Análise Experimental de Abordagens de Preservação de Privacidade para Testes Qui-quadrado em GWAS

Antonio A. Marreiras Neto[1], **Manuel Edvar B. Filho**[1], **Javam C. Machado**[1]

[1]LSBD – Departamento de Computação – Universidade Federal do Ceará (UFC)

{antonio.marreiras, edvar.filho, javam.machado}@lsbd.ufc.br

Abstract. *Demand for genetic sequencing studies has resulted in the increasing production and gathering of genome data from the general population. Said data has a highly sensitive nature, as it is possible to deduct not only information about the individual themselves, but also their close relatives. Therefore there exists a necessity for methods to preserve the privacy of participants without major utility losses. In this paper we conduct an experimental analysis evaluating four different approaches to releasing the chi-squared statistic for contingency tables ensuring privacy and utility in order to determine which differentially private model best preserves the utility of the data.*

Resumo. *Demanda por estudos de sequenciamento genético tem resultado na crescente produção e coleta de dados genômicos da população geral. Tais dados possuem um caráter altamente sensível pois, a partir destes é possível deduzir informações não apenas sobre o indivíduo em si, mas também seus familiares. Portanto, existe a necessidade por métodos de preservar a privacidade de participantes sem grandes perdas de utilidade. Neste artigo conduzimos uma análise experimental avaliando quatro diferentes abordagens para distribuir a estatística resultante do teste qui-quadrado para tabelas de contingência garantindo privacidade e utilidade com o intuito de determinar qual modelo diferencialmente privado melhor preserva a utilidade dos dados.*

1. Introdução

O avanço de tecnologias de sequenciamento genético vem alimentando uma crescente demanda por estudos genômicos. Estes estudos têm o recorrente objetivo de identificar correlações entre alterações incomuns no genoma, geralmente polimorfismos de nucleotídeo único (em inglês *single nucleotide polymorphisms*, SNPs), e a relação destas com a ocorrência de doenças aplicando técnicas de análise estatística em extensos conjuntos de dados com amostras de diversos indivíduos. Dados genômicos no entanto tem um caráter especialmente sensível, [Homer et al. 2008] informa que estando um indivíduo presente no conjunto de dados de um estudo entre SNPs, com participantes divididos em grupos de controles e casos, um atacante tendo conhecimento de SNPs referentes ao indivíduo e das frequências agregadas de alelos do grupo é capaz de identificar estatisticamente se o indivíduo em questão está no grupo de casos.

Além disso, [Wang et al. 2009] demonstram como identificar indivíduos a partir de algumas poucas estatísticas resultantes de um estudo, por meio de estatísticas de correlação entre SNPs. Por fim, devido a natureza hereditária dos dados genômicos, um ataque bem sucedido pode comprometer também a privacidade da família do indivíduo

atacado [Wang et al. 2017]. Em resposta a estes estudos, institutos como o *National Institutes of Health* (NIH) e o *EMBL's European Bioinformatics Institute* (EMBL-EBI) limitam o acesso a dados genômicos brutos, e divulgam dados estatísticos somente após processados por métodos de preservação de privacidade. Porém, preservar a privacidade dos participantes no conjunto de dados vem ao custo da utilidade dos dados. Portanto, existe a demanda por métodos de garantir a privacidade dos indivíduos participantes com mínima perda de utilidade dos dados resultantes de análises estatísticas. Neste artigo, conduzimos experimentos para avaliar e comparar a utilidade dos resultados do teste qui-quadrado após o processamento dos dados por cada abordagem e determinar quais seriam melhores em preservar a utilidade de acordo com métricas de acurácia.

