Universidade Federal da Paraíba
Centro de Ciências Exatas e da Natureza
Departamento de Estatística

Lei de Benford e Regras de Associação no Power BI: Ferramentas Estatísticas Aplicadas à Auditoria

Mateus Passador Bittencourt de Sá

Março/2020

Mateus Passador Bittencourt de Sá

Lei de Benford e Regras de Associação no Power BI: Ferramentas Estatísticas Aplicadas à Auditoria

Monografia apresentada ao Curso de Bacharelado em Estatística da Universidade Federal da Paraíba como requisito parcial para obtenção do Grau de Bacharel. Área de Concentração: Estatística Aplicada.

João Pessoa
Março de 2020

Catalogação na publicação
Seção de Catalogação e Classificação

S111l Sá, Mateus Passador Bittencourt de.
Lei de Benford e Regras de Associação no Power BI: Ferramentas Estatísticas Aplicadas à Auditoria / Mateus Passador Bittencourt de Sá. - João Pessoa, 2020.
65 f. : il.

Coorientação: Telmo de Menezes e Silva Filho.
TCC (Especialização) - UFPB/CCEN.

1. Lei de Benford. 2. Mineração de Dados. 3. Power BI. I. Título

UFPB/CCEN

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
COORDENAÇÃO DE ESTATÍSTICA

ATA DE DEFESA DO TRABALHO DE CONCLUSÃO DE CURSO

"Lei de Benford e Regras de Associação no Power BI: Ferramentas Estatísticas Aplicadas á Auditoria"

Mateus Passador Bittencourt de Sá

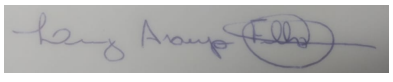
No vigésimo quinto dia do mês de março de 2020 ás 14:30h via internet, a Banca Examinadora do Trabalho de Conclusão de Curso do(a) aluno(a) Mateus Passador Bittencourt de Sá, mat. 11503943, foi composta pelos professores: Dr. Eufrásio de Andrade Lima Neto, Presidente/Orientador(a) (Departamento de Estatística - UFPB), Dr. Marcelo Rodrigo Portela Ferreira, Examinador(a) (Departamento de Estatística – UFPB), Dr. Luiz Medeiros de Araújo Lima Filho, Examinador(a) (Departamento de Estatística – UFPB) e Dr. Ulisses Umbelino dos Anjos, Examinador(a) Suplente (Departamento de Estatística – UFPB). Vale destacar que o professor Dr. Telmo de Menezes e Silva Filho (Departamento de Estatística - UFPB) exerceu o papel de co-orientador do referido trabalho. Dando início aos trabalhos, o presidente da banca cumprimentou os presentes, comunicou aos mesmos a finalidade da reunião e passou a palavra ao candidato para que se fizesse, oralmente, a exposição do trabalho de conclusão de curso intitulado "**Lei de Benford e Regras de Associação no Power BI: Ferramentas Estatísticas Aplicadas á Auditoria**". Concluída a apresentação, o(a) candidato(a) foi arguido(a) pela Banca Examinadora que sugeriu que o(a) aluno(a) fizesse algumas alterações até o dia 1 de abril de 2020. Uma vez entregue a vers ão final do Trabalho de Conclusão de Curso à Coordenação do Bacharelado em Estatística com as alterações solicitadas pela banca examinadora dentro do prazo que o aluno recebeu, o(a) aluno(a) será aprovado com a nota **9,3 (nove vírgula três)**, que é a média aritmética das notas atribuídas pelos membros da Banca Examinadora.

_____ _____

Professor(a) Orientador(a) Dr. Eufrásio de Andrade Lima Neto

_____

Professor(a) Dr. Marcelo Rodrigo Portela Ferreira



_____

Professor(a) Dr. Luiz Medeiros de Araújo Lima Filho

_____

Aluno(a) Mateus Passador Bittencourt de Sá

João Pessoa, 25 de março de 2020.

Este trabalho é dedicado à minha família e amigos, que, certamente, em muito me agregaram ao longo deste curso.

