{,}52\, f_2' + b_5 u_5$$

3.11. O problema da estimação das comunalidades

Nos exemplos anteriores, as comunalidades eram supostas conhecidas ou calculáveis a partir do produto $F\,F^T$, dada a matriz factorial F.

No entanto, na prática, as comunalidades são desconhecidas, iniciando-se a análise pela estimação destas. Uma primeira aproximação grosseira,que servirá como ponto de partida para se obter novas aproximações, é

$$\hat{h}_j = \max_k r_{jk}$$

Retome-se o exemplo numérico 6, supondo agora as comunalidades não conhecidas. Preencha-se a diagonal principal com as correlações de maior valor absoluto de cada coluna (ou linha), isto é, na posição jj ponha-se o valor $\max_k r_{jk}$.

Obtém-se assim uma primeira matriz R_h^* :

	1	2	3	4	5	
1	(0,42)	0,42	0,30	0,00	0,12	
2	0,42	(0,42)	0,35	0,00	0,14	
3	0,30	0,35	(0,35)	0,00	0,10	
4	0,00	0,00	0,00	(0,56)	0,56	
5	0,12	0,14	0,10	0,56	(0,56)	
T	1,26	1,33	1,10	1,12	1,48	$\sum T = 6,29$
$T/\sqrt{\sum T}$	0,502	0,530	0,439	0,447	0,590	
	(a'_{11})	(a'_{21})	(a'_{31})	(a'_{41})	(a'_{51})	

$$\therefore \quad F'_1 = \begin{bmatrix} 0,502 \\ 0,530 \\ 0,439 \\ 0,447 \\ 0,590 \end{bmatrix}$$

Calculando-se sucessivamente

$$F'_1 F'^{T}_1 = \begin{bmatrix} 0,252 & 0,266 & 0,220 & 0,224 & 0,296 \\ 0,266 & 0,281 & 0,233 & 0,237 & 0,313 \\ 0,220 & 0,233 & 0,193 & 0,196 & 0,259 \\ 0,224 & 0,237 & 0,196 & 0,200 & 0,264 \\ 0,296 & 0,313 & 0,259 & 0,264 & 0,348 \end{bmatrix} = R^*_{(1)}$$

$$R_h^* - R_{(1)}^* = \begin{bmatrix} (0,168) & 0,154 & 0,080 & -0,224 & -0,176 \\ 0,154 & (0,193) & 0,117 & -0,237 & -0,173 \\ 0,080 & 0,117 & (0,157) & -0,196 & -0,159 \\ -0,224 & -0,237 & -0,196 & (0,360) & 0,296 \\ -0,176 & -0,173 & -0,159 & 0,296 & (0,212) \end{bmatrix}$$

$$T = \begin{bmatrix} 0,002 & 0,000 & -0,001 & -0,001 & 0,000 \end{bmatrix}$$

A linha T serve para controlo dos cálculos, devendo ser constituída por zeros, a menos de erros de arredondamento.

Em $R_h^* - R_{(1)}^*$, reflectindo as variáveis 4 e 5 (isto é mudando o sinal às variáveis residuais $y_4 = x_4 - a'_{41}f_1$ e $y_5 = x_5 - a'_{51}f_1$) e substituindo os valores da diagonal — valores resultantes de uma estimação grosseira das comunalidades — pelos elementos de maior valor absoluto em cada coluna, obtém-se

	1	2	3	4	5	
1	(0,224)	0,154	0,080	0,224	0,176	
2	0,154	(0,237)	0,117	0,237	0,173	
3	0,080	0,117	(0,196)	0,196	0,159	
4	0,224	0,237	0,196	(0,296)	0,296	
5	0,176	0,173	0,159	0,296	(0,296)	
T	0,858	0,918	0,748	1,249	1,100	$\sum T=4,873$
$T/\sqrt{\sum T}$	0,389	0,416	0,339	0,566 (-)	0,498 (-)	

e . . $F'_2 = \begin{bmatrix} 0,389 \\ 0,416 \\ 0,339 \\ -0,566 \\ -0,498 \end{bmatrix}$

$$F'_2 F'^{T}_2 = \begin{bmatrix} (0,151) & 0,162 & 0,132 & -0,220 & -0,194 \\ 0,162 & (0,173) & 0,141 & -0,235 & -0,207 \\ 0,132 & 0,141 & (0,115) & -0,192 & -0,169 \\ -0,220 & -0,235 & -0,192 & (0,320) & 0,282 \\ -0,194 & -0,207 & -0,169 & 0,282 & (0,248) \end{bmatrix} = R^*_{(2)}$$

$$R^*_h - R^*_{(1)} - R^*_{(2)} = \begin{bmatrix} (0,073) & -0,008 & -0,052 & -0,004 & 0,018 \\ -0,008 & (0,064) & -0,024 & -0,002 & 0,034 \\ -0,052 & -0,024 & (0,081) & -0,004 & 0,010 \\ -0,004 & -0,002 & -0,004 & (-0,024) & 0,014 \\ 0,018 & 0,034 & 0,010 & 0,014 & (0,048) \end{bmatrix}$$

Todos os elementos não diagonais desta última matriz têm valor absoluto pequeno o que permite concluir, em princípio, que correspondem a correlações não significativas. A análise fica então limitada a dois factores, tendo-se a seguinte matriz factorial provisória

$$F' = \begin{bmatrix} 0,502 & 0,389 \\ 0,530 & 0,416 \\ 0,439 & 0,339 \\ 0,447 & -0,566 \\ 0,590 & -0,498 \end{bmatrix} = \begin{bmatrix} \vec{a}_1 \\ \vec{a}_2 \\ \vec{a}_3 \\ \vec{a}_4 \\ \vec{a}_5 \end{bmatrix}$$

Podem-se calcular agora estimativas mais precisas das comunalidades:

$$h_1^2 = \vec{a}_1 . \vec{a}_1 = (0,502)^2 + (0,389)^2 = 0,4$$

$$h_2^2 = \vec{a}_2 . \vec{a}_2 = 0,45 \quad , \quad h_3^2 = \vec{a}_3 . \vec{a}_3 = 0,31$$

$$h_4^2 = \vec{a}_4 . \vec{a}_4 = 0,52 \quad , \quad h_5^2 = \vec{a}_5 . \vec{a}_5 = 0,60$$

estimativas que, como se vê, diferem das primitivas. Inserindo estes valores na diagonal da matriz R^* obtem-se uma nova matriz R^*_h com a qual se

inicia uma segunda análise, não sendo agora necessário reestimar as comunalidades em cada etapa, na medida em que se partiu de estimativas mais aproximadas.

Este processo repete-se até que os valores das comunalidades estabilizem aproximando-se os resíduos de zero.

Ao método iterativo acabado de descrever dá-se o nome de método das aproximações sucessivas.

Outro método de estimação inicial das comunalidades consiste essencialmente numa análise baricêntrica parcial (isto é, usando apenas uma parte das variáveis) - método baricêntrico de estimação.

Considere-se um pequeno subconjunto das variáveis incluindo aquela cuja comunalidade se pretende estimar, sendo as restantes fortemente correlacionadas com esta. A comunalidade da variável no conjunto total das variáveis é, em geral, pouco diferente da comunalidade obtida a partir dum factor geral (primeiro factor baricêntrico) relativo ao sub-conjunto de variáveis escolhido. Calcule-se então o quadrado da saturação da variável no primeiro factor baricêntrico da tabela parcial.

Retome-se a tabela do exemplo numérico 6.. Para estimar h_1^2 construa-se a tabela parcial incluindo as variáveis 1, 2 e 3 (2 e 3 são os testes mais fortemente correlacionados com 1)

	1	2	3	
1	(0,42)	0,42	0,30	
2	0,42	(0,42)	0,35	
3	0,30	0,35	(0,35)	
	1,14	1,19	1,00	$\sum T=3,33$

A diagonal principal foi preenchida com as correlações de maior valor absoluto de cada coluna.

A comunalidade h_1^2 é então estimada por

$$\hat{h}_1^2 = \frac{(1,14)^2}{3,33} = 0,39$$

Compare-se este valor — 0,39 — com o valor 0,4, obtido a partir do método das aproximações sucessivas e com o valor conhecido 0,36.

Note-se que as estimativas obtidas pelo método baricêntrico podem ser usadas, com vantagem, como estimativas iniciais do método de aproximações sucessivas, em lugar das estimativas $\hat{h}_j^2 = \max_k r_{jk}$

Guttman[38] mostrou que a melhor estimativa, num dado sentido, da comunalidade de uma variável é o quadrado da correlação múltipla desta variável com o conjunto das restantes variáveis em estudo.