2. Fundamentação Teórica

O DNA (ácido desoxirribonucleico) é um composto orgânico constituído das bases nitrogenadas. Um gene é uma sequência destas bases com a função de codificar informações que determinam as características físicas de seres vivos. O gene completo de um organismo é seu genoma, que apresenta alelos, que são variações em uma mesma posição do genoma, denominados menores os que ocorrem com menor frequência em uma população. Uma combinação de alelos capaz de afetar uma característica biológica é um genótipo. SNPs (polimorfismos de nucleotídeo único) são a forma mais simples e frequente no genoma humano em que ocorrem os alelos. Sendo a variação em apenas um único par de bases nitrogenadas. Em conjuntos de dados genômicos SNPs são frequentemente representados pelo número de alelos menores 0, 1 e 2 em um par de cromossomos homólogos. Podemos concluir assim que SNPs são informações privadas sensíveis sobre um indivíduo e também familiares, logo está no interesse do participante do conjunto de dados que sua privacidade seja preservada.

Em um estudo de associação de genoma (GWAS) podemos fazer uso da representação típica dos SNPs ou normalizá-los em 0 e 1 de acordo com um modelo genético para construirmos uma tabela de contingência 2x2 ou 3x2 de acordo com o número de ocorrências de genótipos 0 e 1, ou 0, 1 e 2 entre dois grupos, um de caso e outro de controle. A partir desta tabela, podemos extrair o quão correlacionado um SNP está a uma doença aplicando nesta o teste qui-quadrado para detectar a possibilidade de um genótipo ser a causa da condição.

	(0,0)	(0,1)	(1,0)	(1,1)	Total
Casos	$O_{1,1}$	$O_{1,2}$	$O_{1,3}$	$O_{1,4}$	m_1
Controles	$O_{2,1}$	$O_{2,2}$	$O_{2,3}$	$O_{2,4}$	m_2
Todos	s_1	s_2	s_3	s_4	n

A partir desse exemplo de tabela de contingência, seja $E_{i,j}$ o valor esperado de uma célula [i,j] (um cenário onde todos valores das células [i,j] são iguais), podemos calcular χ^2 como

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} V_{i,j}, \tag{1}$$

onde $V_{i,j} = \frac{(E_{i,j} - O_{i,j})^2}{E_{i,j}} \sum_{j=1}^{J} V_{i,j}$, e $E_{i,j} = s_j \cdot \frac{m_i}{n}$. Calculamos o p-valor de χ^2 a partir de sua distribuição, e por fim comparamos o p-valor ao nível de significância da hipótese nula para rejeitar ou não os resultados.

Privacidade diferencial é uma técnica para preservar a privacidade de um indivíduo independente de sua participação ou não em um banco de dados. Por meio de mecanismos diferencialmente privados, que adicionam ruído a uma resposta a consulta em um banco de dados de acordo com sensibilidade global e orçamento determinados, garantindo que a presença ou não de qualquer indivíduo não vai alterar a probabilidade de resposta de uma consulta.

Definição 2.1 (Privacidade Diferencial) *Um mecanismo aleatorizado M é ϵ-diferencialmente privado (PD) se, seja um ϵ um número real positivo, e sejam D_1 e D_2 dois conjuntos de dados quaisquer que diferem no máximo em um elemento, e para qualquer conjunto S de todas as saídas possíveis de um mecanismo M,*

$$Pr[M(D_1) \in S] \leq \exp(\epsilon) \times Pr[M(D_2) \in S] \tag{2}$$

O *budget* ϵ estima o impacto da adição ou remoção de um indivíduo no conjunto de dados. Assim, um valor ϵ pequeno indica que com pouco impacto de qualquer indivíduo na utilidade do conjunto de dados em questão, um mecanismo ϵ-diferencialmente privado M pode introduzir maior ruído para maior privacidade. O Mecanismo de Laplace [Dwork et al. 2006] adiciona ruído gerado a partir da distribuição de Laplace que satisfaça ϵ-DP . Para explicar o funcionamento do Mecanismo de Laplace primeiro introduzimos o conceito de sensibilidade global:

Definição 2.2 (Sensibilidade global de uma função de consulta) *Sejam D_1 e D_2 dois conjuntos de dados quaisquer que diferem no máximo em um elemento. Seja f uma função $f : D \rightarrow \mathbb{R}^d$, onde D é a coleção de conjuntos de dados e d um valor não negativo. Quando f satisfaz*

$$\Delta f = \max_{D_1, D_2} ||f(D_1) - f(D_2)||_1 \tag{3}$$

para todos D_1 e D_2, a sensibilidade global de f é Δf.

