Investigações em Ensino de Ciências – V4(2), pp. 161-180, 1999

UM EXEMPLO DE ANÁLISE MULTIVARIADA APLICADA À PESQUISA QUANTITATIVA EM ENSINO DE CIÊNCIAS: EXPLICANDO O DESEMPENHO DOS CANDIDATOS AO CONCURSO VESTIBULAR DE 1999 DA UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

(An example of multivariate analysis applied to quantitative research in science teaching: explaining the performance of the 1999 entrance exam candidates to the Federal University of Rio Grande do Sul, Brazil)

Fernando Lang da Silveira [lang@if.ufrgs.br]
Instituto de Física da UFRGS
Caixa Postal 15051
91501-970 Porto Alegre, RS, Brasil

Resumo

O objetivo desse trabalho é o de apresentar algumas técnicas de análise quantitativa, potencialmente úteis na abordagem de problemas de pesquisa em ensino de ciências com muitas variáveis, destacando os conceitos e os significados das mesmas. As técnicas apresentadas (Análise de Consistência Interna e Análise da Variância) são exemplificadas através do estudo das relações que quinze variáveis sócio-econômico-culturais tiveram com o desempenho em nove provas respondidas por 35463 candidatos ao Concurso Vestibular de 1999 da Universidade Federal do Rio Grande do Sul. O estudo mostrou que as quinze variáveis conjuntamente explicaram 34,2% da variância do desempenho dos candidatos, sendo 19,0% a explicação das variáveis de escolaridade independentemente das variáveis sócio-econômicas.

Palavras-chave: análise multivariada; pesquisa em ensino de ciências, vestibular.

Abstract

This paper aims at presenting some quantitative analysis techniques that can be potentially useful in approaching research problems in science teaching with many variables, emphasizing their concepts and meanings. The presented techniques (Internal Consistency Analysis and Variance Analysis) are exemplified through the study of the role fifteen social, economic, and cultural variables had on the performance in nine tests that were answered by 35463 candidates of the 1999 Entrance Exam to the Federal University of Rio Grande do Sul. The study showed that the fifteen variables together could explain 34.2% of the variance in the performance of the candidates, being 19% represented by schooling variables that were independent of the socio-economic ones.

Keywords: Multivariate Analysis; research on science education; university entrance exam .

1. Introdução

Na pesquisa quantitativa em ensino de ciências, freqüentemente nos interessa algum fenômeno onde diversas variáveis estão envolvidas, sendo necessário conhecermos as relações entre elas. Remontam às primeiras décadas do século XX o início do desenvolvimento dos procedimentos analíticos para tratar dessas complexas situações multivariadas. Atualmente muitas dessas técnicas estão disponíveis em programas computacionais[1]; outrora, a justificativa para a não aplicação de tais tratamentos era a grande quantidade de cálculos necessários. Todavia esta não pode ser mais a desculpa; com auxílio dos computadores pessoais é possível realizar facilmente a tarefa. Ainda

[1] ? Todas as técnicas de análise estatística apresentadas neste trabalho foram viabilizadas através do programa *"SPSS for Windows ? Release 8.0"*.

assim existem barreiras para a utilização desses métodos; uma delas é o seu desconhecimento pelos pesquisadores interessados na pesquisa quantitativa.

O objetivo principal deste trabalho é apresentar algumas dessas técnicas de análise quantitativa, procurando destacar os conceitos envolvidos. Exemplificaremos alguns procedimentos através de um estudo que visou elucidar as relações que quinze variáveis sócio-econômico-culturais apresentaram com nove variáveis de desempenho, entre 35463 candidatos ao Concurso Vestibular de 1999 da Universidade Federal do Rio Grande do Sul (CV?99/UFRGS). Julgamos ser extremamente importante a elucidação das relações entre esse dois grupos de variáveis, já que opiniões sobre como fatores sócio-econômico-culturais explicam o desempenho nos concursos vestibulares existem; o que falta em nossa realidade (muito possivelmente em outras também) são os estudos concretos. Este estudo, entretanto, tem o objetivo central de exemplificar uma situação multivariada.

As referências bibliográficas apresentadas são, intencionalmente, por vezes redundantes. Queremos, dessa forma, dar indicações alternativas ao leitor interessado em aprofundar o assunto.

2. Um problema de pesquisa multivariado

O problema central que nos motivou a realizar a pesquisa pode assim ser enunciado: Quais são as relações do desempenho dos candidatos no CV?99/UFRGS com variáveis sócio-econômico-culturais?

Esta é uma questão que virtualmente envolve muitas variáveis; os conteúdos de segundo grau constantes no programa do CV?99/UFRGS foram avaliados em nove provas: Língua Portuguesa, Língua Estrangeira, Literatura, História, Geografia, Biologia, Matemática, Física e Química. Cada prova, exceto uma, teve 30 itens de escolha múltipla e resposta única; a prova de Língua Portuguesa incluiu também uma questão de Redação[2]. Ou seja, tínhamos para cada candidato nove escores de desempenho nas provas; estes escores eram variáveis com valor mínimo nulo e máximo trinta, indicando o número de questões respondidas corretamente em cada prova por cada um dos 35463 candidatos.

Adicionalmente, trabalhamos com as respostas emitidas pelos candidatos ao Questionário de Informações sobre o Candidato, respondido no momento da inscrição ao CV?99/UFRGS. Deste questionário, com um total de 21 questões, 15 delas forneceram dados sócio-econômico-culturais (adiante explicitaremos quais foram esse dados). Ou seja, mais 15 variáveis, potencialmente explicativas do desempenho, constaram desta pesquisa. Portanto o nosso estudo caracterizou-se como multivariado, envolvendo 24 variáveis.

3. Quantificando a relação entre duas variáveis: o coeficiente de correlação

Um coeficiente de correlação é uma medida padronizada do grau de associação (variação concomitante) entre duas variáveis. O conhecido coeficiente de correlação de Pearson (Afifi e Clark, 1996; Cronbach, 1996; Ferguson, 1976, Guilford e Fruchter, 1973; Wherry, 1984), cuja fórmula pode ser encontrada em qualquer texto de estatística elementar, quantifica em uma escala adimensional, que em valor absoluto vai de zero à unidade, o grau de interrelacionamento entre duas variáveis (quanto maior o módulo do coeficiente, mais intensa é a associação linear entre as

[2] ? Não utilizamos nesta pesquisa os escores da questão de Redação pois cerca da metade dos vestibulandos tiveram sua Redação avaliada; a outra metade participou do concurso mas foi eliminada por um critério que visava reduzir o número de redações a corrigir. Esta redução já poderia estar operando como um filtro sócio-econômico-cultural; a fim de não arriscar uma perda em variabilidade em tais fatores, decidimos conduzir o estudo apenas com os resultados advindos das nove prova com itens de escolha múltipla.

duas variáveis[3]). É usual representá-lo pela letra R, indexada com os símbolos das duas variáveis ($R_{Y,X}$).

Podemos exemplificar o uso deste coeficiente calculando-o para os escores dos candidatos nas provas de Biologia e História do CV?99. A Figura 1 apresenta o diagrama de dispersão para tais escores; cada "pétala" dos "girassóis" representa quarenta pontos, quarenta pares de escores (os pares de escores em Biologia e História de cada candidato) e o "centro" dos "girassóis" entre um e quarenta pares de escores. Por exemplo, onde se vê um "girassol" com seis "pétalas"[4], há entre 241 e 280 pares de escores.