De facto, retomemos a variável $x_j = \sum_{\ell=1}^{q} a_{j\ell} f_\ell + b_j u_j$ e seja $x'_j = \sum_{\ell=1}^{q} a_{j\ell} f_\ell$ a parte "comum" de x_j; x'_j é a melhor estimação linear de x_j em função de $f_1,\ldots,f_q$. O cálculo da correlação entre x_j e x'_j dá:

$$\rho(x_j,x'_j) = \frac{\operatorname{cov}(x_j,x'_j)}{\sigma_{x_j}\sigma_{x'_j}} = \frac{E(x_j x'_j)}{\sqrt{1 \cdot h_j^2}} = \frac{1}{h_j} E\left[\left(\sum_{\ell=1}^{q} a_{j\ell} f_\ell + b_j u_j\right)\left(\sum_{\ell=1}^{q} a_{j\ell} f_\ell\right)\right] =$$

$$= \frac{1}{h_j} E\left[\left(\sum_{\ell=1}^{q} a_{j\ell} f_\ell\right)^2\right] = \frac{1}{h_j} v\left(\sum_{\ell} a_{j\ell} f_\ell\right) = \frac{1}{h_j} \sum_{\ell} a_{j\ell}^2 = \frac{h_j^2}{h_j} = h_j$$

isto é, a raiz quadrada da comunalidade é igual à correlação entre a variável e a sua parte comum. Por outro lado, a correlação simples entre x_j e $x'_j = \sum_{\ell} a_{j\ell} f_\ell$ é, por definição, a correlação múltipla entre x_j e $f_1,\ldots,f_q$.

Portanto,
$$\rho^2(x_j,x'_j) = \rho^2\left[x_j(f_1,\ldots,f_q)\right] = h_j^2$$

Como as ligações entre os x'_j e os f_ℓ são lineares, tem-se ainda

$$\rho^2\left[x_j(f_1,\ldots,f_q)\right] = \rho^2\left[x_j(x'_1,\ldots,x'_j,\ldots,x'_n)\right] \approx$$
$$\approx \rho^2\left[x_j(x_1,\ldots,x_j,\ldots,x_n)\right].$$

Então, uma estimativa de $\rho^2\left[x_j(x_1,\ldots,x_n)\right]$ é uma estimativa de h_j^2, c.q.d.

---

(38) L. Guttman, "Une solution au problème des communautés", Bull.C.E.R.P. 1956, 5, 123-128.

3.12. Determinação do número de factores

Após um número suficiente de iterações, as comunalidades estabilizam e os resíduos aproximam-se de zero. Impõe-se, então, um critério que permita decidir se uma dada matriz residual é, ou não, significativamente diferente da matriz nula (matriz com todos os seus elementos nulos) determinando-se, deste modo, o número de factores a extraír. Por outras palavras, é necessário testar se os valores residuais se podem considerar como flutuações aleatórias em torno de zero, atribuíveis aos erros(*) de amostragem.

McNemar(39) propõe o seguinte critério empírico:

— a partir da matriz residual calcula-se a quantidade

$$\frac{s_r}{1 - \bar{h}^2}$$

sendo s_r o desvio padrão dos resíduos r_{ij} da matriz e $\bar{h}^2$ a média das comunalidades definitivas, isto é, calculadas a partir das saturações estimadas. Sendo N a dimensão da amostra sobre a qual foram calculadas as correlações r_{jk}, se $\frac{s_r}{1 - \bar{h}^2} \leq \frac{1}{\sqrt{N}}$

considera-se que o número de factores extraídos é suficiente.

Lawley(40) estabelece um critério, válido para amostras grandes, baseado num teste estatístico, aplicável a partir das comunalidades estabilizadas por iteração. Seja r_{ij} o elemento genérico da matriz dos resíduos. A variável

$$X^2 = (N - 1)\sum_{i<j} \frac{r^2_{ij}}{(1-h_i^2)(1-h_j^2)} = (N - 1)\sum_{i<j} \frac{r^2_{ij}}{b_i^2\, b_j^2}$$

tem, aproximadamente, uma distribuição χ^2 com $q = \frac{1}{2}\left[(n-q)^2-(n+q)\right]$ graus de liberdade. (n = nº de variáveis; q = nº de factores).(41)

(*) Erro no sentido estatíst. de variabil. aleatória(ou devida ao acaso).

(39) McNEMAR,Q. - On the number of factors,Psychometrika, 1942, 7, 9-18.

(40) LAWLEY,D. - A statistical examination of the centroid method, Proc. Roy. Soc. Edimb., Secçao A, 1957, 175-189.

(41) Para uma forma mais rigorosa de χ^2, veja-se, por exemplo, Torrens-Ibern, ob. cit. na bib., p.79 .

Se $X^2 > \chi^2_q(\alpha)$, sendo α o nível de significância escolhido, deve-se extraír, pelo menos, mais um factor.

Apêndice I

Quadro resumo das medidas de associação entre duas variáveis

X \ Y	Dicotômica (nominal ou categorial)	Dicotomisada (normal subjacente)	Tricotomisada (normal subjacente)	Discreta finita (nominal ou categorial)	Ordinal	Contínua (intervalos ou razões)	Contínua e Gaussiana
Dicotômica (nominal ou categorial)	Correlação φ^* ou correl. pont. 2x2 $\varphi^* = \frac{f_{11}f_{22} - f_{12}f_{21}}{\sqrt{f_{1.}f_{2.}f_{.1}f_{.2}}}$						
Dicotomisada (normal subjacente)		Coef. tetracórico r_{tet} (Bonnardel) ($\varphi^* < r_{tet}$)					
Tricotomisada (normal subjacente)			Coef. eneacórico (Coumetou)				
Discreta finita (nominal ou categorial)				Contig. quadrada média (K. Pearson) $\varphi^{*2} = \sum_{i,j} \frac{f_{ij}^2}{f_{i.}f_{.j}} - 1$ e razão de correl. ζ^2		Razão de correlaç. $\zeta^{*2}_{yx} = 1 - \frac{s_r^2}{s_y^2}$	Razão de correlaç. ζ^{*2}_{yx}
Ordinal					Correlação ordinal Spearman: $r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$ Kendall: $\tau = \frac{2S}{\binom{n}{2}}$	Razão de correlaç. ζ^{*2}_{yx}	
Contínua (Intervalos ou razões)						Razão de correlaç. $\zeta^{*2}_{yx} = 1 - \frac{s_r^2}{s_y^2}$ (ou ζ^{*2}_{xy})	
Contínua e Gaussiana	Biserial pontual $r_{bp} = \frac{\bar{x}_p - \bar{x}_q}{s_x}\sqrt{pq}$ (r_b > r_{bp}) ⟶	Biserial $r_b = \frac{\bar{x}_p - \bar{x}_q}{s_x}\frac{pq}{z}$ ($r_b = r_{bp}\frac{\sqrt{pq}}{z}$)	Triserial (Burt) $r_t = \frac{\bar{x}_{p'} - \bar{x}_{p''}}{s_x} \times \frac{1}{z'/p' + z''/p''}$				Coef. de correlaç. de Bravais-Pearson $r = \frac{s_{12}}{s_1 s_2}$ ($r^2 = \zeta^{*2}_{yx}$)

Apêndice II

O coeficiente de fiabilidade

Erros de medida

Dada uma população (universo) U, seja X uma característica mensurável dos indivíduos de U (em geral, e no caso que nos interessa, as pontuações de um teste); X é uma variável aleatória definida sobre U. Escolhido ao acaso $u \in U$ seja $X(u)=x$ o valor observado de X para o indivíduo u. Pondo $X=\xi+\varepsilon$, com ξ o valor verdadeiro de X e ε o erro (absoluto) da observação, X aparece como a soma de duas variáveis não observáveis.

No modelo clássico para testes de comprimento fixo admite-se a hipótese de que ε é uma variável aleatória de valor médio nulo e não correlacionada com ξ. As hipóteses do modelo clássico podem então exprimir-se pelas relações seguintes (válidas em U)

$$E(\varepsilon)=0$$
$$\mathrm{cov}(\varepsilon,\xi)=0$$

Dadas duas medições (variáveis) sobre U, X_1 e X_2, com

$$X_1 = \xi_1 + \varepsilon_1 \quad , \qquad X_2 = \xi_2 + \varepsilon_2 \quad \text{é}$$
$$\mathrm{cov}(\varepsilon_1,\varepsilon_2) = 0$$
$$\mathrm{cov}(\varepsilon_i,\xi_j) = 0 \quad , \quad i,j=1,2$$

Das hipóteses do modelo resulta imediatamente que

$$E(X) = E(\xi) = \mu \qquad (1)$$
$$v(X) = v(\xi) + v(\varepsilon) \qquad (2)$$
$$(\text{ou} \quad \sigma_x^2 = \sigma_\xi^2 + \sigma_\varepsilon^2)$$

(1) exprime o facto de que o valor médio observado é igual ao valor médio verdadeiro; (2) mostra que a variância das pontuações observadas é igual à variância das pontuações verdadeiras mais a variância de erro.