Definição 2.3 (Mecanismo de Laplace) *O Mecanismo de Laplace M gera saída $f(D)+Y$ onde Y é uma variável aleatória da distribuição de Laplace amostrada com escala $\frac{\Delta f}{\epsilon}$*

$$Y = Lap(\frac{\Delta f}{\epsilon}) \tag{4}$$

Sendo $Lap(v)$ uma função que retorna variáveis aleatórias independentes de acordo com o parâmetro de escala v.

3. Trabalhos relacionados

Existem diversos trabalhos que propõe modelos capazes de realizar publicações de estatísticas diferencialmente privadas, aqui citamos os relevantes para nossa comparação.

RandChi e RandChiDist [Sei and Ohsuga 2017] pedem a adição de ruído pelo Mecanismo de Laplace sob o valor qui-quadrado de acordo com sensibilidade Δ_R calculada por

$$\Delta_R = \frac{(m_a + m_b)N}{m_a(1 + m_b)},$$

tal que $a = arg_i min\ \ m_i$ e $b = arg_{i \neq a} min\ \ m_i$

sendo esta calculada de acordo com o vetor $m_i(1, ..., I)$ que representa o total de amostras em cada linha *I* da tabela de contingência, necessitando que m_i seja uma informação pública para que as abordagens possam ser aplicadas, o que costuma ser o cenário mais comum em GWAS, ambas possuem também aplicabilidade em tabelas de contingência IxJ de quaisquer dimensões. RandChiDist funciona à semelhança de RandChi porém faz uso de uma tabela modificada da distribuição qui-quadrado para realizar hipótese nula com maior precisão e evitar falsos positivos. Esse passo adicional é referido como teste DP da hipótese nula de acordo com as seguintes equações e algoritmo:

Encontramos t por meio da equação

$$\int_{x=t}^{\infty} g_{v,\Delta,\epsilon}(x) = \alpha,$$

onde

$$g_{v,\Delta,\epsilon}(x) = \mathcal{L}_{\mu,\beta}(x)\mathcal{Z}_v(\mu)d\mu, \ \ \beta = \frac{\Delta}{\epsilon},$$

$$\mathcal{L}_{\mu,\beta}(x) = \begin{cases} \frac{exp(-\frac{x-\mu}{\beta})}{2\beta}, & \text{se } x > 0 \\ \frac{exp(-\frac{\mu-x}{\beta})}{2\beta}, & \text{caso não} \end{cases}$$

$$\mathcal{Z}_v(x) = \begin{cases} \frac{2^{-v/2}exp(-x/2)x^{-1+v/2}}{\Gamma(v/2)}, & \text{se } x > 0 \\ 0, & \text{caso não} \end{cases}$$ e rejeitamos a hipótese nula caso $t > \chi^2 + Lap(\frac{\Delta f}{\epsilon})$.

A abordagem proposta por [Han et al. 2018] adiciona ruído diretamente nos dados brutos do conjunto de dados de acordo com um modelo que as ocorrências dois possíveis genótipos 0 e 1 em uma matriz de dimensões $\frac{N}{2}X2$ onde N é o número de participantes no conjunto, respectivamente dois vetores de casos e controles. Sendo o de casos composto de uma sequência de a zeros casos do genótipo 0, seguido de b números um, de casos do genótipo 1, de um total de $\frac{N}{2}$ elementos. O vetor de controles é similar. Em seguida, após realizar o teste qui-quadrado segundo um limite determinado pelo usuário da abordagem, a tabela de contingência construída a partir da matriz de dados brutos determinamos o resultado esperado do teste, que calculamos assim os possíveis valores de a que preservam o resultado esperado. Aplicamos ruído do Mecanismo de Laplace com escala $\frac{1}{\epsilon}$, normalizando resultados para 0 e 1, encontrando uma matriz de N participantes com um valor de a que determina o restante da tabela para gerar o menor p-valor possível garantindo a utilidade esperada. A partir desses ruídos, calcula-se a métrica (*N3E*) que nos dá a privacidade esperada e assim escolhemos o resultado dentre os *budgets* que nos vai garantir a maior privacidade, assim encontrando um equilíbrio em um jogo não-cooperativo entre privacidade e utilidade, priorizando a maior privacidade possível sem comprometer a utilidade esperada de acordo com o limite de significância usado na abordagem.

