A implicação estatística usada como ferramenta em um exemplo de análise de dados multidimensionais

Régis Gras[1]
Saddo Ag Almouloud[2]

Resumo

O artigo apresenta os resultados de um questionário cujo objetivo é analisar as opiniões dos participantes de um colóquio organizado pela PUC/SP sobre a implicação estatística de análise de dados multidimensionais. O colóquio foi realizado no intuito de estudar as potencialidades, a pertinência e a importância da realização de uma análise implicativa nas investigações das Ciências da Educação. A implicação estatística foi a principal ferramenta para analisar as informações envolvidas neste questionário.

Palavras-chave: implicação estatística, metodologia de análise de dados, hierarquia de similaridade, hierarquia implicativa, avaliação.

Resumé:

L´article présente les résultats d´un questionnaire dont l´objectif est l´étude des opinions des participants d´un colloque organisé à la Pontifícia Universidade Católica de São Paulo, sur l´analyse implicative de données statistiques multidimensionnelles. Le colloque a pour objectif d´étudier les potentialités, la pertinence et l´importance de l´analyse implicative pour les recherches en sciences de l´éducation. L´implication statistique a été l´outil principal de traitement et d´analyse des données issues de ce questionnaire.

Mots – Clés: implication statistique, méthodologie d´analyse de données, hiérarchie de similarité, hiérarchie implicative, évaliation.

I- Introdução

Este artigo discute os resultados de um questionário cujo objetivo é avaliar o colóquio intitulado “O método estatístico implicativo utilizado em estudos qualitativos de regras de associação: contribuição à pesquisa em educação”, realizado em julho de 2003 no Programa de Estudos pós-graduados em Educação Matemática da PUC/SP..

O colóquio tem por objetivo principal realizar um estudo de refinamento sobre as pontecialidades, a pertinência bem como a importância da realização de análises estatísticas

[1] Professeur Emérite à l’Ecole Polytechnique de l'Université de Nantes, La Chantrerie, BP 50609 44306 Nantes cedex 03, e-mail : regisgra@club-internet.fr

[2] Professor do Programa de Estudos Pós-Graduados em Educação Matematica – PUC/SP e-mail : saddoag@pucsp.br.

de dados multidimensionais (análise hierárquica de similaridade, análise implicativa) nas investigações da Educação Matemática e em investigações em Educação de um modo mais abrangente. Foram previstas para o desenvolvimento do trabalho três momentos:

1. Realização de estudo das fases fundamentais de uma análise de dados multidimensionais (instrumentos de coleta de dados, organização e exploração, instrumentos de tratamentos, interpretação, levando em conta a questão e os objetivos da pesquisa);
2. Realização de um breve estudo de carater teórico e intuitivo sobre os diferentes métodos de análise de dados multidimensionais e oferta de oficinas sobre os métodos;
3. Realização de estudo dos exemplos de dados encontrados em pesquisas desenvolvidas no Programa de Pós Graduação em Educação Matemática e em Educação e currículo da PUCSP;
4. organização de um colóquio sobre os métodos, aberto a distintos Programas de Pós-Graduação da PUC/SP, bem como de outros Universidades.

O colóquio conta com a participação ativa do Professor Emérito da Escola Politécnica da Universidade de Nantes Régis Gras. O referido autor e seu grupo de pesquisa, desde 1979, procuraram, entre outros assuntos, colocar à disposição dos pesquisadores (em matemática, em psicologia, em biologia, em educação, etc.) ferramentas estatísticas (a análise implicativa, a hierarquia implicativa) que permitem evidenciar a dinâmica dos comportamentos de sujeitos (alunos, por exemplos) em situação de resolução de problemas, no caso da educação matemática. A análise implicativa (cf. o artigo de Grãs nesta revista), como todos os métodos de análise estatística de dados multidimensionais, permite visualizar, organizar, construir modelos e explicar fenômenos associados aos dados.

II- Estrutura do questionário

Para a constituição do corpo de informações a ser analisado, foi aplicado um questionário-avaliação aos participantes do colóquio. O questionário(cf. anexo 1), além de informações pessoais, solicitava que os participantes respondessem a quatro questões subjetivas, que objetivavam fornecer uma visão de como eles vivenciaram a programação e os conteúdos trabalhados no evento.