AGRADECIMENTOS

Primeiramente gostaria de agradecer aos meus pais, Gislaine e Flávio, obrigado por serem meus maiores motivadores e patrocinadores. Ninguém conquista nada sozinho, por isso vocês tem minha gratidão por todo o esforço e paciência em prol do meu desenvolvimento intelectual, pessoal e profissional.

Ao meu irmão gêmeo Pedro, obrigado pela parceria e paciência ao longo de 5 anos. Passamos por muitas coisas juntos na vida, e a vinda para João Pessoa talvez tenha sido o nosso maior desafio. Com 17 anos em meio ao medo e desconfiança, se propôs em sair de casa e estudar em uma cidade "nova" ficando a mais de 2.200 quilômetros da família e amigos. Certamente não teria conseguido sem seu suporte. Obrigado, Pedrão.

Também vale o agradecimento a minha irmã Paula e minhas avós Marcia e Maria por me incentivarem incondicionalmente em todas minha decisões. Um salve também para a Rosinha, que durante esses 5 anos sempre me apoiou e motivou para que eu pudesse alcançar esse objetivo.

Meus amigos de curso, Natan, Natália, Ullysses, Fernanda, Giovani, Eduardo, Kelfanio, Rodolfo e Marina, vocês são pra vida toda. Minha gratidão se estende a vocês, que em muito me agregaram ao longo desses anos.

Minha gratidão aos amigos do Hospital por todos os ensinamentos especialmente ao João Paulo, Thaís e o Hérico que nunca se cansaram de instruir e sempre me auxiliaram em cada um dos meus passos. Vale lembrar também do Leandro que me deu oportunidade de ter essa experiência, e não deixou eu desistir. Por fim, meus parceiros Helton, Marroney e Daniel por todos os perrengues que passamos.

Os meus agradecimentos também se estendem aos amigos que fiz do curso de Engenharia de Alimentos, mais especificamente aos membros do grupo "rolezeiros". Muito agradecido por todos os momentos de descontração.

Agradeço aos meus orientadores Eufrásio e Telmo por aceitar conduzir o meu trabalho de conclusão de curso ao longo de dois semestres. Obrigado por toda paciência, atenção e incentivo durante esse período.

Não poderia faltar gratidão aos professores do DE em sua grande maioria por serem sempre atenciosos e dedicados, em especial aos que tive o privilégio de cursar alguma disciplina. Aos professores Neir, Eufrásio e Tarciana um agradecimento especial por me darem a oportunidade de ser seu aluno bolsista PIBIC. Quero enfatizar que apesar das dificuldades cada aprendizado valeu muito a pena. Aos professores Ana Flávia e Hemílio minha gratidão por me auxiliarem e incentivarem ao longo do curso.

Agradeço a banca avaliadora deste trabalho pela disposição em contribuir com esta monografia.

Aos amigos e familiares, sintam-se incluídos nesta seção de agradecimentos. Vocês são responsáveis por partilhar de momentos alegres e tristes, sempre com muito apoio. Sintam-se também homenageados Amanda, Cláudia, Cristina, Renata, Gabriela, Deco, Gisele, Patrick, Beatriz, Arthur, Kaike, Daniel, Larissa, Raquel, Diego.

"A verdadeira jornada de
descoberta não consiste
em procurar novas paisagens,
mas em ter novos olhos."

(Marcel Proust)

RESUMO

A expectativa de vida das população vem aumentando e, consequentemente, a área da saúde acaba passando por grandes transformações. O negócio da saúde envolve muito dinheiro tanto no meio público quanto privado, daí as cobranças por uma boa gestão hospitalar vêm à tona. No Brasil, país que sempre está entre os piores no *ranking* mundial de corrupção, não é fácil gerir um negócio. É nesse contexto que a auditoria ganha força. Com o intuito de maximizar os lucros e valorização da empresa, o auditor tem como principal função garantir a veracidade das informações contábeis e patrimoniais de uma corporação. Como resultado do processo de auditoria, muitas vezes, fraudes são detectadas. São diversas as técnicas utilizadas por auditores para a detecção de fraudes. Neste trabalho, destacamos a Lei de Benford. Essa técnica foi proposta por Benford (1938) e se utiliza das frequências com que dígitos aparecem para garantir a autenticidade das informações. Benford concluiu que existe um padrão nas frequências com que os dígitos aparecem, quando são provenientes de eventos aleatórios. O segundo método utilizado neste trabalho para auxiliar na auditoria é proveniente do campo da mineração de dados: as regras de associação. Esse procedimento busca por comportamentos dentro do banco de dados. A aplicação dessas técnicas em conjunto foi feita em dois bancos de dados reais referentes ao faturamento e contas pagas de um hospital particular da cidade de João Pessoa. Durante a análise, algumas anomalias forma encontradas, todavia, ao fazer uma inspeção mais minuciosa não se encontrou nenhum indício de fraude. Todas as técnicas descritas foram implementadas nos *softwares* Power BI e R. A plataforma de *Business Intelligence* se mostrou como uma ferramenta promissora para esse tipo de análise, fornecendo suporte para tomadas de decisão pelos auditores.