[Yamamoto and Shibuya 2021] provam que é possível garantir privacidade diferencial do *p-valor* resultante do teste qui-quadrado no formato $-log_{10}(p - valor)$ e aplicando ruído do Mecanismo de Laplace com escala 2.33 para tabelas $2x2$ e $log_{10}(e).\frac{N}{0.5N+2}$ com N sendo o número de participantes no conjunto para tabelas $3x2$. Se espera tabelas com mesmo número de casos e controles.

4. Experimentos e Resultados

Nesta seção, são apresentados resultados de uma análise experimental entre as abordagens propostas para divulgar o resultado do teste qui-quadrado realizado em GWAS.

Para conduzir nossa análise, fazemos uso de tabelas de contingência com igual número de casos e de controles, necessidade dos métodos HanZiwei e Yamamoto, virtuais porém com base no conjunto de dados *1000 genome project (phase 3)* referente a SNPs do cromossomo 22 em populações reais. Nos nossos experimentos, a divisão entre casos e controles, e o SNP escolhido são ambos dados aleatórios mas escolhidos de acordo com a condição anterior. Assim criamos 200 tabelas para cada budget e checamos se o resultado do teste da hipótese nula realizado sob a estatística processada se trata de um verdadeiro positivo (*true positive*, TP), um falso positivo (*false positive*, FP) ou um falso negativo (*false negative*, FN).

A partir do número de *TP*, *FP* e *FN* obtidos para cada teste em um *budget* calculamos as métricas de precisão, revocação e F1-score (que mede a acurácia geral do teste) de acordo com as fórmulas:precisão $= \frac{TP}{TP+FP}$, revocação $= \frac{TP}{TP+FN}$ e F1-score $= \frac{2TP}{2TP+FP+FN}$ para identificar quanto da utilidade dos dados foi preservada. E por fim construímos um gráfico comparando os resultados que cada método atinge de acordo com uma série de *budgets* em ordem crescente. Comparamos especificamente os métodos aplicáveis a uma tabela 2x2 na figura 1, e em tabelas 3x2 na figura 2.

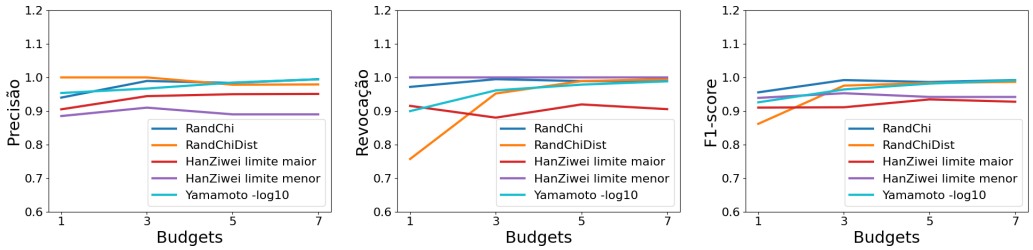

Figura 1. Precisão, revocação e F1-Score para abordagens em tabelas 2x2

Testamos a abordagem HanZiwei em duas formas, com um limite de significância acima do teste de referência, e outro abaixo. No caso em que ambos limites de significância era iguais, HanZiwei sempre entregava resultados esperados, assim optamos por testar estes dois cenários. Notamos que em contraste com a forte acurácia de um limite perfeitamente adequado ao teste, os cenários de HanZiwei testados tiveram o pior desempenho. RandChiDist teve a pior performance em *budgets* menores com alta incidência de falso negativos, logo baixa revocação, mas chegou a se igualar e ultrapassar RandChi com *budgets* maiores. Yamamoto e RandChi ofereceram os resultados mais consistentes para diferentes *budgets*. HanZiwei com limite de significância acima do teste apresentou revocação baixa em todos *budgets* e pior acurácia no geral, resultado de vários falso negativos, e com limite de significância abaixo do teste tendo a segunda pior acurácia e a menor precisão, indicando frequentes falso positivos.