As questões subjetivas foram as seguintes:

1) Suas expectativas foram satisfeitas:

( ) Sim *( ) Parcialmente* *( ) Não*

2) Quais os pontos da programação despertaram mais seu interesse?

3) Quais aspectos você desejaria aprofundar mais?

4) Na lista abaixo, escolhe a(s) expressão(ões) que melhor expressa(m) seus sentimentos a propósito do colóquio.

( ) muito longo	*( ) muito curto*	*( ) Cansativo*	*( ) Estimulante*
( ) Curiosidade	*( ) muito bom*	*( ) Inútil*	*( ) Abertura*
( ) difícil	*( ) gostaria fazer mais*		

As respostas foram codificadas (cf. anexo 1) e as variáveis estatísticas que dizem respeito aos dados pessoais foram consideradas como variáveis suplementares(o código dessas variáveis é do tipo X s).

Analisamos as duas questões abertas ("Quais os pontos da programação despertaram mais seu interesse?" e "Quais aspectos você desejaria aprofundar mais?"), destacando as palavras mais significativas das respostas e essas palavras foram agrupadas por sinonímia. Obtivemos, assim, 6 variáveis para a primeira questão e 4 para a segunda, retomando, as vezes, as alegações que eram comparáveis de uma questão a outra.

Mesmo sabendo que o número de questionários esteja relativamente fraco, algumas estruturas interessantes são interpretáveis sem que seu caráter explicativo seja de uma fidedignidade absoluta, por conseqüência precisa-se ser confirmada por outro estudo. Nos nós contentaremos, então, em destacar as tendências e os pontos mais assegurados.

A tabela do anexo 2 apresenta as ocorrências, as médias e os desvios padrão das variáveis estatísticas estudadas. Observa-se que as variáveis que tiveram maiores ocorrências são: XSS (expectativas satisfeitas), EST (o trabalho foi estimulante) e IMA (manipulação de CHIC). Estas três variáveis expressam o grau de satisfação dos participantes e a pertinência da metodologia de trabalho adotada.

III- Análise de similaridades segundo I.C. Lerman

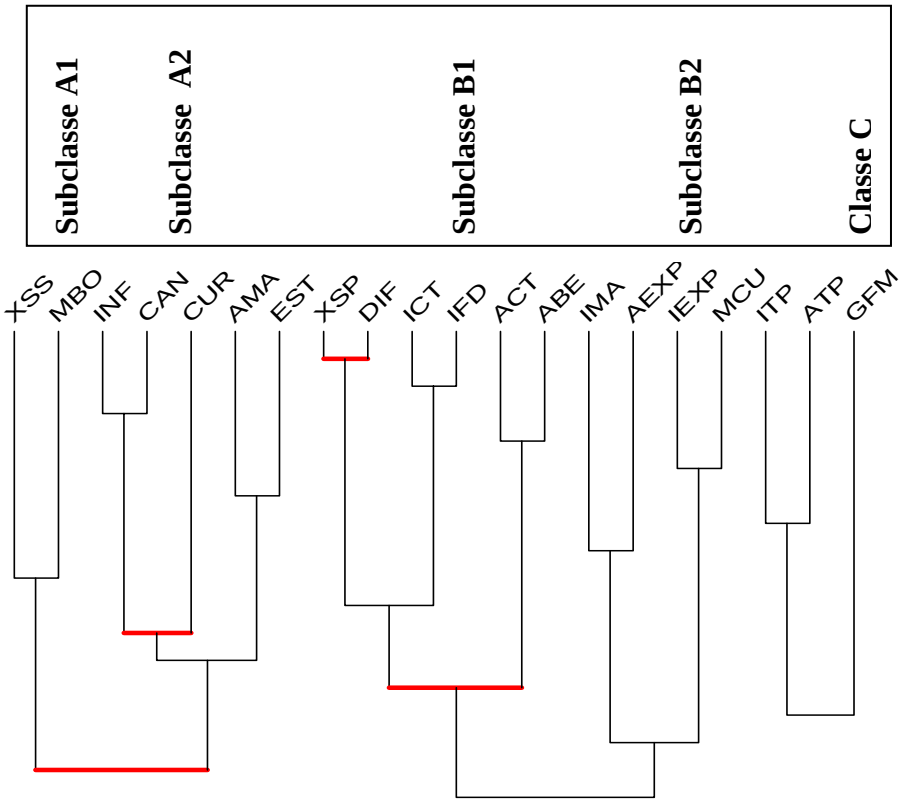
Como em todos os métodos de classificação, procuramos constituir, em um conjunto V de variáveis, partições de V cada vez menos finas, construídas de maneira ascendente. Essas partições encaixadas são representadas por uma árvore construída usando um critério de similaridade ou de semelhança estatística entre variáveis. A similaridade se define a partir do cruzamento do conjunto V das variáveis com um conjunto E de sujeitos (ou de objetos). Este tipo de análise permite ao usuário estudar e interpretar, em termos de tipologia e de