O cálculo da covariância entre X e ξ dá

$$\mathrm{cov}(X,\xi) = E(X\xi) - E(X)E(\xi) = E[(\xi+\varepsilon)\xi] - E^2(\xi) = E(\xi^2) - E^2(\xi) = v(\xi) = \sigma_\xi^2 \qquad (3)$$

e $\rho_{x\xi} = \frac{cov(X,\xi)}{\sigma_x \sigma_\xi} = \frac{\sigma_x^2}{\sigma_x \sigma_\xi} = \frac{\sigma_\xi}{\sigma_x}$; $\rho^2_{x\xi} = \frac{\sigma^2_\xi}{\sigma^2_x} = \frac{v(\xi)}{v(X)}$ (4)

. . o quadrado do coeficiente de correlação entre as pontuações observadas e as pontuações verdadeiras é igual à razão das variâncias das pontuações verdadeiras e das pontuações observadas.

Como $v(X) = v(\xi) + v(\varepsilon)$ é $v(\xi) = v(X) - v(\varepsilon)$ e (4) toma o aspecto

$$\rho^2_{x\xi} = 1 - \frac{v(\varepsilon)}{v(X)} \quad (5)$$

Medições (testes ou formas) paralelas

Duas medições $X = \xi + \varepsilon$ e $X' = \xi' + \varepsilon'$ sobre U dizem-se paralelas se $\xi = \xi'$ e $v(\varepsilon) = v(\varepsilon')$.

Desta definição resulta imediatamente que

$$E(X) = E(X')$$
$$v(X) = v(X')$$

Por outro lado $cov(X,X') = \sigma^2_\xi$ (6)

e $\rho_{xx'} = \frac{\sigma^2_\xi}{\sigma_x \sigma_{x'}} = \frac{\sigma^2_\xi}{\sigma^2_x} \geqslant 0$ (7)

Comparando (4) e (7) vem

$$\rho^2_{x\xi} = \rho_{xx'} \quad (8)$$

isto é, o quadrado da correlação entre as pontuações observadas e as pontuações verdadeiras é igual à correlação entre medições paralelas.

Este resultado permite exprimir a quantidade não observável $\rho^2_{x\xi}$ em função da quantidade $\rho_{xx'}$, parâmetro de uma distribuição (bivariada) observável.

Analogamente, (6) permite exprimir σ^2_ξ (quantidade não observável) como a covariância de duas medições paralelas (quantidade potencialmente observável).

De igual modo, a partir de (2) e de (6) é possível exprimir a variância do erro (não observável directamente) em termos de quantidades observáveis. De facto,

$$v(\varepsilon) = v(X) - v(\hat{\xi}) = v(X) - \mathrm{cov}(X,X') = v(X) - \rho_{xx'}\, v(X) = v(X)(1-\rho_{xx'}) \qquad (9)$$

$$\text{e} \quad \sigma_\varepsilon = \sigma_x \sqrt{1-\rho_{xx'}} \qquad (10)$$

$$\text{Como} \quad \rho_{x\varepsilon} = \frac{\sigma_\varepsilon}{\sigma_x} \quad \text{de (10) vem} \quad \rho_{x\varepsilon} = \sqrt{1-\rho_{xx'}} \qquad (11)$$

Demonstra-se facilmente o seguinte

Teorema: Sejam $X_1, X_2, \ldots$ medições paralelas e Y uma medição arbitrária distinta. Então,

$$\rho_{x_1x_2} = \rho_{x_1x_3} = \ldots = \rho_{x_2x_3} = \ldots$$

$$\text{e} \quad \rho_{x_1y} = \rho_{x_2y} = \ldots$$

As mesmas igualdades são válidas para as covariâncias. Em notação mais compacta tem-se

$$\rho_{x_i x_j} = \text{const.} \quad \forall\, i,j \quad , \; i \neq j$$

$$\rho_{x_i y} = \text{const.} \quad \forall\, i$$

$$\mathrm{cov}(X_i, X_j) = \text{const.} \quad \forall\, i,j \quad , \quad i \neq j$$

$$\mathrm{cov}(X_i, Y) = \text{const.} \quad \forall\, i$$

Em resumo: as intercorrelações entre testes paralelos são todas iguais e testes paralelos têm iguais correlações com qualquer outro teste.

Fiabilidade de um teste

O coeficiente de fiabilidade, R, de um teste define-se como o quadrado da correlação entre as pontuações observadas e as pontuações verdadeiras:

$$R = \rho^2_{x\hat{\xi}} = \frac{v(\hat{\xi})}{v(X)} \leq 1 \qquad (12)$$

O coeficiente de fiabilidade aparece como uma medida do grau de variação das pontuações verdadeiras relativamente à variação das pontuações observadas. Por outras palavras, R indica a proporção da variância calculada que corresponde à variância verdadeira, sendo a proporção complementar atribuível a erros de medida.

Pondo $R = \rho^2_{x\xi} = 1 - \frac{v(\varepsilon)}{v(X)}$ vê-se que, se a variância do erro fôr nula, a fiabilidade do teste é igual a 1 (por outras palavras, as medições são feitas sem erro: de facto, se $v(\varepsilon)=0$, a variável aleatória ε é constantemente igual a $E(\varepsilon)=0$).

Se $v(\varepsilon)=v(X)$, então, $v(\xi)=0$, anulando-se a fiabilidade do teste. Exceptuando estes casos limites, a fiabilidade é um número entre 0 e 1. Por exemplo, $R=0,90$ significa que 90% da variância observada, $v(X)$, corresponde à variância verdadeira $v(\xi)$; os 10% restantes resultam de erros de medição.

De acordo com (4), (7) e (8) tem-se

$$R = \rho^2_{x\xi} = \frac{v(\xi)}{v(X)} = \rho_{xx'} \qquad (13)$$

com X e X' testes paralelos.

A relação (13) evidencia três maneiras de definir o coeficiente de fiabilidade. Usando $\rho_{xx'}$, a fiabilidade aparece como a correlação entre duas medições (ou duas formas paralelas) do mesmo teste.

Por outras palavras, a <u>fiabilidade</u> exprime em que medida duas aplicações de um teste, a um indivíduo, dão o mesmo resultado.

A definição $R = \rho^2_{x\xi}$ põe a ênfase no facto de que a fiabilidade pode ser definida sem recurso ao conceito de medições paralelas.

A $\sqrt{R} = \rho_{x\xi} = \frac{\sigma_\xi}{\sigma_x}$ (correlação entre as pontuações observadas e verdadeiras) dá-se o nome de <u>índice de fiabilidade</u>.

<u>Testes compostos: 2 componentes</u>

Sejam $Y_1 = \eta_1 + \varepsilon_1$ e $Y_2 = \eta_2 + \varepsilon_2$ duas medições e $X = Y_1 + Y_2$ a medição (teste) composta. Pondo $X = \xi + \eta$ é $\xi = \eta_1 + \eta_2$, $\varepsilon = \varepsilon_1 + \varepsilon_2$, e

$$E(X) = E(Y_1)+E(Y_2) \qquad (14)$$

$$v(X) = v(Y_1)+v(Y_2)+2\ cov(Y_1,Y_2) \qquad (15)$$

$$v(\xi) = v(\eta_1)+v(\eta_2)+2\ cov(\eta_1,\eta_2) \qquad (16)$$

$$v(\varepsilon) = v(\varepsilon_1)+v(\varepsilon_2) \qquad (17)$$

Em particular, sejam Y e Y' duas medições paralelas (isto é, $Y = \xi + \varepsilon_1$, $Y' = \xi' + \varepsilon_1'$, $v(\varepsilon_1) = v(\varepsilon_1')$) e $X = Y + Y'$

De (15), (16) e (17) vem

$$v(X) = 2v(Y)(1+\rho_{yy'}) \qquad (15')$$

$$v(\eta) = 4v(\xi) \qquad (16')$$

$$v(\varepsilon) = 2v(\varepsilon_1) \qquad (17')$$

Sejam Y, Y', Y_1, Y_1' testes paralelos e construam-se os testes compostos $X = Y + Y_1$, $X' = Y' + Y_1'$

Teorema. a) X e X' são testes paralelos

b) A fiabilidade de X é $\rho_{xx'} = \dfrac{2\rho_{yy'}}{1+\rho_{yy'}}$ (18)

Dem: a) Pondo $X = \eta + \varepsilon_x$, $X' = \eta' + \varepsilon_{x'}$

é $\eta = \xi + \xi_1$, $\eta' = \xi' + \xi_1'$

e como $\xi = \xi_1 = \xi' = \xi_1'$ segue-se que $\eta = \eta'$.