Na figura 2 apresentamos somente os métodos aplicáveis a tabelas 3x2. Aqui o comportamento de RandChi e RandChiDist permaneceu idêntico, com a abordagem Yamamoto tendo performance abaixo daquela no cenário anterior e portanto acurácia um pouco mais baixa que os outros dois métodos escolhidos, resultante de valores de

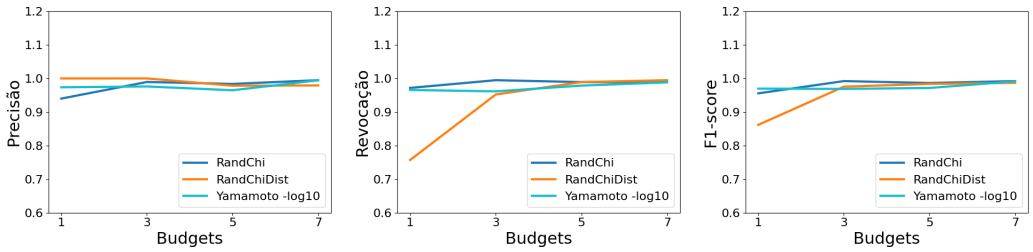

Figura 2. Precisão, revocação e F1-Score para abordagens em tabelas 3x2

revocação menores, indicando maior ocorrência de falsos negativos em comparação a quando usado para tabelas 2x2.

5. Conclusão

Em conclusão, as abordagens RandChi e RandChiDist são as mais versáteis dentre as testadas. Com um mesmo algoritmo aplicável a tabelas de contingência de quaisquer dimensões IxJ seja I e J maiores ou iguais a 2. Yamamoto teve melhores resultados quando aplicado a tabelas 2x2 porém com acurácia não tão distante de RandChi e RandChiDist mesmo em tabelas 3x2, útil para cenários onde $-log_{10}(pvalor)$ possa ser divulgado como alternativa ao valor de χ^2. HanZiwei é a única abordagem que permite divulgar dados brutos de SNPs, porém apesar de perfeita acurácia no cenário que o limite de significância desejado venha a ser idêntico àquele usado pela abordagem, no cenário que estes divergem os resultados tem acurácia baixa comparados aos de outras abordagens, limitando sua aplicabilidade.

Referências

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer.

Han, Z., Liu, H., and Wu, Z. (2018). A differential privacy preserving framework with nash equilibrium in genome-wide association studies. In *2018 International Conference on Networking and Network Applications (NaNA)*, pages 91–96. IEEE.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., and Craig, D. W. (2008). Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8):e1000167.

Sei, Y. and Ohsuga, A. (2017). Privacy-preserving chi-squared testing for genome snp databases. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3884–3889. IEEE.

Wang, M., Ji, Z., Wang, S., Kim, J., Yang, H., Jiang, X., and Ohno-Machado, L. (2017). Mechanisms to protect the privacy of families when using the transmission disequilibrium test in genome-wide association studies. *Bioinformatics*, 33(23):3716–3725.

Wang, R., Li, Y. F., Wang, X., Tang, H., and Zhou, X. (2009). Learning your identity and disease from research papers: information leaks in genome wide association study. In *Proceedings of the 16th ACM conference on Computer and communications security*, pages 534–544.

Yamamoto, A. and Shibuya, T. (2021). More practical differentially private publication of key statistics in gwas. *Bioinformatics Advances*, 1(1):vbab004.