Palavras-chave: Lei de Benford; Mineração de Dados; Power BI.

ABSTRACT

The life expectancy of the population has been increasing and, consequently, the healthcare system ends up undergoing major transformations. The healthcare business involves a lot of money in both the public and private sectors, hence the demand for good hospital management gains importance. In Brazil, a country that is always among the worst in the world ranking of corruption, it is not easy to run a business. It is in this context that audits gain strength. In order to maximize profits, the auditor's main function is to guarantee the accuracy of a corporation's accounting and equity information. As a result of the audit process, frauds are often detected. There are several techniques used by auditors to detect fraud. In this work, we highlight Benford's Law. This technique was propose by Benford (1938) and uses the frequencies on which the digits appear to guarantee the authenticity of the information. Benford concluded that there is a pattern in the frequencies with which digits appear, when they arise from random events. The second method used in this work to assist in auditing comes from the field of data mining: rules of association. This procedure searches for behaviors within the database. The joint application of these techniques was made in two real databases referring to the billing and paid bills of a private hospital in the city of João Pessoa. During the analysis, we found some anomalies, however, upon closer inspection, there was no evidence of fraud. All the techniques were implemented in the Power BI and R project software. The Business Intelligence platform proved to be a promising tool for this type of analysis, providing support for decision making by auditors.

Keywords: Benford's Law; Data Mining; Power BI.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

CAPÍTULO 1

INTRODUÇÃO

1.1 Definição e Motivação

Com os avanços tecnológicos e da medicina, a expectativa de vida da população vem aumentando e consequentemente acarretando grandes mudanças sociais e econômicas. Uma das principais áreas afetadas é da saúde. Por conta disso, cada vez mais o negócio da saúde vem a pauta. Ainda existe uma certa dificuldade de visualizar os hospitais como uma empresa de negócios devido a sua origem (PORTELA; SCHMIDT, 2008). Porém, esse panorama vem sendo alterado devido ao grande fluxo de capital envolvido.

A gestão hospitalar vem se tornando cada vez mais rigorosa para maximizar os resultados. Os maiores cuidados são com os controles dos custos e receitas operacionais. Essas duas vertentes são responsáveis pela manutenção do equilíbrio financeiro de um hospital e tem impacto direto nas margens de lucro. Os custos podem ser tanto diretos quanto indiretos. Os custos diretos estão ligados aos gastos vinculados aos pacientes como: medicamentos, materiais, etc. Já os indiretos, são os custo no qual não tem saída direta para os pacientes, porém são necessários para o funcionamento do hospital, são eles: água, luz, eletricidade, dentre outros. Em contrapartida aos custo temos as receitas operacionais, essas são o resultado de todo recurso proveniente da venda de um produto, no caso do hospital esse produto são procedimentos para a manutenção da saúde.

No Brasil, país que sempre está entre os piores países no *ranking* de corrupção do mundo não se é fácil gerir um negócio. Fala-se muito em corrupção na política, no entanto ela está presente em todas esferas, inclusive no meio hospitalar. A Agência Brasil publicou um artigo feito por Abdala (2020) mostrando uma operação da Polícia Federal

(PF) que investigou compras de medicamentos feita no município de Nova Friburgo, Rio de Janeiro. A PF suspeita que o desvio tenha sido de mais de meio milhão de reais. É nesse contexto que a auditoria ganha força.