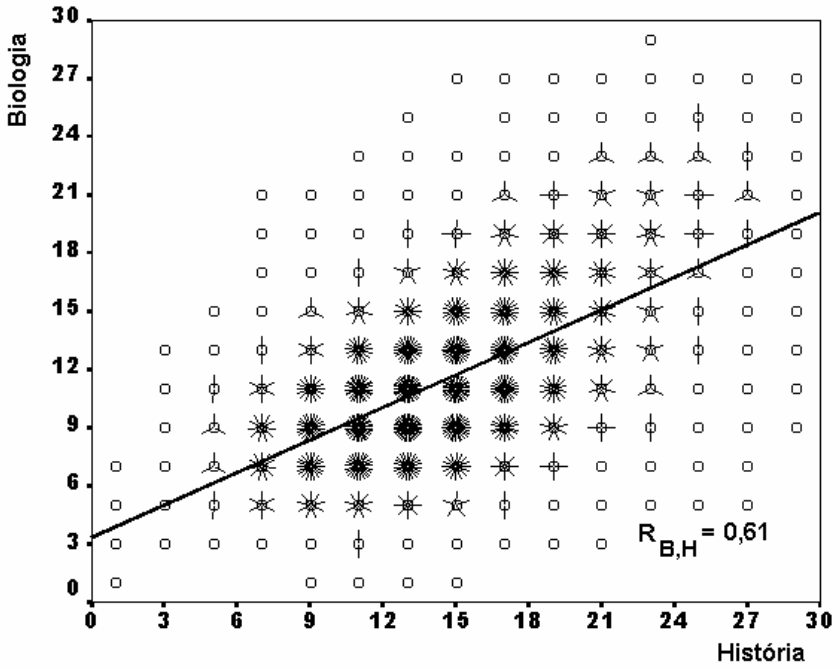

Figura 1 ? Diagrama de dispersão, coeficiente de correlação e reta de regressão dos escores em Biologia e História para os candidatos ao CV?99 da UFRGS.

O diagrama de dispersão mostra haver uma associação entre os dois escores: quando cresce o escore em História tende a crescer o escore em Biologia e vice-versa. Tal tendência está quantificada no coeficiente de correlação (0,61); o fato dele ser inferior à unidade, indica que nem toda a variação de uma variável é concomitante com a da outra. Em outras palavras, se um candidato possui escore elevado (baixo) em História, é provável que ele tenha um escore elevado (baixo) em Biologia. Entretanto, o leitor encontrará no diagrama de dispersão casos onde tal não ocorre e, por isso, o coeficiente de correlação é inferior a um.

A reta que está representada no diagrama de dispersão (denominada reta de regressão dos escores em Biologia contra os escores em História) é a reta dos mínimos quadrados. O coeficiente de correlação é a declividade da reta de regressão, com ambas as variáveis padronizadas em escores[5] z (Cronbach, 1996; Guilford e Fruchter, 1973; Wherry, 1984). Sendo o coeficiente de

[3] ? Na verdade o coeficiente de correlação de Pearson constitui-se em uma medida quase-universal de relação entre duas variáveis pois ele em módulo é: 1 ? invariante frente à transformações lineares em qualquer das variáveis; 2 ? quase-invariante frente a transformações monotônicas em qualquer das variáveis (Nunnally, 1978; Silveira, 1993).

[4] ✳

[5] ? Padronizar uma variável em escores z significa transformá-la linearmente de tal forma que a sua média seja nula e o desvio padrão igual a um. Para isso calcula-se a razão entre o resíduo da variável (diferença entre cada valor da variável e a média) e o desvio padrão. A variável padronizada z, que não tem unidade de medida (é adimensional), possui propriedades importantes em conseqüência da desigualdade de Chebychev (vide adiante a nota 13).

correlação a declividade da reta de regressão com as variáveis padronizadas, ele possui sinal. O sinal positivo indica a tendência das duas variáveis crescerem ou diminuírem concomitantemente; o sinal negativo expressa a tendência para que crescendo uma variável, a outra diminua.

Outra propriedade notável do coeficiente de correlação é que o seu quadrado ($R^2_{Y,X}$) determina a percentagem da variância de uma variável compartilhada com a outra. Esta propriedade permite a construção do diagrama de Venn (Kerlinger, 1980) da Figura 2, onde os círculos representam 100% da variância de cada variável e a interseção representa a percentagem da variância de Y que está associada com X ou é explicada[6] por X.

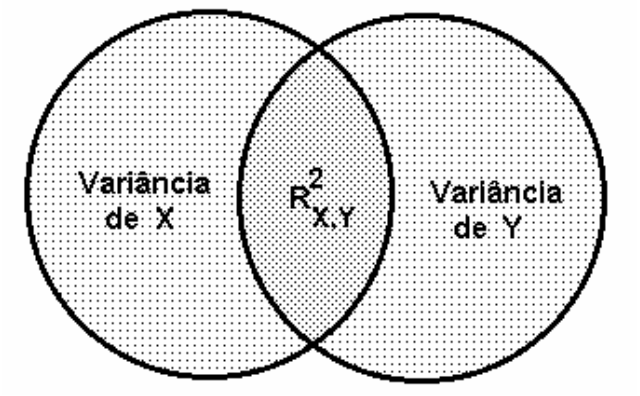

Figura 2 ? Diagrama de Venn representando a percentagem variância compartilhada por duas variáveis.

O conceito de correlação abrange a quantificação do grau de associação entre outros tipos de variáveis, além de variáveis quantitativas ou intervalares. Especificamente nos interessa o coeficiente de correlação entre uma variável quantitativa Y e uma variável categórica ou nominal X: o coeficiente eta ($?_{Y,X}$). A relação que este coeficiente possui com o de Pearson pode ser encontrada, por exemplo, em Wherry (1984). O coeficiente eta resulta sempre em um valor no intervalo fechado de zero a um; ele é nulo quando todas as categorias possuem a mesma média, crescendo quando a variância das médias de Y nas diversas categorias crescer; ele é um quando, dentro de cada categoria da variável X, os escores Y são iguais. O quadrado de eta é a percentagem da variância da variável Y explicada pela (compartilhada com a) variável nominal X. O quadrado de eta pode ser obtido dividindo-se a variância das médias de Y nas diversas categorias de X pela variância total de Y; maiores detalhes sobre o cálculo pode-se encontrar em Ferguson (1976), Guilford e Fruchter (1973) e Wherry (1984).

Exemplificamos a utilização deste coeficiente no estudo da relação entre o desempenho em Biologia no CV?99/UFRGS e o tipo de ensino médio que os candidatos cursaram. A Figura 3 constitui-se em um gráfico onde estão representadas as médias de acertos em Biologia nos grupos de candidatos, discriminados de acordo com a modalidade de ensino médio cursado; a barra se estende, em torno da média, por um desvio padrão dos escores de Biologia, dando-nos uma idéia sobre a variabilidade desses escores nos diversos tipos de ensino médio.

[6] ? Notamos anteriormente que o coeficiente de correlação é a declividade da reta de regressão, com ambas as variáveis padronizadas em escores z. Essa padronização torna a variância de cada variável unitária; por isso, o quadrado do coeficiente de correlação é o percentual da variância compartilhada. Ou seja, nos diagramas de Venn, os círculos sempre têm a mesma área, que representa 100% da variância de cada variável.