semelhança (e não semelhança) decrescente, classes de variáveis, constituídas significativamente a certos níveis da árvore e se opondo a outras nestes mesmos níveis.

O **critério de similaridade** se exprime da maneira seguinte nos casos das variáveis binárias (presença – ausência, verdadeiro – falsa, sim – não, etc...): 2 variáveis a e b, satisfeitas respectivamente por dois sub-conjuntos A e B de E, são muito semelhantes quando o número k dos sujeitos de $A \wedge B$ é importante de um lado, pelo que teria sido no caso da ausência de ligação entre a e b, e por outro lado, com relação aos cardinais de E, A e B. Medimos esta semelhança pela probabilidade de que k seja superior ao número aleatório esperado nesta situação.

A modelagem probabilista da variável aleatória, cujo k é a realização, pode ser uma distribuição **binomial** ou de **Poisson** à escolha do usuário. A segunda distribuição supõe que E seja uma amostra de uma população grande, o que a primeira não supõe. Se E não tem nenhuma razão estatística a priori de ser representativo, é preferível usar o modelo binomial que analisa a estrutura de E enquanto tal. Quando os parâmetros o permitem, uma aproximação gaussiana destas duas leis é efetuada.

O índice de similaridade entre variáveis é usado para definir um índice de similaridade entre duas classes de variáveis segundo o princípio de comparação entre a observação e o que seria dado pelo acaso. Um índice, dito de coesão, permite reagrupar as classes de variáveis.



Arbre des similarités : C:\chcicsp\CHIC 3.1Sao Paulo\AVALIA2.csv

Assim, para construir uma árvore de similaridade(cf. árvore acima), reunimos em uma classe de primeiro nível, primeiramente, as 2 variáveis que são mais similares no sentido do índice de similaridade, depois 2 outras variáveis ou uma variável e a classe já formada no sentido do índice da classe, e depois outras variáveis ou classes de variáveis.

Interpretação sucinta

No primeiro nível, aparece uma relação de similaridade entre as únicas duas variáveis mais ou menos restritivas: “expectativas parcialmente satisfeitas” e “difícil”. O ritmo do colóquio pode explicar o cansaço, tanto mais que a carga teórica era muito importante. Identificamos aqui, talvez, o desequilibro entre aportes teóricos e ilustrações. Mas a representação deste fenômeno é muito minoritária.

Sublinhamos, por outro lado e a contrário, que a relação entre satisfação (84¨%) e o fato de reconhecer que o colóquio foi satisfatório de modo geral aparece em um outro nível. É este caráter (MBO) que consideramos como indício de satisfação. A variável EF é a mais típica desta relação.

Globalmente, distinguimos na hierarquia, da esquerda para direita, três grandes classes de variáveis: classe A subdividida em A1 e A2, a classe B, subdividida em B1 e B2 e enfim C.