A auditoria tem como objetivo averiguar os registros de uma entidade em busca da garantia da veracidade das informações de resultados patrimoniais. Os objetos da auditoria são todos e quaisquer documentos que comprovam esses registros contábeis. Para uma avaliação minuciosa desses documentos o auditor recorre a diversas técnicas. Caso não seja comprovado a autenticidade dessas informações, temos um cenário de possível fraude fiscal. Vale enfatizar que o auditor não tem a obrigação de detectar fraudes, mais sim de apontar possíveis fraudes durante a processo.

Uma das técnicas utilizadas pelos auditores é a Lei de Benford. Essa técnica foi proposta por Benford (1938) e busca mostrar a aleatoriedade de um evento, através da distribuições dos dígitos. Vale ressaltar que, assim como qualquer outra técnica de auditoria, que caso tenhamos alguns indícios de não conformidade das informações, não podemos necessariamente concluir que haja uma fraude, sendo então necessária uma análise mais detalhada. A Lei de Benford também tem a aplicabilidade em outras áreas, além da contabilidade, sendo elas: detecção de fraudes eleitorais, dados de genoma, detecção de fraude em artigos científicos, verificação de provas judiciais, entre outros.

Outro método que pode auxiliar na auditoria segundo Lee e Stolfo (1998) é a técnica de mineração de dados denominada de regras de associação. Esse procedimento faz a busca por itens em um conjunto de dados que impliquem na presença de outros. Os comportamentos obtidos através da aplicação do método são chamados de regras. Muitas dessas regras encontradas não são facilmente perceptíveis, por conta desse fato, essa técnica acaba se tornando uma ferramenta conveniente para os auditores. Silva e Ralha (2011) mostram a efetividade dessa metodologia ao fazer a aplicação dessa técnica para auditar licitações públicas, afim de detectar cartéis.

A aplicação de métodos de detecção de fraude pelos auditores pode ser feita por diversos *softwares*. Entretanto as plataformas de *Business Intelligence* (BI) vem ganhando força no mercado. Com o BI, pode ser feita uma automatização da utilização dessas técnicas de auditoria, além de ser possível por gatilhos e alertas sobre possíveis fraudes. Essa automatização do processo implica em uma melhor disseminação e velocidade de acesso aos usuários. Espera-se que o uso de plataformas de BI auxiliem a tomada de decisão por parte do auditor.

1.2 Objetivos

Gerais

Aplicar a lei de Benford e regras de associação como ferramentas de auxílio à auditoria.

Específicos

- Identificar possíveis atividades fraudulentas a partir dos resultados de medidas e testes de conformidade;
- Encontrar regras que possam implicar na não conformidade coma lei de Benford;
- Implementar as técnicas da lei de Benford e regras de associação no *software* Power BI;

1.3 Organização do Trabalho

Além do capítulo de introdução, este trabalho é composto por mais outros três. No segundo capítulo é apresentada uma revisão geral sobre as técnicas utilizadas, assim como é feita toda contextualização do tema. O Capítulo 3 é onde serão apresentados todas as aplicações feitas. E por fim, o Capítulo 4, no qual são apresentadas as considerações finais acerca do trabalho e algumas sugestões para trabalhos futuros.

CAPÍTULO 2

REFERENCIAL TEÓRICO

2.1 Lei de Benford

Nesta seção, apresentaremos uma visão geral sobre a Lei de Benford, introduzindo sua história, representações e propriedades matemáticas, além de conceitos básicos para compreensão das aplicações que serão feitas. Conjuntamente serão introduzidas medidas de perfomance que serão usados para avaliação do método.

2.1.1 História

O fenômeno conhecido por Lei de Benford foi primeiramente idealizado por Simon Newcomb, um astrônomo altamente honrado em sua época. Segundo Jamain (2001), Newcomb observou através das páginas de uma tabela logarítmica que, as primeiras páginas apresentavam mais marcas de uso que as últimas. A partir daí surgiu a ideia de que as frequências dos dígitos não eram iguais. Mais que isso, que os dígitos iniciais eram mais comuns do que os finais. Está concepção pode causar um certo estranhamento visto que, intuitivamente, acredita-se que os dígitos de eventos aleatórios sejam uniformemente distribuídos.