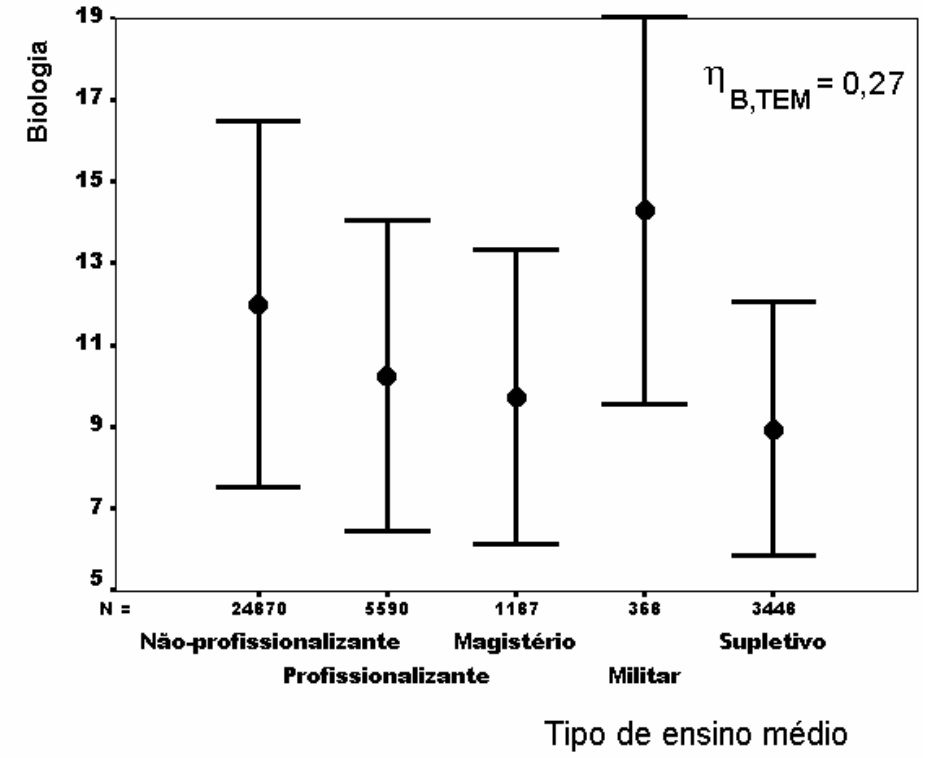

Figura 3 ? Relação do desempenho na prova de Biologia no CV? 99/UFRGS com o tipo de ensino médio que os candidatos cursaram.

Conforme indicado na figura, o coeficiente de correlação entre o desempenho em Biologia (B) e o tipo de ensino médio (TEM) cursado pelos candidatos é 0,27 (? $_{B,TEM}$=0,27). Esse coeficiente demonstra haver alguma relação entre as duas variáveis; o significado da relação pode ser estabelecido através das diferenças entre as médias do desempenho nas 5 categorias: os candidatos oriundos de escolas militares são os que, em média possuem o maior desempenho; a seguir, aparecem os alunos que cursaram ensino médio não-profissionalizante; depois, os que fizeram ensino médio profissionalizante, seguido por magistério e finalmente supletivo. O fato do coeficiente ser distante da unidade (valor máximo possível) se deve a que dentro de cada modalidade de ensino médio existe uma variabilidade grande no desempenho, conforme mostram as barras centradas nas médias.

Os coeficientes de correlação possibilitam comparar poderes explicativos de diferentes variáveis. Assim, observando os coeficientes apresentados nas figuras 1 e 3, conclui-se que há uma associação mais intensa entre o desempenho em Biologia com o desempenho em História do que com o tipo de ensino médio cursado pelos candidatos. Precisando melhor, o desempenho em Biologia compartilha cerca de 37% ($0,61^2 = 0,37$) da sua variância com o desempenho em História, enquanto compartilha apenas 7% ($0,27^2 = 0,07$) da sua variância com o tipo de ensino médio realizado pelos candidatos.

4. Construção de uma única medida de desempenho a partir dos escores nas nove provas do CV? 99/UFRGS

Conforme já relatado, tínhamos nove escores de desempenho no CV?99/UFRGS para cada candidato, um para cada prova. Por uma questão de parcimônia e inteligibilidade gostaríamos, se possível, de reduzir a apenas uma medida estes nove escores; esta variável única expressaria então o desempenho global de cada candidato naquele concurso. Se for possível a construção de tal medida

única de desempenho[7], esta será analisada em termos das relações com as variáveis sócio-econômico-culturais dos candidatos[8].

Para discutir a licitude ou não de uma única medida de desempenho, buscamos primeiramente os coeficientes de correlação entre as nove provas. Esses coeficientes são apresentados em uma matriz de correlações na Tabela 1.

Tabela 1 ? Matriz de correlações entre as nove provas do CV? 99/UFRGS.

PROVA	Biol.	Fís.	Geo.	Hist.	L. Estr.	Liter.	Mat.	Port.	Quí.
Biologia		0,63	0,55	0,61	0,46	0,60	0,54	0,53	0,58
Física	0,63		0,63	0,55	0,49	0,53	0,67	0,53	0,69
Geografia	0,55	0,63		0,62	0,57	0,56	0,53	0,61	0,58
História	0,61	0,55	0,62		0,51	0,65	0,50	0,58	0,51
Língua Estrangeira	0,46	0,49	0,57	0,51		0,55	0,42	0,63	0,48
Literatura	0,60	0,53	0,56	0,65	0,55		0,47	0,61	0,51
Matemática	0,54	0,67	0,53	0,50	0,42	0,47		0,51	0,57
Português	0,53	0,53	0,61	0,58	0,63	0,61	0,51		0,53
Química	0,58	0,69	0,58	0,51	0,48	0,51	0,57	0,53	
Coeficiente de correlação médio	**0,56**	**0,59**	**0,58**	**0,57**	**0,51**	**0,56**	**0,53**	**0,57**	**0,56**

Observa-se na Tabela 1 que qualquer uma das provas apresenta correlação positiva com todas as outras. Também é notório que tais correlações são bastante homogêneas, situando-se entre 0,42 e 0,67; o diagrama de dispersão para os escores em qualquer par de provas será semelhante ao apresentado na Figura 1. Em média cada prova correlaciona-se entre 0,51 e 0,59 com as demais. Portanto, há uma tendência para que candidatos com escore elevado (baixo) em alguma prova, possuam escore elevado (baixo) em qualquer outra. Realmente isso não se constitui em uma especificidade das nossas medidas de desempenho pois, reiteradamente, por quase um século, a partir dos estudos de Alfred Binet em 1905 (Nunnaly, 1978), resultados semelhantes têm sido encontrados. Em nossa realidade, Silveira (1996 e 1997) encontrou correlações semelhantes.

Como os escores nas provas estão todos relacionados positivamente, um escore total nas nove provas (somatório do número de acertos nas duzentos e setenta questões constituintes das nove provas) condensará todos eles em uma única medida[9]. Esta única medida de desempenho guardará correlações importantes com cada um dos nove escores parciais; ou seja, com uma única medida de desempenho conseguiremos representar muito bem os nove escores parciais. A forma de verificar que o escore em cada uma das nove provas está muito bem representado pelo escore total é

[7] ? Denomina-se *"Análise de Consistência Interna"* (Nunnaly, 1978; Silveira, 1993) o procedimento através do qual se estuda a possibilidade de condensar diversas variáveis em uma única.

[8] ? É importante notar que o objetivo desse trabalho é estudar como o desempenho dos candidatos, quantificado nos escores das diversas provas do CV-99, está relacionado com variáveis sócio-econômico-culturais. Não pretendemos explicar o sucesso (classificação para algum curso) ou o fracasso dos concorrentes naquele concurso. O sucesso ou fracasso, apesar de guardar alguma relação com o desempenho, depende também da taxa candidato/vaga, de tal forma que candidatos com alto desempenho (elevados escores nas nove provas) podem não ser classificados porque disputam com alta concorrência, ou, candidatos com desempenhos não tão altos se classificam em cursos de menor disputa.