A classe A corresponde a um conjunto de participantes que estão muito satisfeitos por que se investiram muito durante as jornadas, apesar da extensão e da variedade da programação do colóquio Se alguns participantes se dizem cansados, isto parece estar no próprio e total engajamento do sujeito nas atividades. Cansados, certo, mas apreenderam muito e a curiosidade deles foi plenamente satisfeita. A variável suplementar G é típica de A. A1 é a subclasse precedentemente analisada (XSS e MBO)

A subclasse A2 evoca um cansaço (CAN) relacionado com novas aprendizagens e, então a uma curiosidade plenamente preenchida (tipicalidade G).

A classe B contém as nuanças explicando a percepção da dificuldade e, então da satisfação parcial:

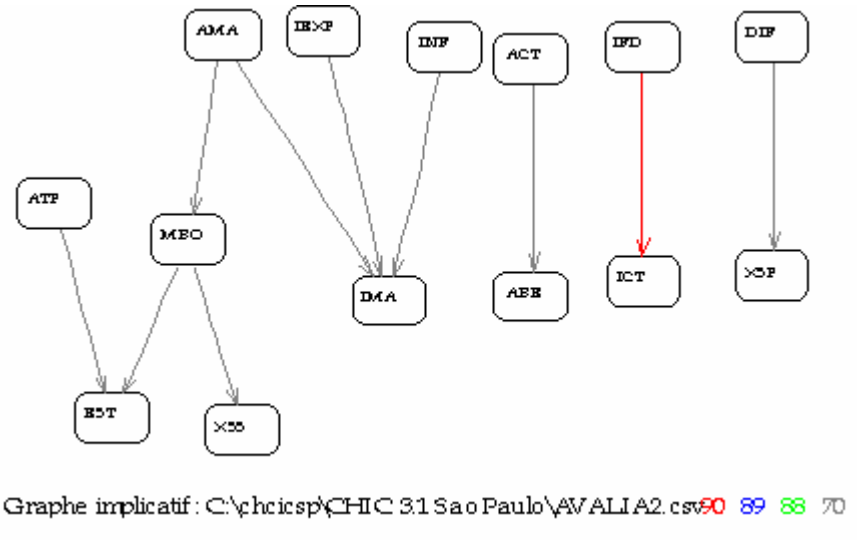
- A subclasse B1 é justamente já foi análisada (XSP e DIF). GME é a variável a mais típica desta subclasse e, globalmente de B.
- A subclasse B2 explica que a origem da restrição vem da complexidade teórica, o que não impede que haja demanda posterior (ACT) de mais atividades sobre o método em razão das possibilidades (ABE) que a teoria permite. A variável M é a mais típica.

A classe C revela a satisfação e uma nova demanda (GFM) de uma programação na qual a teoria e prática (ITP e ATP) seriam intimamente associadas. GMA é a variável a mais típica.

IV- Análise implicativa

Análise do grafo implicativo obtido

Consideramos a construção do grafo implicativo no valor mínimo de 0,70 que permite apresentar uma estrutura interessante e ainda significativa do ponto de vista estatístico. É possível diminuir este valor a fim de obter mais relações implicativas, relações estas se tornando ainda mais fracas. Mas pode-se igualmente aumentar esse número para conservar as implicações mais fortes.



Interpretação sucinta

Duas sub-estruturas aparecem, sem relação entre elas.

1. A subclasse S1 corresponde à satisfação total (XSS), a subclasse S2 à satisfação parcial (XSP). Estas duas variáveis são colocadas nas extremidades dos caminhos implicativos, significando portanto que elas são mais as conseqüências dos sentimentos expressos.

A subclasse S1 reúne dois caminhos:

O primeiro corresponderia, antes de tudo, a futuras expectativas, o segundo corresponderia a satisfações de expectativas a priori.

A satisfação dessas últimas provém da manipulação de CHIC (IMA) e particularmente da descoberta de suas novas funcionalidades (INF) (EM variável típica), mas igualmente das aplicações e experiências associadas (IEXP). São essas relações que deram sentido à manipulação (IEXP =>IMA) (G é a variável mais típica).

Notamos portanto que é a variável AMA que articula os dois caminhos. Pode-se interpretá-la de modo contraposto assim: “se não estivesse satisfeito pelas manipulações, então não desejaria ter outras na ocasião de um próximo colóquio”.
O segundo mostra que a satisfação total (MBO) está muito ligada ao caráter estimulante das atividades (EST).
Observamos que são as variáveis suplementares EF (Ensino Fundamental) e G que parecem ser mais responsáveis desta estrutura S1.