Por outro lado,

$v(\varepsilon_x) = v(\varepsilon) + v(\varepsilon_1) = v(\varepsilon') + v(\varepsilon_1') = v(\varepsilon_{x'})$ c.q.d.

b) $\rho_{xx'} = \dfrac{\operatorname{cov}(X,X')}{\sigma_x \sigma_{x'}} = \dfrac{\operatorname{cov}(Y+Y_1, Y'+Y'_1)}{v(X)} = \dfrac{4\operatorname{cov}(Y,Y')}{v(Y+Y_1)} =$

$$= \frac{4\rho_{yy'}\sigma_y\sigma_{y'}}{v(Y)+v(Y_1)+2\operatorname{cov}(Y,Y_1)} = \frac{4\rho_{yy'}v(Y)}{2v(Y)+2\rho_{yy'}v(Y)} =$$

$$= \frac{2\rho_{yy'}}{1+\rho_{yy'}} \quad \text{c.q.d.}$$

(De outro modo, como X e X' são medições paralelas, de (7) vem $\rho_{xx'} = \dfrac{v(\eta)}{v(X)}$ e de (15') e (16') vem, respectivamente, $v(X) = 2v(Y)(1+\rho_{yy'})$ e $v(\eta) = 4v(\xi) = 4\operatorname{cov}(Y,Y') =$

$= 4v(Y)\rho_{yy'}$, obtendo-se imediatamente (18)).

A fórmula (18) exprime a fiabilidade $\rho_{xx'}$, de um teste composto com duas componentes paralelas, em função da fiabilidade comum, $\rho_{yy'}$, das componentes.

$R = \rho_{xx'} = \frac{2\rho_{yy'}}{1 + \rho_{yy'}}$ é a fórmula de Spearman-Brown para a fiabilidade de um teste de comprimento duplo.

Como $\rho_{yy'} < 1$ é $\frac{2}{1 + \rho_{yy'}} > 1$

e $\therefore \frac{\rho_{xx'}}{\rho_{yy'}} = \frac{2}{1 + \rho_{yy'}} > 1$ isto é, $\rho_{xx'} > \rho_{yy'}$ (19)

(19) mostra que a duplicação do comprimento de um teste implica um aumento da fiabilidade.

Testes compostos: k componentes

Seja agora $X = \sum_{j=1}^{k} Y_j = \sum_{j=1}^{k} (\zeta_j + \varepsilon_j)$

com $\xi = \sum_j \zeta_j$, $\varepsilon = \sum_j \varepsilon_j$ obtendo-se imediatamente

$$E(X) = \sum_j E(Y_j) = E(\xi) \tag{20}$$

$$E(\varepsilon) = \sum_j E(\varepsilon_j) = 0 \tag{21}$$

$$v(X) = \sum_j v(Y_j) + 2 \sum_{\ell<m} \sum \operatorname{cov}(Y_\ell, Y_m) \tag{22}$$

$$v(\xi) = \sum_j v(\zeta_j) + 2 \sum_{\ell<m} \sum \operatorname{cov}(\zeta_\ell, \zeta_m) \tag{23}$$

$$v(\varepsilon) = \sum_j v(\varepsilon_j) \tag{24}$$

Sendo as k medições paralelas, tem-se

$$v(Y_j) = v(Y) \quad , \quad v(\zeta_j) = v(\zeta) \quad , \quad j=1,2,\ldots,k$$

$$\operatorname{cov}(Y_\ell, Y_m) = \text{const.} = \operatorname{cov}(Y, Y') = v(Y)\,\rho_{yy'}$$

e

$$v(X) = k\, v(Y)\left[1 + (k-1)\rho_{yy'}\right] \tag{22'}$$

$$v(\xi) = k^2 v(\zeta) \tag{23'}$$

$$v(\varepsilon) = k\, v(\varepsilon_1) \tag{24'}$$

De (23') e (24') conclui-se que a variância do valor verdadeiro cresce mais rapidamente que a variância do erro, o que mostra que é vantajoso aumentar o comprimento de um teste.

O coeficiente α e a fiabilidade de testes compostos

Retome-se o teste composto com duas componentes, $X = \xi + \varepsilon = Y_1 + Y_2$ com $Y_1 = \eta_1 + \varepsilon_1$, $Y_2 = \eta_2 + \varepsilon_2$. De (4) e (6) tem-se

$$\rho_{xx'} = \rho^2_{x\xi} = \frac{v(\xi)}{v(X)} = \frac{v(\eta_1) + v(\eta_2) + 2\,\mathrm{cov}(\eta_1,\eta_2)}{v(X)} \qquad (25)$$

Teorema. a) $$\rho^2_{x\xi} = R \geqslant 2\left[1 - \frac{v(Y_1) + v(Y_2)}{v(X)}\right] \qquad (26)$$

b) Em particular, se $Y_1 = Y$ e $Y_2 = Y'$ são testes paralelos, tem-se

$$R = \rho^2_{x\xi} = \frac{2\rho_{yy'}}{1 + \rho_{yy'}} \qquad (27)$$

Dem.: Considere-se a diferença $(\sigma_{\eta 1} - \sigma_{\eta 2})^2 =$

$$= v(\eta_1) + v(\eta_2) - 2\sigma_{\eta 1}\sigma_{\eta 2} \geqslant 0$$

ou $v(\eta_1) + v(\eta_2) \geqslant 2\,\sigma_{\eta 1}\sigma_{\eta 2}$

mas, pela desigualdade de Schwartz, $\left|\mathrm{cov}(\eta_1,\eta_2)\right| \leqslant \sigma_{\eta 1}\sigma_{\eta 2}$

$$\therefore\ v(\eta_1)+v(\eta_2) \geqslant 2\sigma_{\eta 1}\sigma_{\eta 2} \geqslant 2\left|\mathrm{cov}(\eta_1,\eta_2)\right| \geqslant 2\,\mathrm{cov}(\eta_1,\eta_2) \qquad (28)$$

Então,

$$v(\xi) = v(\eta_1)+v(\eta_2)+2\,\mathrm{cov}(\eta_1,\eta_2) \geqslant 4\,\mathrm{cov}(\eta_1,\eta_2) = 4\,\mathrm{cov}(Y_1,Y_2)$$

e $$\rho^2_{x\xi} \geqslant \frac{4\,\mathrm{cov}(Y_1,Y_2)}{v(X)} = 2.\frac{2\,\mathrm{cov}(Y_1,Y_2)}{v(X)} = 2.\frac{v(X)-v(Y_1)-v(Y_2)}{v(X)} =$$

$$= 2\left[1 - \frac{v(Y_1)+v(Y_2)}{v(X)}\right] \quad \text{c.q.d.}$$

A generalização para um teste composto com k componentes é imediata. Com $X = \xi + \varepsilon = \sum_{j=1}^{k} Y_j = \sum_{j=1}^{k}(\eta_j + \varepsilon_j)$ de (23) vem

$$\rho^2_{x\xi} = \frac{v(\xi)}{v(X)} = \frac{\sum_j v(\eta_j) + 2\sum_{\ell<m}\sum_{m}\mathrm{cov}(\eta_\ell,\eta_m)}{v(X)} \qquad (29)$$

tendo lugar o

Teorema. a) $$\rho^2_{x\xi} = R \geqslant \frac{k}{k-1}\left[1 - \frac{\sum_{j=1}^{k} v(Y_j)}{v(X)}\right] \qquad (30)$$

b) Em particular, se Y_j , $j=1,2,\ldots k$ são testes paralelos, tem-se

$$R = \rho^2_{x\xi} = \frac{k\rho_{yy'}}{1-(k-1)\rho_{yy'}} \qquad (31)$$

Dem.: De (28) tem-se $\forall$ ℓ e m

$$v(\xi_\ell) + v(\xi_m) \geq 2\,\mathrm{cov}(\xi_\ell, \xi_m)$$

Somando em ℓ e m vem

$$\sum_\ell \sum_m \left[v(\xi_\ell)+v(\xi_m)\right] = 2k \sum_j v(\xi_j) \geq 2 \sum_\ell \sum_m \mathrm{cov}(\xi_\ell, \xi_m)$$

isto é,

$$k \sum_j v(\xi_j) \geq \sum_\ell \sum_m \mathrm{cov}(\xi_\ell, \xi_m) = \sum_{\ell \neq m} \sum \mathrm{cov}(\xi_\ell, \xi_m) + \sum_j v(\xi_j)$$

$$\therefore\ (k-1) \sum_j v(\xi_j) \geq \sum_{\ell \neq m} \sum \mathrm{cov}(\xi_\ell, \xi_m) = 2 \sum_{\ell < m} \sum \mathrm{cov}(\xi_\ell, \xi_m)$$

ou $$\sum_j v(\xi_j) \geq \frac{2}{k-1} \sum_{\ell < m} \sum \mathrm{cov}(\xi_\ell, \xi_m)$$