[9] ? Caso a matriz de correlação mostrasse que alguns escores parciais se correlacionavam mais fortemente entre si do que com outros, dois ou mais escores totais construídos a partir das variáveis mais interrelacionadas seriam necessários para representar as nove medidas. As técnicas de Análise Fatorial ou Análise de Fatores (Mulaik, 1972; Nunnally, 1978; Spearritt, 1997), que não discutiremos aqui, são apropriadas em tais situações.

calculando o coeficiente de correlação de cada escore parcial com o total; a Tabela 2 apresenta estes coeficientes.

Tabela 2 ? Coeficientes de correlação do escore em cada prova com o escore total no CV? 99/UFRGS.

Prova	Coeficiente de correlação com o escore total
Biologia	0,77
Física	0,81
Geografia	0,81
História	0,79
Língua Estrangeira	0,75
Literatura	0,78
Matemática	0,73
Português	0,80
Química	0,77

O escore total, por correlacionar-se no mínimo com um coeficiente de 0,73 com cada prova, as representa bastante bem; além disso, possui a propriedade de ser uma medida mais estável, fidedigna que os escores parciais. A variância do escore total, conforme demonstrou Cronbach (1967), pode ser decomposta em uma parcela atribuída ao que há de comum entre os nove escores parciais e uma outra parte de erro de medida[10]. A estimativa desta parcela estável, fidedigna, comum às nove provas, é quantificada no coeficiente alfa (?) de Cronbach (Cronbach, 1996; Lord e Novick, 1968; Silveira, 1993; Thorndike e Thorndike, 1997). A Tabela 3 apresenta a média, o desvio padrão e o coeficiente alfa do escore total nas nove provas entre os 35463 candidatos ao CV? 99/UFRGS.

Tabela 3 ? Características do escore total nas nove provas para os 35463 candidatos ao CV? 99/UFRGS.

Média	Desvio padrão	Coeficiente de fidedignidade (coeficiente ?)
111,87	34,19	0,92

O fato do coeficiente de fidedignidade ser elevado (0,92) significa que a variância do escore total nas nove provas é virtualmente explicável em quase sua totalidade[11]; apenas 8% da variância é

[10] ? Este teorema respalda teoricamente a construção de escores totais e médias, seja em provas individuais, seja sobre diversas provas. Preferimos quase sempre avaliar nossos alunos através de medidas que acumulam escores parciais em diversos itens (questões) e depois em diversas provas; este procedimento, normalmente realizado de maneira tácita e acrítica, encontra suporte na teoria da medida psicológica e educacional.

[11] ? A relação do coeficiente de fidedignidade de uma variável com os coeficientes de correlação que ela pode apresentar com outras variáveis é discutida em Cronbach (1967; 1996), Guilford e Fruchter (1973), Lord e Novick (1968), Nunnally (1978), Thorndike e Thorndike, (1997) e Wherry (1984).

atribuída a erros de medida, não podendo ser compartilhada com qualquer variável potencialmente explicativa do desempenho no CV?99/UFRGS.

Finalmente, para tornar o escore total de desempenho facilmente interpretável, o padronizamos. A padronização adotada foi uma transformação linear[12] que o levou a ter média 500 e desvio padrão[13] 100. O histograma do escore total bruto e padronizado é apresentado na Figura 4, bem como a equação que calcula o escore padronizado (EP) a partir do escore total bruto (Total).

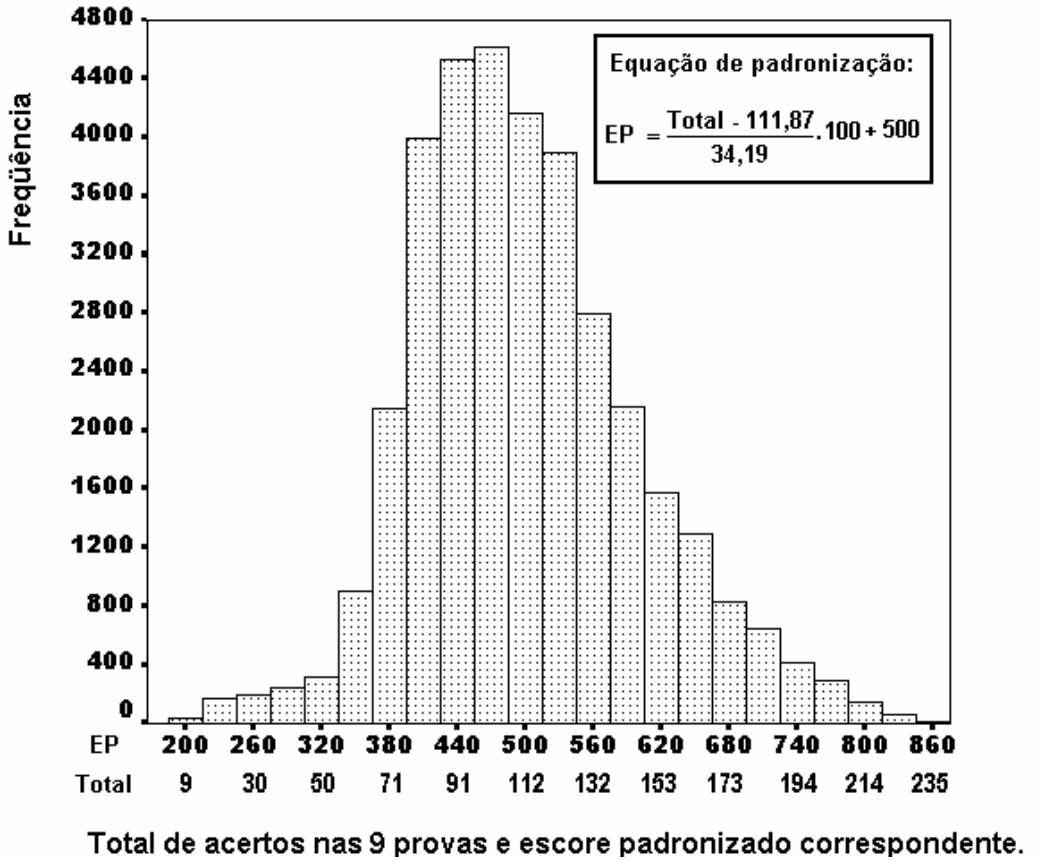

Figura 4 ? Histograma dos escores total bruto e padronizado para os 35463 candidatos ao CV? 99/UFRGS.

Desta forma, mostramos como condensar as nove variáveis de desempenho em uma única. Esta única terá a sua variância analisada pelos fatores sócio-econômico-culturais dos candidatos.

[12] Transformações lineares da variável a ser explicada (aqui o escore total) não afetam as correlações com as variáveis explicativas (aqui as sócio-econômico-culturais).

[13] ? O escore padronizado é facilmente interpretável pois cerca de dois terços dos candidatos têm tal escore compreendido entre 400 e 600, cerca de 95% dos candidatos entre 300 e 700 e a quase totalidade dos mesmos entre 200 e 800. Estas proporções independem da média e do desvio padrão dos escores brutos, dependendo apenas da forma da distribuição (suposta como aproximadamente gaussiana). Mesmo que a distribuição não seja normal (gaussiana), os escores padronizados ainda são interpretáveis através da desigualdade de Chebychev (Bock, 1975; Sveshnikov, 1978). Esta afirma que, independentemente da forma da distribuição, haverá no mínimo 75% dos candidatos com escore padronizado entre 300 e 700, no mínimo 89% dos candidatos com escore padronizado entre 200 e 800 e no mínimo 94% dos candidatos com escore entre 100 e 900.