Newcomb (1881) publicou 2 páginas no Jornal Americano de Matemática a respeito de sua descoberta. Todavia, não deixou formalizado matematicamente as equações das frequências esperadas para cada dígito que, futuramente, viriam a ser apresentadas por Benford. Porém, certamente, Newcomb estava ciente das mesmas, em razão de em seu

artigo ter apresentado uma tabela com as frequências relativas esperadas para o primeiro e o segundo dígito. Também foi Newcomb quem propôs a convicção de que a medida que se deslocam os dígitos para a esquerda, a distribuição do dígito vai convergindo para uniforme, tornando-se uma das propriedades de maior relevância da Lei de Benford. Por conta de todos esses fatores muitas pessoas se referem a Lei de Benford como Lei de Newcomb-Benford.

Assim como Newcomb, o físico Frank Benford fez a mesma observação sobre as marcas de usos das páginas da tabela logarítmica (NIGRINI, 2012). O mesmo realizou um experimento no qual coletou 20.299 observações de 20 diferentes eventos "naturais"e computou a frequência dos dígitos. Os dados contemplam informações de áreas de rios, constantes físicas e matemáticas, pesos atômicos, *etc.* O resultado desse estudo é apresentado pela Tabela 2.1.

Tabela 2.1: Experimento de Benford

			First Digit								
Group	Description	Count	1	2	3	4	5	6	7	8	9
A	Rivers, Area	335	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1
B	Population	3259	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2
C	Constants	104	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6
D	Newspapers	100	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0
E	Spec.Heat	1389	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1
F	Pressure	703	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7
G	H.P.Lost	690	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6
H	Mol.Wgt.	1800	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2
I	Drainage	159	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9
J	AtomicWgt.	91	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5
K	$\frac{1}{n}, \sqrt{n}$	5000	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9
L	Design	560	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6
M	Digest	308	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2
N	CostData	741	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1
O	X-RayVolts	707	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8
P	Am.League	1458	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0
Q	BlackBody	1165	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4
R	Addresses	312	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0
S	$n_1, n_2, \ldots, n!$	900	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5
T	DeathRate	418	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1
	Average	1011	30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7
	ProbableError	-	±0.8	±0.4	±0.4	±0.3	±0.2	±0.2	±0.2	±0.2	±0.3

Benford fez a média das proporções para cada dígito do conjunto de dados. Ele observou que 30,6 por cento dos números tinham um primeiro dígito 1 e 18,5 por cento dos números tinham um primeiro dígito 2. A medida que os dígitos iam aumentando menores eram suas frequências relativas.

A partir desse estudo, Benford (1938) publicou o artigo *The Law of Anomalous Numbers* que é o grande responsável pelo que conhecemos por lei de Benford. A partir dele, outros estudos foram desenvolvidos fazendo surgir abundantes aplicações e propriedades.

Segundo Nigrini (2012) para uma boa aplicação da lei de Benford é necessário que os dados sejam provenientes de eventos aleatórios ou registros como por exemplo o tamanho de uma população por município ou dados dos níveis de transações como contas pagas e contas à pagar. Em contrapartida a lei não tem aplicabilidade em números não aleatórios como é o caso do registros geral no qual os números tem relação com a localidade e com a ordem que foi tirado esse documento, e também não é indicado para dados com limites máximos ou mínimos incorporados.

2.1.2 Distribuição dos dígitos

Considere uma amostra aleatória $x_1, x_2, ..., x_n$ de uma variável aleatória X. Defina D_{ki} como sendo a variável aleatória relacionada ao k-ésimo digito da i-ésima observação da amostra aleatória, onde $i = 1, ..., n$ e $D_{ki} \in \{1, 2, ..., \infty\} \forall k, i$. Adicionalmente, defina $D_k D_{k+1}$ como sendo a variável aleatória relacionada a dois dígitos consecutivos. Nesse sentido, D_1 é a variável aleatória relacionada com o primeiro dígito, D_2 é a variável aleatória relacionada com o segundo dígito e $D_1 D_2$ é a variável aleatória relacionada com os dois primeiros dígitos de um número.