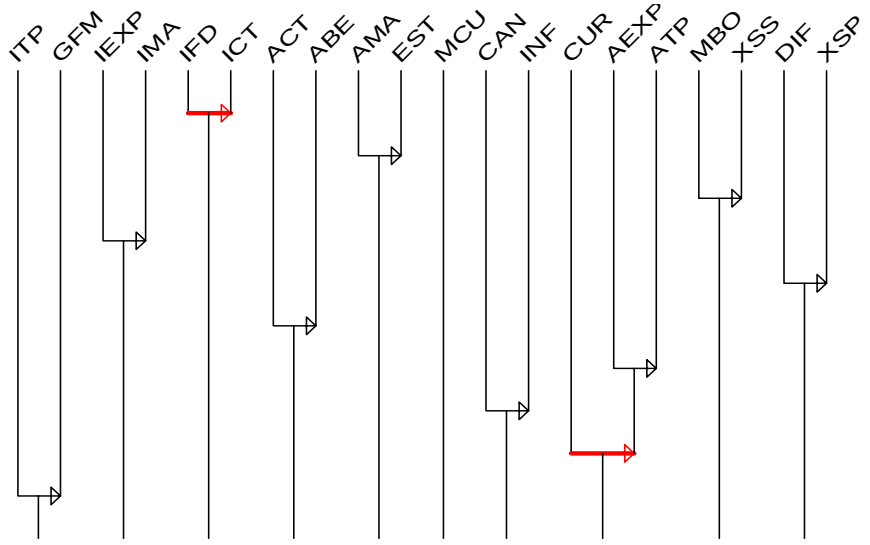
A subclasse S_2 é constituída de três implicações simples:

- ACT => ABF : que significa que a abertura do método implicativo está muito condicionada à teoria (tipicalidade dos GDO, doutores)
- IFD => ICT : que significa a importância atribuída à teoria teria vindo da forma dinâmica de sua apresentação (tipicalidade dos professores do ensino superior).
- DIF => XSP : nos já analisamos esta relação implicativa, mas a análise implicativa nós mostra em qual sentido devemos ler: É por que é difícil que a satisfação apareceu mais fraca (tipicalidade GME ou mestre). Uma interpretação psicológica poderia ser tentada aqui: os participantes que começaram a se engajar na pesquisa parecem ter dificuldade no caminho a percorrer; pois eles devem perceber os aportes teóricos como uma fronteira a ultrapassar para acessar ao estatuto de pesquisador. Isto não é o caso dos mais jovens, não ainda implicados no uso do método, nem os mais avançados ou já iniciados.

2º Hierarquia coesitiva observada

Observa-se que poucas grandes classes se formaram. Todas as classes, exceto uma, são constituídas de dois elementos. Este fato é devido muito ao número relativamente importante de variáveis em relação ao número de sujeitos. Analisaremos, então, brevemente cada uma dessas classes respeitando a sua ordem de qualidade de coesão decrescente.

- IFD => ICT : a forma dinâmica e eficiente é fonte de interesse no que tange ao aporte teórico. Esta observação enfatiza a prudência com a qual um curso magistral deve ser pensado Um curso não acompanhado de exemplo nos quais os sujeitos encontram um domínio já conhecido e manipulado, se torna desestimulante e sem grande efeito sobre a aprendizagem (tipicalidade de ES).