Então, (39) toma a forma

$$\rho^2_{x\xi} \geq \frac{\frac{2}{k-1} \sum_{\ell < m} \sum \mathrm{cov}(Y_\ell, Y_m) + 2 \sum_{\ell < m} \sum \mathrm{cov}(Y_\ell, Y_m)}{v(X)} =$$

$$= \frac{\frac{2k}{k-1} \sum_{\ell < m} \sum \mathrm{cov}(Y_\ell, Y_m)}{v(X)}$$

mas, $$v(X) = v\left(\sum_j Y_j\right) = \sum_j v(Y_j) + 2 \sum_{\ell < m} \sum \mathrm{cov}\,(Y_\ell, Y_m)$$

ou seja, $$2 \sum_{\ell < m} \sum \mathrm{cov}\,(Y_\ell, Y_m) = v(X) - \sum_j v(Y_j)$$

vindo finalmente

$$\rho^2_{x\xi} \geq \frac{k}{k-1} \cdot \frac{v(X) - \sum_j v(Y_j)}{v(X)} = \frac{k}{k-1}\left[1 - \frac{\sum_j v(Y_j)}{v(X)}\right] \text{ c.q.d.}$$

O valor $$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum_j v(Y_j)}{v(X)}\right] \qquad (32)$$

toma o nome de coeficiente α de Cronbach[42].

O coeficiente α, quantidade calculável a partir de uma única aplicação de um teste, dá um minorante para a fiabilidade do teste.

A expressão (31), $\alpha = R = \rho_{xx'} = \rho^2_{x\xi} = \dfrac{k\rho_{yy'}}{1 + (k-1)\rho_{yy'}}$, é conhecida por fórmula de Spearman-Brown para a fiabilidade de um teste composto com k componentes paralelas.

Quando os Y_j são items dicotómicos, o coeficiente α reduz-se à fórmula 20 de Kuder-Richardson (KR20):

$$\alpha_{20} = \frac{k}{k-1}\left(1 - \frac{\sum_j p_j q_j}{v(X)}\right) \qquad (33)$$

Sendo os p_j aproximadamente iguais, pode usar-se a fórmula 21 de Kuder-Richardson (KR21):

$$\alpha_{21} = \frac{k}{k-1}\left(1 - \frac{n\,\bar{p}\,\bar{q}}{v(X)}\right) \qquad (34)$$

com $\bar{p} = \frac{1}{k}\sum_j p_j$, $\bar{q} = 1-\bar{p}$. Note-se que $\alpha_{20} \geqslant \alpha_{21}$ com igualdade se e só se $p_1 = p_2 = \ldots = p_k$.

Métodos de estimação da fiabilidade

Estimar a fiabilidade de um teste é procurar estimar o grau de existência da realidade psicológica definida pelo teste[43].

A avaliação da fiabilidade é usualmente feita a partir de três métodos principais, considerados como processos de avaliação da precisão métrica do teste, suposto este um instrumento de medida aplicável a uma realidade objectiva e independente do teste. Por outras palavras, a fiabilidade é a qualidade que se traduz em que uma mesma prova aplicada duas vezes a um mesmo indivíduo dê resultados idênticos (ou praticamente idênticos).

---

(42) L. Cronbach, "Coefficient alpha and the internal structure of tests", Psychometrika, 1951, 16, 297-334.

(43) Renchlin, 1962, ob. cit. na bib. p.82.

a) Método teste-reteste — um mesmo instrumento de medida é aplicado duas vezes a uma mesma amostra, correlacionando-se os dois conjuntos de valores obtidos. Ao coeficiente de correlação obtido dá-se o nome de coeficiente de constância.

b) Método dos testes (ou formas) paralelas — formas paralelas de um teste são aplicadas a uma mesma amostra. O coeficiente de correlação das pontuações associadas às duas formas toma o nome de coeficiente de equivalência.

c) Método da partição ou das metades (split-half method) — divide-se o teste em duas partes equivalentes, calculando-se a correlação entre as pontuações obtidas nas duas metades, sobre uma mesma amostra.
Em muitos casos, a partição é feita por items pares e ímpares — método par-ímpar (odd-even method).
O coeficiente de fiabilidade toma neste caso o nome de coeficiente de homogeneidade.

Note-se que c) é um caso particular de b): um teste com 2k items dará origem, por partição, a duas formas paralelas com k items cada.

A correlação das duas metades (ou formas paralelas) dá como resultado um coeficiente de fiabilidade para meio teste, $R^*_{1/2}$.

A fiabilidade para a totalidade do teste pode ser estimada pela fórmula de Spearman-Brown (cf. com (18)):

$$R^* = \frac{2\,R^*_{1/2}}{1 + R^*_{1/2}} = \frac{2\,r_{yy'}}{1 + r_{yy'}} \qquad (35)$$

sendo $r_{yy'}$ o coeficiente de correlação da amostra, estimativa do coeficiente de correlação teórico $\rho_{yy'}$.

Um método para obter coeficientes de fiabilidade usando estatísticas teste-item foi desenvolvido por Kuder e Richardson(44).

(44) G. Kuder, M. Richardson, "The theory of the estimation of test reliability, Psychometrika, 1937, 2, 151-160.

Para testes com k items dicotômicos uma estimativa da fiabilidade é obtida a partir da fórmula 20 de Kuder-Richardson (KR20).

$$R^* = \alpha^*_{20} = \frac{k}{k-1}\left(1 - \frac{\sum_j p^*_j q^*_j}{s^2_x}\right) \qquad (36)$$

sendo p^*_j a proporção de indivíduos da amostra com sucesso no item j, $q^*_j = 1 - p^*_j$ e s^2_x a variância empírica das pontuações do teste; α^*_{20} é uma estimativa de α_{20} dado por (33).

Com $p^*_j \simeq p^* = \text{const.}$ obtém-se, de acordo com KR21 dada por (34),

$$R^* = \alpha^*_{21} = \frac{k}{k-1}\left(1 - \frac{k\, p^* q^*}{s^2_x}\right) \qquad (37)$$

Para items não dicotómicos de (32) obtém-se

$$R^* = \alpha^* = \frac{k}{k-1}\left(1 - \frac{\sum_j s^2_j}{s^2_x}\right) \qquad (38)$$

sendo s^2_j a variância empírica do item j e α^* uma estimativa do coeficiente α de Cronbach.

O coeficiente α possui a seguinte propriedade importante:

— retome-se a população teórica U e seja X um teste com 2k items; no método das metades, este teste pode ser subdividido de $\frac{1}{2}\binom{2k}{k} = \frac{1}{2}\frac{(2k)!}{(k!)^2}$ maneiras, obtendo-se $\frac{1}{2}\binom{2k}{k}$ testes compostos (com 2 componentes)

$$X_i = Y_i + Y'_i \quad , \quad i = 1, 2, \ldots, \frac{1}{2}\binom{2k}{k}$$

Para cada teste composto X_i calcule-se, a partir de (32) com k=2,

$$\alpha_i = 2\left[1 - \frac{v(Y_i) + v(Y'_i)}{v(X_i)}\right]$$

Seja α_x o coeficiente α calculado para o teste original com 2k items. É válido então o

Teorema. $E^*(\alpha_i) = \alpha_x$, sendo E^* o operador valor médio sobre a população dos $\frac{1}{2}\binom{2k}{k}$ testes compostos.