5. Quantificando o poder explicativo de duas ou mais variáveis sobre outra

Na seção 3 vimos como é possível quantificar a relação entre duas variáveis através do coeficiente de correlação. Este procedimento pode ser generalizado para quantificação do poder explicativo que duas ou mais variáveis possuem sobre uma outra variável de interesse. No nosso caso, gostaríamos de saber quanto da variância do desempenho no CV?99/UFRGS é compartilhada com as 15 variáveis sócio-econômico-culturais.

O conceito aplicável a este problema mais geral é o da correlação múltipla (Afifi e Clark, 1996; Ferguson, 1976, Guilford e Fruchter, 1973; Nunnally, 1978, Wherry, 1984; Tatsuoka, 1997). Caso as variáveis explicativas sejam ortogonais (não-correlacionadas) entre si, a percentagem da variância explicada por todas elas em conjunto (quadrado do coeficiente de correlação múltipla) é o somatório das variâncias que cada uma delas individualmente compartilha com a variável em questão. A Figura 5 representa tal possibilidade com apenas duas variáveis ortogonais (X_1 e X_2).

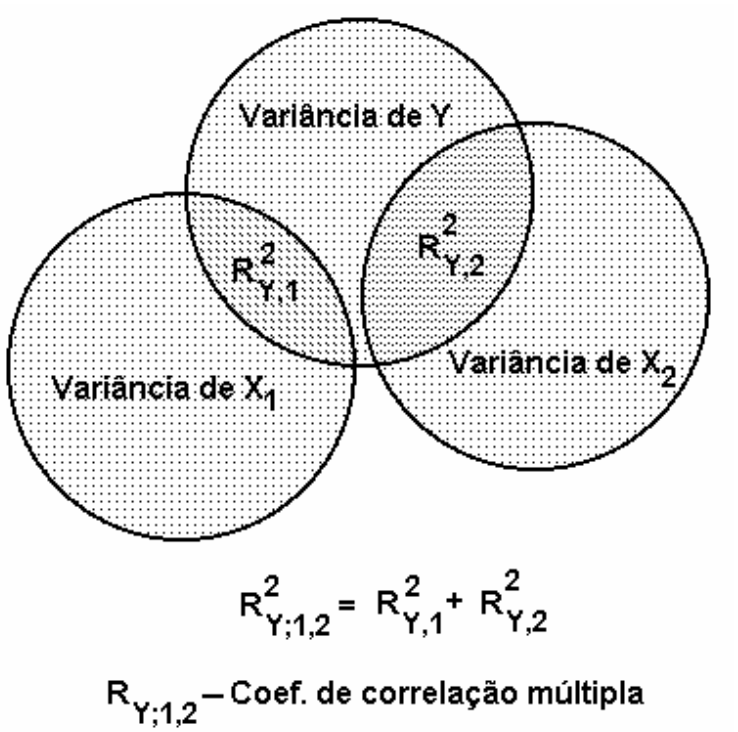

Figura 5 ? Diagrama de Venn representando a explicação da variável Y por duas variáveis ortogonais entre si.

Quando as variáveis explicativas forem correlacionadas (não-ortogonais) entre si, a variância explicada por todas elas conjuntamente (quadrado do coeficiente de correlação múltipla) envolverá cálculos mais complexos; esse coeficiente depende das correlações que cada variável explicativa tem com a explicada e das correlações entre as variáveis explicativas (Bock, 1975; Nunnally, 1978; Tatsuoka, 1997). A Figura 6 representa a explicação de uma variável Y por duas variáveis correlacionadas.

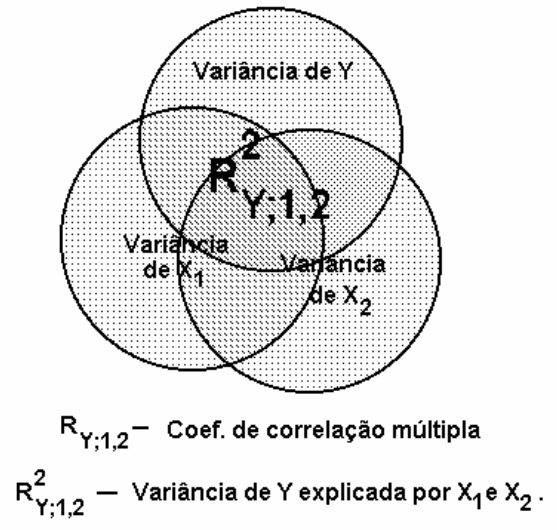

Figura 6 ? Diagrama de Venn representando a explicação da variável Y por duas variáveis correlacionadas entre si.

É interessante notar na Figura 6 que a interseção da variância de Y com aquelas das duas variáveis é composta por três regiões distintas. Uma região central, interseção das três variáveis, representando redundância de explicação por X_1 e X_2, e outras duas regiões laterais, representando explicações exclusivas de X_1 e de X_2. Essas duas regiões laterais estão associadas com os coeficientes de correlação parciais (Guilford e Fruchter, 1973; Nunnaly, 1978; Wherry, 1984).

Caso Y venha a ser explicada por mais de duas variáveis, a representação gráfica torna-se complexa pois o espaço das variáveis é multidimensional, com dimensão igual ao número total de variáveis. Entretanto, o quadrado do coeficiente de correlação múltipla de Y com todas as variáveis explicativas continua sendo a proporção da variância de Y explicada por todas elas. Os cálculos envolvidos em uma situação multivariada, apesar de extensos e complexos, são rapidamente realizados por programas computacionais de estatística (Afifi e Clark, 1996).

6. As variáveis explicativas do desempenho no CV? 99/UFRGS

No momento da inscrição ao CV? 99/UFRGS os candidatos responderam ao Questionário de Informações sobre o Candidato. Essas respostas foram codificadas em 15 variáveis nominais, divididas em dois grandes grupos. O primeiro grupo, constituído por 8 variáveis *prima facie* de nível sócio-econômico. As variáveis desse grupo estão descritas sucintamente na Tabela 4; no Apêndice (tabelas A1 a A8) elas estão apresentadas de forma mais completa, incluindo também estatísticas relevantes aos propósitos desse estudo.

Tabela 4 ? Variáveis sócio-econômicas dos candidatos ao CV? 99/UFRGS.

Nome da variável	Número de categorias
Renda familiar	6
Dependentes da renda familiar	6
Exercício de atividade remunerada pelo candidato	4
Ocupação principal do candidato	30
Ocupação principal do pai do candidato	30
Ocupação principal da mãe do candidato	30
Nível de instrução do pai do candidato	8
Nível de instrução da mãe do candidato	8

O segundo grupo de variáveis, integrado por 7 variáveis, forneceu informações sobre a escolaridade do candidato. No Apêndice (tabelas A9 a A15) elas estão descritas de forma mais completa do que na tabela que se segue.

Tabela 5 ? Variáveis culturais ou de escolaridade dos candidatos ao CV? 99/UFRGS.

Nome da variável	Número de categorias
Tipo de ensino médio cursado	5
Tipo de estabelecimento de ensino médio freqüentado	2
Turno em que realizou o ensino médio	2
Realização de curso pré-vestibular	4
Realização de concursos vestibulares anteriores	6
Nome da escola de ensino médio freqüentada	143
Nível de instrução do candidato	5

As categorias de cada variável são mutuamente exclusivas entre si. Assim, um particular candidato constou em apenas uma categoria de cada variável.