Distribuição do primeiro dígito

A função de probabilidade proposta por Benford para o primeiro dígito é dada pela seguinte expressão:

$$\text{Prob}(D_1 = d_1) = log\left(1 + \frac{1}{d_1}\right); d_1 \in \{1, 2, \ldots, 9\} \tag{2.1}$$

Por exemplo, a probabilidade do primeiro dígito ser 1 é dada por:

$$\text{Prob}(D_1 = 1) = log\left(1 + \frac{1}{1}\right) = log(2) = 0,30103$$

Tabela 2.2: Função de Probabilidade para D_1

Dígito	Probabilidade
1	0,30103
2	0,17609
3	0,12494
4	0,09691
5	0,07918
6	0,06695
7	0,05799
8	0,05115
9	0,05115

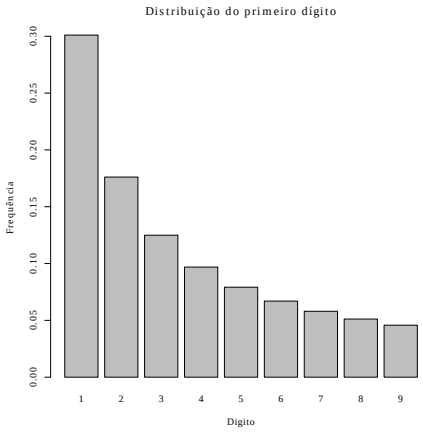


Figura 2.1: Distribuição do primeiro dígito

Distribuição do segundo dígito

Segundo Benford (1938), a função de probabilidade para o segundo dígito e dada por:

$$\text{Prob}(D_2 = d_2) = \sum_{d_1=1}^{9} log\left(1 + \frac{1}{d_1 d_2}\right); d_2 \in \{0, 1, \ldots, 9\} \tag{2.2}$$

Logo, de acordo com a expressão 2.2, a probabilidade do segundo dígito ser igual a 1 é dada por:

$$\begin{aligned}\text{Prob}(D_2 = 1) = \sum_{d_1=1}^{9} log\left(1 + \frac{1}{d_1 d_2}\right) &= log\left(1 + \frac{1}{11}\right) \\ +log\left(1 + \frac{1}{21}\right) + log\left(1 + \frac{1}{31}\right) &+ log\left(1 + \frac{1}{41}\right) \\ +log\left(1 + \frac{1}{51}\right) + log\left(1 + \frac{1}{61}\right) &+ log\left(1 + \frac{1}{71}\right) \\ +log\left(1 + \frac{1}{81}\right) &+ log\left(1 + \frac{1}{91}\right) \\ &= 0,11389\end{aligned}$$

Tabela 2.3: Função de Probabilidade para D_2.

Dígito	Probabilidade
0	0,11968
1	0,11389
2	0,10882
3	0,10433
4	0,10031
5	0,09668
6	0,09337
7	0,09035
8	0,08757
9	0,08500

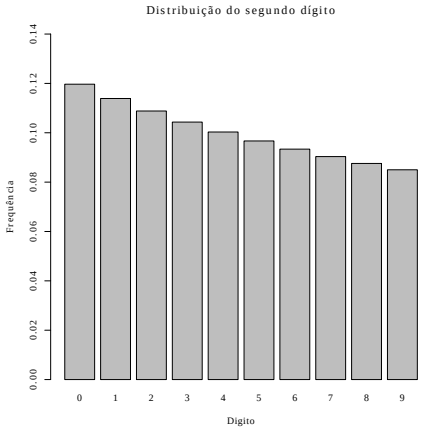


Figura 2.2: Distribuição do segundo dígito.

Distribuição dos dois primeiros dígitos

A função de probabilidade para os dois primeiros dígitos é dada por:

$$\text{Prob}(D_1D_2 = d_1d_2) = log\left(1 + \frac{1}{d_1d_2}\right); d_1d_2 \in \{10, 11, \ldots, 99\} \tag{2.3}$$

As probabilidades dos dois primeiros dígitos não são independentes, logo, vale salientar que $\text{P}(D_1D_2 = 11)$ é diferente de $\text{P}(D_1 = 1) \times \text{P}(D_2 = 1)$. Temos então, pela expressão (2.3) que a probabilidade dos dois primeiros dígitos serem iguais a 11 é:

$$\text{Prob}(D_1D_2 = 11) = log\left(1 + \frac{1}{11}\right) = log\left(\frac{12}{11}\right) = 0,03779$$

Tabela 2.4: Função de Probabilidade para D_1D_2.