Este resultado pode ser intepretado de duas maneiras:

1) calculando os $\frac{1}{2}\binom{2k}{k}$ valores α_i, a sua média aritmética é igual a α_x e
2) distribuindo os items aleatoriamente pelas duas metades, o valor médio de α_i é α_x (45)

As teorias recentes relativas à fiabilidade dos testes inserem este problema num quadro bastante geral. Procura-se estimar não apenas o grau de invariabilidade dum instrumento, ou o grau de invariabilidade de um objecto, mas, fundamentalmente, estimar o grau de repetibilidade duma relação estabelecida quando se usa um instrumento de medida e o grau de existência do objecto definido pelo instrumento (*).

(45) V. Lord and Novick, ob. cit. na bib., cap.4.
(*) Cf. cap. 1

Apêndice III

Vectores. Espaços vectoriais

Um conjunto ordenado de n números $\vec{a} = (a_1, a_2, \ldots, a_n)$ diz-se um vector a n dimensões (ou dum espaço de dimensão n). Os números $a_1, \ldots, a_n$ são as componentes ou coordenadas de $\vec{a}$.

Dois vectores $\vec{a} = (a_1, \ldots, a_n)$ e $\vec{b} = (b_1, \ldots, b_n)$ são iguais se e só se $a_i = b_i$, $i = 1, \ldots, n$.

Soma de vectores: $\vec{a} + \vec{b} = (a_1, \ldots, a_n) + (b_1, \ldots, b_n) =$

$$= (a_1 + b_1, \ldots, a_n + b_n).$$

A soma de vectores é comutativa e associativa. O elemento neutro da adição é o vector nulo $\vec{o} = (o, \ldots, o)$.

Produto do número λ (escalar) pelo vector $\vec{a}$: $\lambda \vec{a} = (\lambda a_1, \ldots, \lambda a_n)$

Ao produto por λ também se dá o nome de homotetia de razão λ, se $\lambda > 0$ (dilatação se $\lambda > 1$, contracção se $0 < \lambda < 1$) ou homotetia de razão $|\lambda|$ (com inversão de sentido) se $\lambda < 0$.

Das definições dadas resultam as seguintes propriedades:

1) $\vec{a} + \vec{b} = \vec{b} + \vec{a}$

2) $(\vec{a} + \vec{b}) + \vec{c} = \vec{a} + (\vec{b} + \vec{c})$

3) $\vec{a} + \vec{o} = \vec{a}$

4) $\vec{a} + \vec{a}' = \vec{o}$ com $\vec{a}' = (-1)\vec{a} = -\vec{a}$

5) $1.\vec{a} = \vec{a}$

6) $\lambda (\mu \vec{a}) = (\lambda \mu) \vec{a}$

7) $\lambda (\vec{a} + \vec{b}) = \lambda \vec{a} + \lambda \vec{b}$ (μ, λ - escalares)

8) $(\lambda + \mu) \vec{a} = \lambda \vec{a} + \mu \vec{a}$

Ao conjunto de todos os vectores a n dimensões, de componentes

reais, munido das operações de adição e multiplicação por um escalar, satisfazendo às propriedades 1-8, dá-se o nome de espaço vectorial (ou afim) a n dimensões, e representa-se por V^n.

Dependência e Independência Linear

Dados dois vectores de V^n, $\vec{a}$ e $\vec{b}$, diz-se que $\vec{b}$ é proporcional a $\vec{a}$ se existir um escalar λ (isto é um número real) tal que $\vec{b} = \lambda \vec{a}$. Também se diz que $\vec{a}$ e $\vec{b}$ são colineares.

O vector $\vec{o}$ é proporcional a todo o vector $\vec{a}$.

Se $\vec{b}$ é proporcional a $\vec{a}$ então $\vec{a}$ é proporcional a $\vec{b}$ (isto é, a relação de proporcionalidade é simétrica).

Um vector $\vec{b}$ diz-se combinação linear dos vectores $\vec{a}_1, \ldots, \vec{a}_s$ se existem escalares (números reais) $\lambda_1, \ldots \lambda_s$ tais que

$$\vec{b} = \lambda_1 \vec{a}_1 + \lambda_2 \vec{a}_2 + \ldots + \lambda_s \vec{a}_s$$

r vectores ($r \geq 2$) $\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_r$ dizem-se linearmente dependentes se pelo menos um destes vectores é combinação linear dos restantes. Caso contrário os vectores $\vec{a}_1, \ldots, \vec{a}_r$ dizem-se linearmente independentes.

Um conjunto de vectores linearmente independentes (dependentes) também se diz uma família livre (não livre).

Por outras palavras, os vectores $\vec{a}_1, \ldots, \vec{a}_r$ são linearmente independentes se uma relação do tipo

$$\lambda_1 \vec{a}_1 + \ldots + \lambda_r \vec{a}_r = 0$$

implica $\lambda_1 = \ldots = \lambda_r = 0$. Com algum(s) λ não nulo(s) os vectores são linearmente dependentes.

Os seguintes resultados são fáceis de estabelecer:

1. O conjunto constituído pelo único vector $\vec{a}$ é linearmente dependente se e só se $\vec{a} = \vec{o}$.

2. Se um subconjunto dum conjunto de vectores é linearmente dependente então o conjunto também é linearmente dependente (equivalentemente: se um conjunto é linearmente independente todo o subconjunto também o é).

3. Todo o conjunto de dois vectores iguais ou, em geral, de dois vectores proporcionais (colineares), assim como todo o conjunto con

tendo o vector nulo são famílias não livres.

Teorema. Num espaço vectorial a n dimensões é sempre possível encontrar n vectores linearmente independentes.

Por exemplo, em V^n

$$\vec{e}_1 = (1,0,\ldots,0)$$
$$\vec{e}_2 = (0,1,0,\ldots 0)$$
$$\cdots\cdots\cdots\cdots$$
$$\vec{e}_n = (0.0,\ldots,0,1)$$

Teorema. Toda a família de s vectores de V^n é não livre se $s>n$.

Por outras palavras, n é o número maximal de vectores linearmente independentes.

Definição. Toda a família $B = \left\{\vec{b}_1, \vec{b}_2,\ldots,\vec{b}_n\right\}$ de n vectores linearmente independentes de V^n se diz uma base de V^n.

Teorema. Todo o vector $\vec{a}$ de V^n se pode representar como combinação linear dos vectores de B, sendo esta representação única.

Seja $\vec{a} = \lambda_1 \vec{b}_1 + \lambda_2 \vec{b}_2 + \ldots + \lambda_n \vec{b}_n$. Os coeficientes λ_i são as coordenadas ou componentes de $\vec{a}$ na base B; $(\lambda_1,\ldots,\lambda_n)$ é a representação de $\vec{a}$ na base B e escreve-se $(\vec{a})_B = (\lambda_1,\ldots,\lambda_n)$.

Em cada base $\vec{a}$ é, então, representado pela sucessão das suas coordenadas, relativas a essa base.

Como vimos, dar um vector $\vec{v}$ de V^n é dar um conjunto ordenado de n números: $\vec{v} = (v_1,v_2,\ldots,v_n)$. É de verificação imediata que $v_1,\ldots,v_n$ são as coordenadas de $\vec{v}$ na base $\left\{\vec{e}_1,\vec{e}_2,\ldots,\vec{e}_n\right\}$ que toma o nome de base canónica ou natural.

Então, dar um vector $\vec{v} = (v_1,\ldots v_n)$ equivale a dar a sua representação na base canónica.

Sendo $\vec{v}$ dado pela sua representação numa base não canónica, esta deve ser explicitada.

Por exemplo, em V^2:

$$(\vec{v})_B = (2, 3) \quad \text{com} \quad B = \left\{\vec{b}_1 = (0,1),\ \vec{b}_2 = (2, 2)\right\}$$

A representação canónica de $\vec{v}$ obtem-se imediatamente:

$$\vec{v} = 2\,\vec{b}_1 + 3\vec{b}_2 = 2(0,1) + 3(2,2) = (0,2) + (6,6) = (6,8) = 6(1,0) + 8(0,1) = 6\,\vec{e}_1 + 8\,\vec{e}_2$$

Tem lugar o resultado:

Se na base $\{\vec{b}_1,\ldots,\vec{b}_n\}$ o vector $\vec{v} = (v_1,\ldots,v_n)$ tem a decomposição $\vec{v} = v_1\,\vec{b}_1 + \ldots + v_n\,\vec{b}_n$, então $\{\vec{b}_1,\ldots,\vec{b}_n\}$ é base canónica $\{\vec{e}_1,\ldots,\vec{e}_n\}$

Definição. Um subconjunto w dum espaço vectorial V diz-se um subespaço vectorial se w e um espaço vectorial relativamente às operações de adição de vectores e multiplicação por um escalar, definidas em V.