Arbre cohésitif : C:\chcicsp\CHIC3.1Sao Paulo\AVALIA2.csv

- AMA => EST: as futuras manipulações de CHIC poderiam ser estimulantes(tipicalidade de EM).
- MBO => XSS : a significação vem sobretudo do caráter total, completo (teoria – prática), adequado do colóquio (tipicalidade de EF).
- IEXP => IMA : as manipulações se justificam por que elas foram acompanhadas da experimentação não formal (tipicalidade de G)
- DIF => XSP : já foi analisada
- ACT =>ABE : já foi analisada
- CAN => INF : lemos esta implicação de modo contraposto, pois ela é mais significativa nesse sentido: “Se CHIC não tivesse apresentado novas funcionalidades, então não teríamos terminado tão cansados”. Sabe-se bem que são os desequilíbrios dos conhecimentos que são dolorosamente ressentidos (tipicalidade de GDA).
- CUR =>(AEXP => ATP) : Esta relação é estatisticamente significativa. Assim, a curiosidade será tanto quanto satisfeita que a regra: “as experiências permitem esclarecer a relação teoria-prática” será também satisfeita. O sujeito é curioso de ver como a teoria é utilizada como ferramenta para a prática (tipicalidade de GDA)..
- ITP => GFM : as demandas posteriores são principalmente induzidas pelo interesse em trabalhar mais a relação teoria-prática (tipicalidade de EM).

Observamos que a variável MCU (tempo muito curto) é neutra como se ela fosse dividida entre todas as relações implicativas. Isto é, claro, compatível com o sentimento de satisfação geral e uma demanda de intervenção posterior deste tipo.

Para concluir, incentivamos os leitores a interpretar por si mesmo os gráficos dados esquecendo, se julgaram necessário, as interpretações que propusemos.

Bibliografia

AG ALMOULOUD S., *L'ordinateur, outil d'aide à l'apprentissage de la démonstration et de traitement de données didactiques,* thèse de l'Université de Rennes 1, 2 novembre 1992.

BODIN A., "Modèles sous-jacents à l'analyse implicative et outils complémentaires". *Prépublication IRMAR. n°97-32,* (1997)

BODIN A. et GRAS R., 1999 : "Analyse du préquestionnaire enseignants avant EVAPM-Terminales", *Bulletin n° 425 de l'Association des Professeurs de Mathématiques de l'Enseignement Public, Paris* (1999), 772-786

COUTURIER R. et GRAS R., "Introduction de variables supplémentaires dans une hiérarchie de classes et application à CHIC", *Actes des 7èmes Rencontres de la Société Francophone de Classification, Nancy,* (15-17 septembre 1999), 87-92

GRAS R., *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques*, Thèse d'Etat, Université de Rennes I, (1979)

GRAS R. et LARHER A., L'implication statistique, une nouvelle méthode d'analyse de données, *Mathématique, Informatique et Sciences Humaines, E.H.E.S.S. Paris, n°120* (1992), 5-31

GRAS R. et RATSIMBA-RAJOHN H. "Analyse non symétrique de données par l'implication statistique". *RAIRO-Recherche Opérationnelle, 30-3, AFCET, Paris*, (1996), 217-232

GRAS R., BRIAND H. et PETER P. "Structuration sets with implication intensity", *Proceedings of the International Conference on Ordinal and Symbolic Data Analysis - OSDA 95, E.Diday, Y.Chevallier, Otto Opitz, Eds., Springer, Paris* (1996), 147-156

GRAS R. et coll., *L'implication Statistique,* Collection Associée à "Recherches en Didactique des Mathématiques", La Pensée Sauvage, Grenoble, (1996)

GRAS R., BRIAND H., PETER P., PHILIPPE J., 1997 - "Implicative statistical analysis", *Proceedings of International Congress I.F.C.S., 96, Kobé, Springer-Verlag, Tokyo* (1997), 412-419

GRAS R., KUNTZ P., COUTURIER R. et GUILLET F..- « Une version entropique de l'intensité d'implication pour les corpus volumineux ». *Extraction des Connaissances et*

Apprentissage (ECA), vol. 1, n° 1-2, 69-80. Hermès Science Publication, 2001

GRAS R., KUNTZ P. et BRIAND H.: Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données, *Mathématiques et Sciences Humaines*, n° 154-155, p 9-29, ISSN 0987 6936, 2001

GRAS R., DIDAY E., KUNTZ P., COUTURIER R.: Variables sur intervalles et variables-intervalles en analyse statistique implicative, *Actes du 8ème Congrès de la Société Francophone de Classification, Université des Antilles-Guyane, Pointe-à-Pitre, 17-21 décembre 2001, pp 166-173*