O número total de inscritos no CV?99/UFRGS foi 39411; entretanto, o nosso estudo realizou-se com 35463 candidatos: aqueles que efetivamente participaram do concurso e que responderam ao Questionário de Informações (o preenchimento do questionário não era obrigatório).

Ao dividirmos as variáveis em dois grupos, não estamos supondo que esses dois grupos sejam ortogonais, não-correlacionados. Inclusive admitimos *a priori* que as variáveis de escolaridade estejam relacionadas com as sócio-econômicas; da mesma forma, admitimos *a priori* haver relações entre as variáveis de cada grupo. Entretanto, a questão de saber qual era de fato a intensidade dessas relações e como elas afetavam o poder explicativo sobre o desempenho no CV?99/UFRGS, será estudado adiante. Aliás, os procedimentos multivariados constituem-se em poderosas formas de análise dessas relações complexas.

7. As relações das variáveis sócio-econômicas com o desempenho no CV? 99/UFRGS

A Tabela 6 apresenta a percentagem da variância do desempenho no CV?99/UFRGS que cada uma das variáveis sócio-econômicas explicou (quadrado do coeficiente de correlação).

Tabela 6 ? Percentagem da variância do desempenho no CV? 99 da UFRGS explicada por cada variável sócio-econômica.

Nome da variável	Percentagem da variância explicada
Renda familiar	10,5*
Dependentes da renda familiar	1,2*
Exercício de atividade remunerada pelo candidato	2,9*
Ocupação principal do candidato	5,5*
Ocupação principal do pai do candidato	4,5*
Ocupação principal da mãe do candidato	3,5*
Nível de instrução do pai do candidato	8,1*
Nível de instrução da mãe do candidato	7,6*

* ? estatisticamente significativa em nível inferior a 0,001.

A relação de cada variável com o desempenho pode ser expressa também pelas médias do desempenho através das diferentes categorias (essas encontram-se nas tabelas A1 a A8 do Apêndice), mostrando o padrão da relação. Por exemplo, as médias do desempenho crescem quando a faixa de renda familiar (vide Tabela A1 do Apêndice) aumenta. Para a faixa de menos de 1 salário mínimo, o desempenho médio é 440, atingindo a média de 559 na faixa de 30 salários ou mais.

De um modo geral, a inspeção das tabelas do Apêndice mostram que candidatos oriundos de estratos sócio-econômicos mais elevados possuem, em média, desempenhos no CV? 99/UFRGS maiores. Entretanto, nenhuma variável sócio-econômica individualmente explicou mais do que 10,5% da variância do desempenho. Se as variáveis sócio-econômicas fossem ortogonais entre si, o poder explicativo conjunto seria 43,8% (somatório das variâncias explicadas da Tabela 6).

A Tabela 7 mostra o efetivo poder explicativo das 8 variáveis sócio-econômicas. Ele foi obtido através de uma Análise da Variância ? ANOVA ? (Afifi e Clark, 1996; Bock, 1975; Wherry, 1984; Tatsuoka, 1997), tendo o desempenho no CV? 99/UFRGS como variável dependente e aquelas 8 variáveis como fatores (variáveis nominais).

Tabela 7 ? Explicação conjunta das variáveis sócio-econômicas sobre o desempenho dos candidatos ao CV? 99/UFRGS.

Variáveis	Coeficiente de correlação múltipla	Percentagem da variância explicada
Oito variáveis sócio-econômicas	0,390*	15,2*

* ? estatisticamente significativo em nível inferior a 0,001.

O fato da explicação conjunta ser de 15,2% ? apenas pouco mais de um terço de 43,8%, proporção que ocorreria se as variáveis fossem não-correlacionadas entre si ? mostra que as variáveis sócio-econômicas, conforme admitíamos *a priori*, estavam realmente interrelacionadas.

8. A relação das variáveis culturais ou de escolaridade com o desempenho no CV? 99/UFRGS

A Tabela 8 apresenta a percentagem da variância do desempenho no CV? 99/UFRGS que cada uma das variáveis culturais ou de escolaridade explicou (quadrado do coeficiente de correlação).

Tabela 8 ? Percentagem da variância do desempenho no CV? 99/UFRGS explicada por cada variável de escolaridade.

Nome da variável	Percentagem da variância explicada
Tipo de ensino médio cursado	7,5*
Tipo de estabelecimento de ensino médio freqüentado	4,3*
Turno em que realizou o ensino médio	4,2*
Realização de curso pré-vestibular	12,5*
Realização de concursos vestibulares anteriores	6,0*
Nome da escola de ensino médio freqüentada	16,6*
Nível de instrução do candidato	1,6*

* ? estatisticamente significativa em nível inferior a 0,001.

A relação de cada variável com o desempenho pode ser expressa também pelas médias do desempenho através das diferentes categorias (vide as tabelas A9 a A15 do Apêndice), mostrando o significado da relação. Por exemplo, a relação que o tipo de estabelecimento de ensino médio freqüentado pelo candidato teve com o desempenho (explicando 4,3% da variância) também pode ser vista na diferença entre as médias dos candidatos que freqüentaram escola pública e escola particular (as duas categorias da variável): respectivamente 478 e 519 (vide a tabela A10 do Apêndice).

Observa-se na Tabela 8 que o nome da escola de ensino médio freqüentada foi a variável com maior poder explicativo (16,6%), seguida da variável que indicava se o candidato realizou ou não curso pré-vestibular (12,5%). Se as variáveis de escolaridade fossem ortogonais entre si, a percentagem da variância explicada por todas elas conjuntamente seria 52,7% (somatório das variâncias explicadas). Uma ANOVA do desempenho no CV? 99/UFRGS, tendo como fatores as 7 variáveis de escolaridade mostrou qual foi o efetivo poder explicativo (vide a Tabela 9).

Tabela 9 ? Explicação conjunta das variáveis de escolaridade sobre o desempenho dos candidatos ao CV? 99/UFRGS.

Variáveis	Coeficiente de correlação múltipla	Percentagem da variância explicada
Sete variáveis de escolaridade	0,559*	31,3*

* ? estatisticamente significativo em nível inferior a 0,001.

O fato da explicação conjunta ser 31,3% ? apesar de pouco mais da metade de 52,7%, proporção que ocorreria se as variáveis fossem ortogonais entre si ? mostra que as variáveis de escolaridade, conforme admitíamos *a priori*, estavam de fato interrelacionadas.

9. A relação de todas as variáveis com o desempenho no CV? 99/UFRGS

A Figura 7 sintetiza os resultados encontrados na duas seções anteriores, mostrando que as 7 variáveis de escolaridade tiveram praticamente o dobro da explicação das 8 variáveis sócio-econômicas.

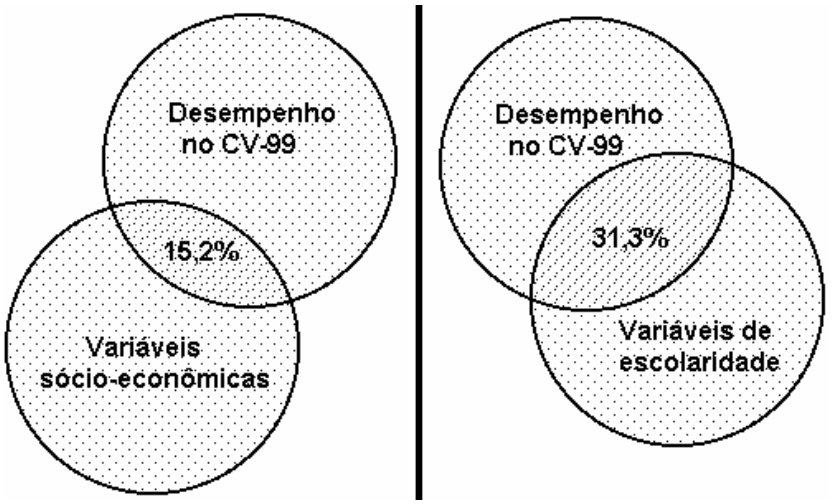

Figura 7 ? Diagramas representando a explicação das variáveis sócio-econômicas e de escolaridade separadamente sobre o desempenho no CV? 99 da UFRGS.