Dígitos	Probabilidade
10	0,0414
11	0,0378
12	0,0348
13	0,0322
14	0,0300
15	0,0280
⋮	⋮
96	0,0045
97	0,0045
98	0,0044
99	0,0044

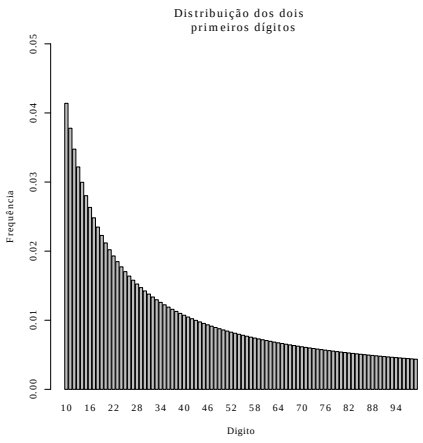


Figura 2.3: Distribuição dos dois primeiros dígitos.

2.1.3 Teoremas e Propriedades

A partir do artigo publicado por Benford (1938), outros trabalhos foram desenvolvidos e propriedades importantes foram descobertas. Nessa sub-seção serão apresentadas os mais importantes teoremas propostos. Uma boa parte desses teoremas demonstram que, mesmo com transformações nos dados, os mesmos tendem a seguir conformidade ou não com a Lei de Benford, de acordo com seu estado original. Vale destacar que essa propriedade somente ocorre em para alguns tipos de transformações, todavia, esses casos já são suficientes para provar a robustez do método em relação a tentativa de manipular dados para omitir fraudes.

Escala Invariante

Se os números $x_1, x_2, \ldots, x_N$ estão conformes à lei de Benford, qualquer nova variável formada pelo produto de x_i por uma constante c, diferente de zero, também estará conforme à lei de Benford (PINKHAM, 1961).

Além de ser útil contra a tentativa de se encobrir fraudes, esse teorema é útil no tratamento dos dados para a aplicação da Lei de Benford. Por exemplo, quando se trabalha com números decimais muitos pequenos, a virgula pode ser um problema para a extração dos dígitos, dependendo do software que se trabalha. Portanto, uma das possíveis soluções é a multiplicação dos dados originais por uma constante. O mesmo ocorre quando se trabalha com números negativos.

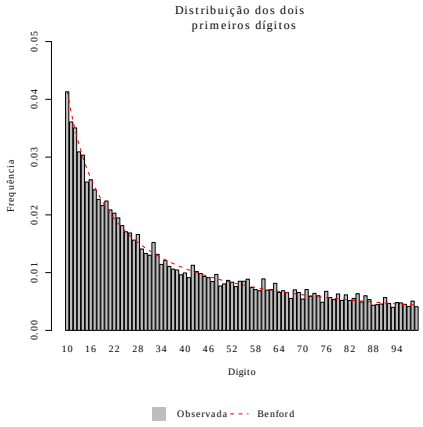


Figura 2.4: Distribuição dos dois primeiros dígitos do número de habitantes.

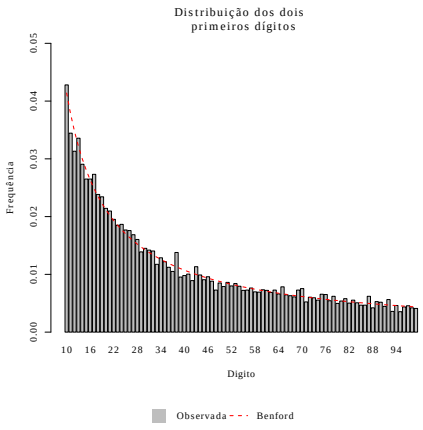


Figura 2.5: Distribuição dos dois primeiros dígitos do produto do número de habitantes por π.