GRAS Régis., GUILLET Fabrice, GRAS Robin et PHILIPPE Jacques Réduction des colonnes d'un tableau de données par quasi-équivalence entre variables*, Extraction des connaissances et apprentissage, Hermès, Volume 1, n°4/2001, p 197-202, ISBN 2-7462-0406-1, 2002*

GRAS R., KUNTZ P., BRIAND H.: Hiérarchie orientée de règles généralisées en analyse implicative, *Extraction des Connaissances et apprentissage, Hermès, p 145-157, ISSN 0992-499X, ISBN 2-7462-0631-5, 2003.*

LERMAN I.C., GRAS R. et ROSTAM H., (1981) : Elaboration et évaluation d'un indice d'implication pour des données binaires, I et II, Mathématiques et sciences Humaines, n°75, Paris.

Anexo 1

1ª Parte: Identificação(variáveis suplementares)

Data: _____

1) Sexo: ( ) Masculino **M s** ( ) Feminino **F s**

2) Em que grau(s) de ensino leciona neste ano?

( ) Ensino Fundamental **EF s**

( ) Ensino Médio **EM** s

( ) Ensino Superior **ES s**

( ) Pós-Graduação **PG s**

3) Qual a sua formação acadêmica?

( ) Graduado **G s**

( ) Pós-Graduado – Mestrando **GMA s**

( ) Pós-Graduado – Mestre **GME s**

( ) Pós-Graduado – Doutorando **GDA s**

( ) Pós-Graduado – Doutor **GDO s**

2ª Parte: Opinião (varáveis principais)

4) Suas expectativas foram satisfeitas:

( ) Sim **XSS** ( ) Parcialmente **XSP** ( ) Não **XSN**

5) Quais os pontos da programação despertaram mais seu interesse?

ICT (teoria, conceitual), INF (novas funcionalidades de CHIC), IMA (manipulação de CHIC), ITP (relação teoria-prática), IEXP (aplicações, experiências), IFD (forma dinâmica eficaz)

6) Quais aspectos você desejaria aprofundar mais?

ACT (teoria, conceitual), AMA (manipulação de CHIC), ATP (relação teoria-prática), AEXP (aplicações, experiências)

7) Na lista abaixo, escolhe a(s) expressão(ões) que melhor expressa(m) seus sentimentos a propósito do colóquio.

( ) muito longo **MLO** ( ) muito curto **MCU** ( ) Cansativo **CAN**

( ) Estimulante **EST** ( ) Curiosidade **CUR** ( ) muito bom **MBO**

( ) Inútil **INU** ( ) Abertura **ABE** ( ) difícil **DIF**

( ) gostaria fazer mais **GFM**

Anexo 2

	Ocorrência	Média	Desvio padrão
XSS	21.00	0.84	0.37

XSP	: 4.00	0.16	0.37
ICT	: 10.00	0.40	0.49
INF	: 5.00	0.20	0.40
IMA	: 17.00	0.68	0.47
ITP	: 3.00	0.12	0.32
IEXP	: 11.00	0.44	0.50
IFD	: 4.00	0.16	0.37
ACT	: 9.00	0.36	0.48
AMA	: 8.00	0.32	0.47
ATP	: 16.00	0.64	0.48
AEXP	: 3.00	0.12	0.32
MLO	: 0.00	0.00	0.00
MCU	: 4.00	0.16	0.37
CAN	: 1.00	0.04	0.20
EST	: 18.00	0.72	0.45
CUR	: 11.00	0.44	0.50
MBO	: 13.00	0.52	0.50
INU	: 0.00	0.00	0.00
ABE	: 10.00	0.40	0.49
DIF	: 2.00	0.08	0.27
GFM	: 19.00	0.76	0.43
M s	: 6.00	0.24	0.43
F s	: 19.00	0.76	0.43
EF s	: 4.00	0.16	0.37
EM s	: 10.00	0.40	0.49
ES s	: 15.00	0.60	0.49
PG s	: 4.00	0.16	0.37
G s	: 4.00	0.16	0.37
GMA s	: 9.00	0.36	0.48
GME s	: 4.00	0.16	0.37
GDA s	: 8.00	0.32	0.47
GDO s	: 1.00	0.04	0.20