Se os dois conjuntos de variáveis fossem ortogonais entre si, a explicação das 15 variáveis atingiria 46,5% da variância do desempenho (15,2% + 31,3%). Realizamos uma ANOVA do desempenho tendo como fatores as 15 variáveis, para encontrarmos o poder explicativo efetivo desse conjunto (vide Tabela 10).

Tabela 10 ? Explicação conjunta das variáveis sócio-econômicas e de escolaridade sobre o desempenho dos candidatos ao CV? 99/UFRGS.

Variáveis	Coeficiente de correlação múltipla	Percentagem da variância explicada
Oito variáveis sócio-econômicas e sete variáveis de escolaridade	0,585*	34,2*

* ? estatisticamente significativo em nível inferior a 0,001.

Destaca-se que o poder explicativo das 15 variáveis excede por apenas um pouco (2,9%) o das 7 variáveis de escolaridade. Tal se deve, conforme admitido *a priori*, às interrelações entre os dois conjuntos de variáveis (sócio-econômicas e culturais). A Figura 8 representa esse importante resultado.

Figura 8 ? Diagrama representando a explicação conjunta das variáveis sócio-econômicas e de escolaridade sobre o desempenho no CV? 99/UFRGS.

A variância explicada pelas 15 variáveis pode ser decomposta em três partes (vide a Figura 9). Um delas, perfazendo 12,3% da variância do desempenho, representa a parcela da explicação redundante, isto é, comum aos dois grupos de variáveis. A outra, perfazendo apenas 2,9%, representa a explicação das variáveis sócio-econômicas não superposta, independente das variáveis de escolaridade. Finalmente, a terceira parcela, perfazendo 19%, representa a explicação das variáveis de escolaridade não superposta, independente das variáveis sócio-econômicas.

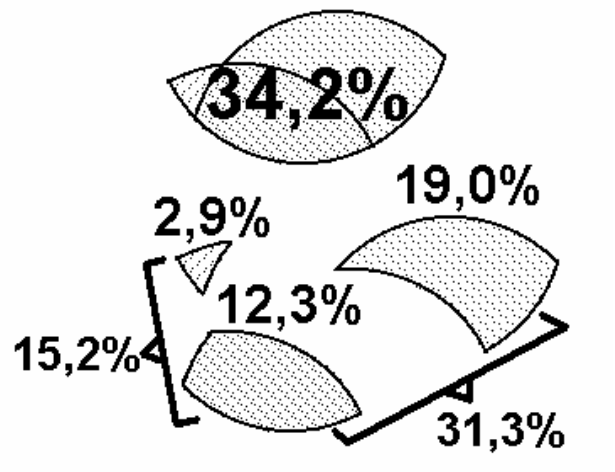

Figura 9 ? Decomposição da variância explicada do desempenho no CV? 99 da UFRGS em três parcelas.

Assim, a maior parte da variância explicada (19,0%) é atribuída à escolaridade dos candidatos independentemente de seus níveis sócio-econômicos. As variáveis sócio-econômicas, mesmo admitido um "efeito" indireto através da escolaridade, explicam no máximo 15,2% da variância do desempenho (2,9% exclusivamente e mais 12,3% superposta com a escolaridade).

10. Conclusão

O objetivo principal desse trabalho foi o de apresentar técnicas de análise quantitativa multivariada, enfatizando os aspectos conceituais das mesmas. A exemplificação dos procedimentos utilizados aconteceu em uma situação concreta com 24 variáveis de 35463 candidatos ao Concurso Vestibular de 1999 da UFRGS. O estudo teve como objetivo estabelecer o poder explicativo que 15 variáveis sócio-econômico-culturais tiveram sobre o desempenho nas 9 provas daquele concurso.

A técnica utilizada na determinação do poder explicativo foi a Análise da Variância (ANOVA); a quantificação das relações entre as variáveis foi efetivada via coeficientes de correlação e médias do desempenho nas categorias das variáveis explicativas. Mostramos também uma Análise de Consistência Interna, tendo como alvo a condensação das 9 variáveis de desempenho em uma única. Queremos ainda alertar o leitor para o fato de que a Análise da Variância pode ser aplicada em situações que envolvam mais de uma variável explicada e diversas variáveis explicativas; as variáveis explicativas não necessitam ser apenas variáveis nominais como no nosso caso.

Julgamos o próprio resultado do estudo que serviu de exemplo como extremamente importante, pois, como destacamos no início, em nossa realidade proliferam opiniões mas faltam estudos objetivos sobre o poder que fatores sócio-econômicos e culturais têm nos resultados dos concursos vestibulares. Mostramos que o conjunto das quinze variáveis sócio-econômico-culturais explicaram 34,2% da variância, isto é, o restante da variância do desempenho (65,8%) não pode ser atribuída a essas variáveis, dependendo talvez de fatores pessoais, psicológicos e vivenciais dos candidatos. Adicionalmente, encontramos dentro da variância explicada a maior parcela associada à escolaridade dos candidatos, independentemente dos fatores sócio-econômicos. Tais resultados são incompatíveis com posicionamentos teóricos reducionistas que pretendem ser o desempenho nos concursos vestibulares exclusivamente determinado por fatores sócio-econômicos.

Agradecimento

Agradeço à professora Maria Cristina Varriale pela leitura crítica deste trabalho e pelas valiosas sugestões que permitiram aprimorá-lo.

Bibliografia

AFIFI, A. A. e CLARK, V. *Computer-aided multivariate analysis.* London: Chapman & Hall, 1996.

BOCK, R. D. *Multivariate statistical methods.* New York: McGraw-Hill, 1975.

CRONBACH, L.J. Coefficient alpha and the internal structure of tests. In: MEHRENS, W. A. e EBEL, R. L. (org.) *Principles of educational and psychological measurement.* Chicago: Rand McNally, 1967.

_____ *Fundamentos da testagem psicológica.* Porto Alegre: Artes Médicas, 1996.

FERGUSON, G. A. *Statistical analysis in psychology and education.* Tokyo: McGraw-Hill Kogagusha, 1976.

GUILFORD, J. P. e FRUCHTER, B. *Fundamental statistics in psychology and education.* New York: McGraw-Hill, 1973.

KERLINGER, F. N. *Metodologia da pesquisa em ciências sociais: um tratamento conceitual.* São Paulo: EDUSP, 1979.

LORD, F. M. e NOVICK, M. R. *Statistical theories of mental test scores.* Menlo Park: Addison?Wesley, 1968.

MULAIK, S. A. *The foundations of factor analysis.* New York: McGraw-Hill, 1972.

NUNNALLY, J. C. *Psychometric theory*. New York: McGraw-Hill, 1978.

SILVEIRA, F. L. Validação de testes de papel e lápis. In: MOREIRA, M. A. e SILVEIRA, F.L. *Instrumentos de pesquisa em ensino e aprendizagem.* Porto Alegre: EDIPUCRS, 1993.