Definição. Seja V um espaço vectorial e $\Delta = \{M, N, O, P, Q, \ldots\}$ um conjunto cujos elementos tomam o nome de pontos.

Suponha-se que a cada par ordenado de pontos de Δ corresponde um e um só vector de V, de acordo com as propriedades.

1. Dois pontos P e Q definem um vector $Q-P = \overrightarrow{PQ} = \vec{v} \in V$

2. Dado um ponto P e um vector $\vec{v}$ existe um e um só ponto Q tal que $\overrightarrow{PQ} = \vec{v}$ (também se diz que a soma do ponto P com o vector $\vec{v}$ é o ponto Q)

3. Três pontos O,P,Q satisfazem à relação

$$\overrightarrow{OP} + \overrightarrow{PQ} = \overrightarrow{OQ}$$

Ao conjunto R de todos os pontos de Δ e todos os vectores de V dá-se o nome de espaço vectorial pontual ou espaço linear. A dimensão do espaço linear é a dimensão do espaço vectorial associado.

Seja R^n um espaço linear n-dimensional. Referencial de R^n é um sistema $(O, \vec{u}_1, \vec{u}_2,\ldots,\vec{u}_n)$ formado pelo ponto O de R^n, escolhido arbitrariamente, e por n vectores linearmente independentes de V^n (espaço vectorial associado a R^n)

Dado um ponto P de R^n tem-se

$$\overrightarrow{OP} = P - O = x_1\vec{u}_1 + x_2\vec{u}_2 + \ldots + x_n\vec{u}_n$$

os x_i dizem-se as coordenadas do ponto P no referencial $(O, \vec{u}_1,\ldots,\vec{u}_n)$.

Em particular o ponto O tem por coordenadas $(0,0,\ldots,0)$

Consequências imediatas das definições são:

1. Se $\overrightarrow{MN} = \overrightarrow{OP}$ então $\overrightarrow{MO} = \overrightarrow{NP}$

2. Dados dois pontos $A(a_1,\ldots,a_n)$ e $B(b_1,\ldots,b_n)$ de R^n as coordenadas do vector $\overrightarrow{AB} = B-A$ são $(b_1-a_1,b_2-a_2,\ldots,b_n-a_n)$.

Seja R_n um espaço linear. O conjunto de pontos de R^n satisfazendo a uma equação da forma $\phi(x_1,x_2,\ldots,x_n)=0$ diz-se uma hipersuperfície. Sendo ϕ linear, i. é, da forma $a_1x_1 +\ldots+a_nx_n - b = 0$ a hipersuperfície toma o nome de hiperplano.

Um hiperplano é um subespaço de R^n de dimensão n-1.

Ao conjunto de pontos satisfazendo ao sistema de r equações independentes (v.nota)

$$\begin{cases} a_{11}x_1 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + \ldots + a_{2n}x_n = b_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{r1}x_1 + \ldots + a_{rn}x_n = b_r \end{cases}$$

dá-se o nome de plano a k = n-r dimensões

Como cada uma das r equações do sistema representa um hiperplano segue-se que todo o plano k-dimensional é a intersecção de r=n-k hiperplanos linearmente independentes.

A equação do hiperplano passando pelo ponto P de coordenadas $(t_1,\ldots,t_n)$ ainda se pode escrever na forma

$$a_1(x_1-t_1)+a_2(x_2-t_2)+\ldots+a_n(x_n-t_n)=0$$

(Nota. r funções $\varphi_1,\ldots,\varphi_r$ de uma ou mais variáveis, são linearmente dependentes se existe uma relação $\lambda_1\varphi_1 +\ldots+\lambda_r\varphi_r = 0$ com algum λ não nulo, para todos os valores das variáveis. Caso con

trário as funções são independentes.

Várias relações do tipo $\lambda_1\varphi_1 + \ldots + \lambda_r\varphi_r = 0$ dizem-se independentes se os vectores $\vec{\lambda} = (\lambda_1, \ldots, \lambda_r)$ são linearmente independentes).

Espaços Euclideanos

Num espaço vectorial introduza-se uma métrica, i.é, um processo permitindo medir comprimentos e ângulos.

A métrica será definida pela introdução dum produto escalar ou interno.

Definição: Num espaço vectorial V define-se um produto escalar fazendo corresponder a todo o par de vectores $\vec{u}$, $\vec{v} \in V$ um número real $(\vec{u},\vec{v})$ de acordo com as propriedades:

1. $(\vec{u},\vec{v}) = (\vec{v},\vec{u})$

2. $(\alpha\vec{u},\vec{v}) = \alpha\,(\vec{u},\vec{v})$

3. $(\vec{u}+\vec{v},\vec{w}) = (\vec{u},\vec{w}) + (\vec{v},\vec{w})$

Em geral designaremos o produto escalar de $\vec{u}$ e $\vec{v}$ por $\vec{u}.\vec{v}$

Um espaço vectorial V munido dum produto escalar diz-se um espaço vectorial euclideano , V_E

De 1.,2. e 3. resulta imediatamente

1'. $(\vec{u}, \beta\vec{v}) = \beta\,(\vec{u},\vec{v})$

2'. $(\vec{u},\vec{v} + \vec{w}) = (\vec{u},\vec{v}) + (\vec{u},\vec{w})$

Definição: Dado $\vec{v} \in V_E$ define-se norma (módulo ou comprimento) de $\vec{v}$ por

$$\|\vec{v}\| = \sqrt{\vec{v}.\vec{v}}$$

Dados os vectores $\vec{a}$, $\vec{b} \in V_E$ tem-se

$$\|\vec{a} + \vec{b}\|^2 = \|\vec{a}\|^2 + \|\vec{b}\|^2 + 2\vec{a}.\vec{b}$$

Define-se ângulo θ dos vectores $\vec{u}$ e $\vec{v}$ pela igualdade

$$\cos\theta = \frac{\vec{u}.\vec{v}}{\|\vec{u}\|\|\vec{v}\|}$$

É válida a igualdade

$$\cos(\widehat{\vec{u},\vec{v}}) = \cos(\widehat{\lambda\,\vec{u},\mu\,\vec{v}})$$

Dois vectores $\vec{u}$ e $\vec{v}$ dizem-se ortogonais se $\vec{u}.\vec{v} = 0$

Uma base $\{\vec{e}_1,\ldots,\vec{e}_n\}$ dum espaço euclideano é ortogonal se $\vec{e}_i.\vec{e}_k = 0$ para $i \neq k$, $i,k=1,\ldots,n$. Se ainda for $\|\vec{e}_i\| = 1 \ \forall_i$ a base diz-se ortonormada.

Aos vectores de módulo 1 dá-se o nome de vectores unitários.

Vectores não nulos ortogonais dois a dois são linearmente independentes.

Dado um vector $\vec{v}$ e um vector $\vec{e}$ ligados pela relação $\vec{v} = \|\vec{v}\|\vec{e}$ tem-se que $\|\vec{e}\| = 1$. A $\vec{e}$ dá-se o nome de versor ou vector unitário de $\vec{v}$

Expressão do produto escalar em função das coordenadas

Seja $\{\vec{e}_1,\ldots,\vec{e}_n\}$ uma base qualquer de V_E^n . Dados dois vectores $\vec{a}$ e $\vec{b}$ tem-se $\vec{a} = a_1\vec{e}_1 + \ldots + a_n\vec{e}_n$, $\vec{b} = b_1\vec{e}_1 + \ldots + b_n\vec{e}_n$

e $(\vec{a},\vec{b}) = \vec{a}.\vec{b} = \sum_i \sum_k a_i b_k \, \vec{e}_i.\vec{e}_k = \sum_i \sum_k g_{ik}\, a_i b_k$

com $g_{ik} = (\vec{e}_i,\vec{e}_k) = \vec{e}_i.\vec{e}_k$

Sendo a base $\{\vec{e}_1,\ldots,\vec{e}_n\}$ ortogonal tem-se $g_{ik} = 0$ para $i \neq k$ e

$$\vec{a}.\vec{b} = \sum_i g_{ii}\, a_i b_i$$

Sendo a base ortonormada ainda se tem $g_{ii} = 1$ e

$$\vec{a}.\vec{b} = \sum_i a_i b_i$$

Por exemplo, sejam em V^3 os vectores $\vec{a}=(1,-1,1)$ e $\vec{b}=(2,4,2)$.