_____ Relação do desempenho no concurso vestibular da Universidade Federal do Rio Grande do Sul com diversas variáveis. *Estudos em Avaliação Educacional*, São Paulo, 14, pp. 83-103, 1996.

_____ Comparação entre três argumentos de concorrência para o concurso vestibular da Universidade Federal do Rio Grande do Sul. *Estudos em Avaliação Educacional*, São Paulo, 16, pp. 43?57, 1997.

SPEARRITT, D. Factor analysis. In: KEEVES, J. P. (org.) *Educational research, methodology, and measurement: an international handbook.* Cambridge: Pergamon, 1997.

SVESHNIKOV, A. A. *Problems in probability theory, mathematical statistics and theory of random functions.* New York: Dover, 1978.

WHERRY, R. J. *Contributions to correlational analysis.* London: Academic Press, 1984.

TATSUOKA, M. M. Regression analysis of quantified data. In: KEEVES, J. P. (org.) *Educational research, methodology, and measurement: an international handbook.* Cambridge: Pergamon, 1997.

THORNDIKE, R. L. E THORNDIKE, R. M. Reliability. In: KEEVES, J. P. (org.) *Educational research, methodology, and measurement: an international handbook.* Cambridge: Pergamon, 1997

Recebido em: 26.10.99
Aceito em: 10.12.99

APÊNDICE

Apresentamos a seguir as variáveis sócio-econômicas e de escolaridade desse estudo. Em cada tabela, indicamos o nome da variável, as diversas categorias que a compuseram, o número de candidatos em cada categoria (N) e a média desses candidatos no escore padronizado de desempenho no CV?99/ UFRGS. As categorias estão colocadas em ordem decrescente pela média; em variáveis com mais de vinte categorias, apresentamos apenas as dez primeiras e as dez últimas.

Tabela A1 ? Renda familiar.

Categoria	Média	N
Mais de 30 salários	559	3814
De 20 a 30 salários	536	3618
De 10 a 20 salários	517	8019
De 5 a 10 salários	490	10653
De 1 a 5 salários	460	8798
Até 1 salário	440	561

Tabela A2 ? Número de dependentes da renda familiar.

Categoria	Média	N
Cinco ou seis	510	9761
Quatro	507	11576
Sete ou mais	495	964
Três	492	6529
Dois	482	4409
Um	481	2224

Tabela A3 ? Exercício de atividade remunerada pelo candidato.

Categoria	Média	N
Não exerce	513	22645
Exerce eventualmente	489	1834
Exerce em tempo parcial	478	4851
Exerce em tempo integral	474	6133

Tabela A4 ? Ocupação principal do candidato.

Categoria	Média	N
Servidor público de nível superior	571	90
Técnico de nível superior	546	57
Proprietário de estabelecimento industrial	523	7
Oficial militar	519	51
Estudante	515	22926
Professor ensino médio	506	107
Proprietário de estabelecimento prestador de serviço	505	85
Outro servidor público	503	656
Diretor ou gerente de empresa	500	80
Técnico de nível superior	490	908
Administrador de empresa	474	1046
Proprietário de estabelecimento comercial	471	102
Trabalhador informal	468	1379
Desempregado	463	1334
Trabalhador da produção industrial	461	296
Outra	456	1110
Comerciário	454	1886
Do lar	448	376
Proprietário de estabelecimento agrícola	446	17
Trabalhador no setor primário	433	63

Tabela A5 ? Ocupação principal do pai do candidato.

Categoria	Média	N
Professor de ensino superior	559	423
Servidor público de nível superior	532	1165
Profissional liberal	531	4246
Diretor ou gerente de empresa	527	1644
Proprietário de estabelecimento industrial	527	323
Técnico de nível superior	525	531
Professor ensino médio	524	370
Oficial militar	524	455
Membro de um dos 3 Poderes	518	326
Proprietário de estabelecimento prestador de serviço	516	729
Técnico de nível médio	491	553
Trabalhador em navegação aérea ou marítima	491	108
Militar não-oficial	487	412
Comerciário	486	1836
Trabalhador da produção industrial	482	761
Outro servidor público	478	1009
Desempregado	478	817
Do lar	477	67
Outra	472	3822
Trabalhador informal	468	1339

Tabela A6 ? Ocupação principal da mãe do candidato.

Categoria	Média	N
Servidor público de nível superior	549	893
Professor de ensino superior	540	444
Profissional liberal	536	1980
Proprietário de estabelecimento prestador de serviço	534	202
Membro de um dos 3 Poderes	532	177
Oficial militar	530	3
Professor de ensino médio	524	1830
Técnico de nível superior	524	231
Diretor ou gerente de empresa	524	303
Proprietário de estabelecimento industrial	524	74
Técnico de nível médio	489	237
Do lar	488	12698
Desempregado	485	442
Trabalhador informal	484	802
Militar não-oficial	483	5
Comerciário	483	1327
Outra	475	2190
Trabalhador do setor primário	472	170
Trabalhador da produção industrial	468	244
Trabalhador em navegação aérea ou marítima	466	13

Tabela A7 ? Nível de instrução do pai do candidato.

Categoria	Média	N
Pós-graduação	556	2551
Superior completo	530	8812
Superior incompleto	512	3485
Ensino médio completo	491	6693
Ensino médio incompleto	482	2846
Ensino fundamental completo	474	3013

Ensino fundamental incompleto	469	7570
Não freqüentou escola	447	493

Tabela A8 ? **Nível de instrução da mãe do candidato.**

Categoria	Média	N
Pós-graduação	546	2360
Superior completo	533	7904
Superior incompleto	520	2862
Ensino médio completo	496	7972
Ensino médio incompleto	485	3020
Ensino fundamental completo	475	3522
Ensino fundamental incompleto	468	7336
Não freqüentou escola	449	487

Tabela A9 ? **Tipo de ensino médio freqüentado pelo candidato.**

Categoria	Média	N
Militar	585	368
Não-profissionalizante	514	24870
Profissionalizante	480	5590
Magistério	463	1187
Supletivo	434	3448

Tabela A10 ? **Tipo de estabelecimento de ensino médio freqüentado pelo candidato.**

Categoria	Média	N
Escola particular	519	18828
Escola pública	478	16635

Tabela A11 ? **Turno em que o candidato cursou o ensino médio.**

Categoria	Média	N
Diurno	509	29735
Noturno	454	5728

Tabela A12 ? **Realização de curso pré-vestibular pelo candidato.**

Categoria	Média	N
Realizou por mais de 1 ano	582	3515
Realizou por 1 ano	531	5164
Realizou por menos de 1 ano	503	11873
Não realizou	468	14911

Tabela A13 ? **Realização de concursos vestibulares anteriores.**

Categoria	Média	N
Realizou mais de quatro	543	1216
Realizou quatro	531	1064
Realizou três	527	2786
Realizou dois	523	5205
Realizou um	516	9084
Não realizou	474	16108

Tabela A14 ? **Nome da escola de ensino médio freqüentada.**

Categoria	Média	N
	592	833
	589	93
	588	258
	583	450
	578	110
	574	168
	572	416
	570	202
	568	88
	564	123
	440	112
	435	50
	434	15
	427	31
	426	12
	426	17
	422	34
	418	25
	418	77
	413	10

Observação: o nome das escolas foi omitido.

Tabela A15 ? **Nível de instrução do candidato.**

Categoria	Média	N
Pós-graduação	558	152
Superior completo	543	728
Superior incompleto	523	5217
Médio completo	497	20915
Médio incompleto	489	8451