Considerando os vectores representados na base canónica é

$$\vec{a}.\vec{b} = 1x2 - 1x4 + 1x2 = 2-4+2 = 0$$

Seja R^n um espaço linear e V^n o espaço vectorial correspondente munido de uma métrica euclideana (isto é, dum produto escalar satisfazendo aos postulados indicados)

A introdução da métrica em R^n permite definir a distância d(P,Q) entre dois pontos, P e Q, a partir da igualdade

$$d(P,Q) = \|\overrightarrow{PQ}\|$$

Ao espaço R^n munido da métrica euclideana dá-se simplesmente o nome de espaço euclideano e representa-se por E^n.

Da definição de distância resulta que

1. $d(P,Q) \geq 0$
2. $d(P,Q) = d(Q,P)$
3. $d(P,Q) \leq d(P,O) + d(O,Q)$

(Em que condições é válida a igualdade?)

——— /// ———

Referências bibliográficas

A.C.Aitken (1956), "Determinants and matrices", Oliver and Boyd.(C3)

H.R.Alker (1969), "Introduction à la sociologie mathématique", Larousse Université. (C2)

F.Ayres,Jr. (1962), "Matrices", Schaum's Outline Series, McGraw-Hill (C3)

G.Bachelard (1953), "Le matérialisme rationnel",P.U.F. (C1)

— (1972), "L'engagement rationaliste", P.U.F. (C1)

— (1974), "Épistémologie", P.U.F. (textos escolhidos por D. Lecourt) (C1)

R.Blanché (1970), "L'axiomatique", P.U.F. (C1)

— (1975), "A epistemologia", Edit. Presença (C1)

R.Boudon (1970), "Tendances principales de la recherche dans les sciences sociales et humaine. Partie 1: sciences sociales - Modèles et méthodes mathématiques", Mouton-Unesco (Edição portuguesa, Bertrand, 1973) (C1)

W.Bridgman (1927), "The logic of modern physics", McMillan. (C1)

N.R.Campbell (1920), "Physics: the elements", Cambridge Univ. Press. (Publicado pela Dover em 1957 sob o título "Foundations of science: the philosophy of theory and experiment"). (C1)

M.Castells, C.Dan e outros (1976), "Epistemologia e ciências sociais", Rés. (C1)

M.Caveing, Juliot-Curie e outros (1976), "Problemática da Ciência", Rés. (C1)

W.J.Conover (1971), "Pratical nonparametric statistics",John Wiley. (C1, C2)

H.Coombs, M.Dawes, A.Tversky (1975), "Psychologie mathématique",P.U.F. (2 volumes) (C1)

H.Cramér (1946), "Mathematical methods of statistics", Princeton Univ. Press. (C2)

F.B. Davis (1966), "Analyse des items", Edit. Nauwelaerts, Louvain. (A II)

J.Faverge (1971 - 6ª ed. actualizada), "Méthodes statistiques en psychologie appliquée", P.U.F. (3 volumes)(C1, C2, C3, A II)

G.A. Ferguson (1971), "Statistical analysis in psychology & education", McGraw-Hill. (C2, C3, A II)

B. Fruchter (1954), "Introduction to factor analysis", Van Nostrand. (C3)

F.Galvão de Melo (1973), "Introdução aos métodos estatísticos", vol.II Liv. Escolar Editora. (C2)

G.Glass, J.Stanley (1970), "Statistical methods in education and psychology", Prentice-Hall. (C2)

L.I. Golovina (1974), "Algebra lineal y algunas de sus aplicaciones", Edit. Mir. (A III)

G.G. Granger (1975), "Pensamento formal e ciências do homem", Edit.Presença. (2 volumes) (C1)

Pierre Gréco (1976), "Epistemologia da psicologia", Edit. Nova Crítica. (C1)

P.R. Halmos (1958), "Finite-dimensional vector spaces", Van Nostrand. (A III)

H.H. Harman (1967), "Modern factor analysis", Univ. of Chicago Press. (C3)

H.L.Hays (1972), "Statistics"London, Holt,Rinehart and Winston.(C2)

K.J. Holzinger, H.H.Harman (1941), "Factor analysis. A synthesis of factorial methods", Univ. of Chicago Press. (C3)

M.Kendall (1971), "Rank Correlation methods", Charles Griffin. (C2)

— (1975), "Multivariate Analysis", Charles Griffin. (C3)

A. Kurosh (1973), "Cours d'algèbre supérieure", Edit. Mir (C3, A III)

S. Lipschutz (1968), "Linear algebra", Schaum's outline series, MacGraw-Hill. (C3, A III)

F.M. Lord, M.R. Novick (1968), "Statistical theories of mental test scores", Addison-Wesley. (C2, A II)

Q. McNemar (1969), "Psychological statistics", John Wiley. (C2)

N. Mouloud (1968), "Les structures, la recherche et le savoir", Payot. (C1)

J.C. Nunnally (1959), "Tests and measurements", McGraw-Hill.(C1, C2, A II)

— (1970), "Introduction to psychological measurement", McGraw-Hill. (C1, C2, A II)

J. Piaget (1970), "Psychologie et epistémologie", Gonthier. (Edição portuguesa, Pub. D.Quixote, 1972) (C1)

— (1970), "Tendances principales de la recherche dans les sciences sociales et humaines. Partie 1: sciences sociales - La situation des sciences de l'homme dans le système des sciences" Mouton-Unesco. (C1)

— (1970), "La psychologie", Mouton-Unesco. (Edições em português, Bertrand, 1972) (C1, C3)

P. Pichot (1968), "Les tests mentaux", P.U.F. (Edição portuguesa, Pub. D.Quixote, 1969) (C3)

Max Planck (1949), "L'image du monde dans la physique moderne", Gonthier. (C1)

M. Reuchlin (1962), "Les méthodes quantitatives en psychologie", P.U.F. (C1, C2, C3, A II)

— (1964), "Méthodes d'analyse factorielle", P.U.F. (C3)

— (1972), "La psychologie différentielle" P.U.F. (Ediç. portuguesa, Estúdios Cor, 1972) (C3)

— (1976), "Précis de statistique", P.U.F. (C1, C2)

J. Sebastião e Silva (1959), "Introdução à lógica simbólica e aos fundamentos da matemática", separata da Rev.Palestra nº6 (C2)

— (1975), "Compêndio de matemática", Edição G.E.P.1ºvol. (C2) ; 2º Volume (C3, A III)

S. Siegel (1956), "Nonparametric statistics for the behavioral sciences", McGraw-Hill (C1, C2)

C. Spearman (1927), "The abilities of man", McMillan. (C3)

S.S. Stevens (1946), "On the theory of scales of measurement", Science, 103, 677-680. (C1)

— (1951), "Handbook of experimental psychology", John Wiley. (C1)

R.R. Stoll (1961), "Sets, logic and axiomatic theories", Freeman. (C1, C2)

G. Thompson (1951), "The factorial analysis of human ability", Univ. London Press. (C3)

— (1954), "The geometry of mental measurement", Univ. London Press. (C3)

L.L. Thurstone (1935), "The vectors of mind", Univ. of Chicago Press. (C3)

— (1938), "Primary mental abilities", Univ. of Chicago Press. (C3)

— (1947), "Multiple factor analysis", Univ. of Chicago Press. (C3)

J. Torrens-Ibern (1972), "Modèles et méthodes de l'analyse factorielle", Dunod. (C3)

J. Ullmo (1958), "La pensée scientifique moderne", Flammarion. (Edição portuguesa, Coimbra Editora, 1967) (C1)

P.H. Vernon (1950), "The structure of human abilities", John Wiley. (C3)

Executado nas Oficinas Gráficas
da Editorial do M. E. U. – Algueirão

Obra de referência útil, quer para o matemático aplicado quer para o psicólogo ou sociólogo que nos seus trabalhos tenham que recorrer aos métodos quantitativos de análise de dados.

O texto está dividido em três partes. Numa primeira, procura-se fundamentar o uso dos modelos matemáticos nas ciências do homem, nomeadamente na psicologia, fazendo-se o estudo das escalas (níveis) de medida e suas condições de aplicabilidade.

Na segunda e terceira partes expõem-se algumas técnicas matemático-estatísticas de grande importância nas aplicações.

Descrevem-se as medidas de associação (correlação) entre duas variáveis, percorrendo a gama dos métodos oferecidos pela estatística e adaptados aos diferentes tipos de escalas, cobrindo a maior parte das situações postas pela prática. Finalmente, desenvolvem-se alguns aspectos da Análise Factorial, tratando e discutindo as hipóteses de Spearman, Holzinger e Thurstone e descrevendo, com algum pormenor, as soluções de Hotteling e Thurstone.

obras avulsas

1. Oa

ESCOLA NACIONAL DE SAÚDE PÚBLICA

EDIÇÕES DE SAÚDE