Utilizando a variável número de habitantes por municípios provenientes do Censo populacional de 2009 dos Estados Unidos, disponível em Nigrini (2012), fizemos o produto dessa variável pela constante π. As Figuras 2.4 e 2.5 apresentam a distribuição dos dois primeiros dígitos para a variável original e para seu produto com a constante. Note que o cenário de conformidade de Benford é mantido, mesmo com a multiplicação pela constante, satisfazendo assim o teorema de escala invariante proposta por Pinkham (1961).

Produto de Uniformes

Ao fazer produto de variáveis aleatórias uniformemente distribuídas, quanto maior o número de multiplicações, maior será a conformidade com a lei de Benford (ADHIKARI; SARKAR, 1968).

Para ilustrar este resultado, gerou-se 5 amostras pseudo-aleatórias com distribuição uniforme no intervalo 0 a 1, cada uma contendo 25.000 observações. Fez-se então o produto delas, uma a uma. Veja, na Figura 2.6, que a medida que é adicionada uma dessas variáveis uniformes no produto, maior será a conformidade com a distribuição de Benford.

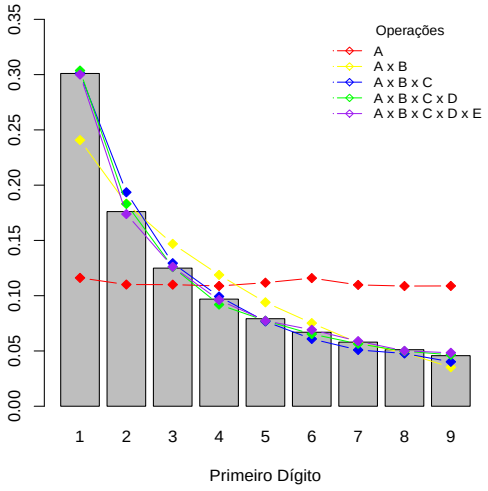


Figura 2.6: Produtos de Uniformes.

Multiplicação por $1/X$

Assim como no teorema anterior, este também foi enunciado por Adhikari e Sarkar (1968). Caso um conjunto de números $x_1, x_2, \ldots, x_N$ estão conformes a lei de Benford,

então, um conjunto de números formados por $1/X$ ou c/X onde $c > 0$ também estará conforme.

Utilizando variável do número de habitantes por município da base de dados do Censo populacional norte americano de 2009, fizemos a razão de um sobre a quantidade de habitantes por municípios. O resultado dessa transformação é dado na Figura (2.7). A manutenção da conformidade é mantida assim como o teorema previa.

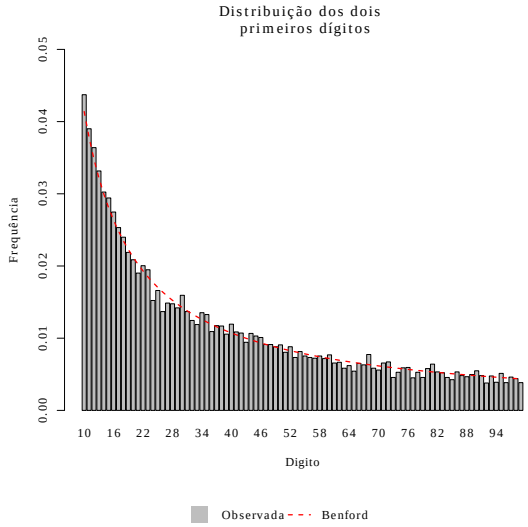


Figura 2.7: Transformação $1/X$ do Censo 2009.

O produtos de variáveis

Considere $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_k, (k = 2, \ldots, p)$ um conjunto de k amostras aleatórias de tamanho n provenientes de k distribuições de probabilidade independentes, onde $\boldsymbol{X}_j = [x_{1j}, x_{2j}, \ldots, x_{nj}]$, $j = 1, \ldots, k$. Seja $\boldsymbol{Y}_j = \prod_{j=1}^{k} \boldsymbol{X}_j$ uma V.A., ou seja, o produto de k variáveis aleatórias anteriores. A distribuição dos dígitos da variável aleatória $\mathbf{Y}$ seguirá a Lei de Benford.

Tomando como base a Tabela 2.5, geramos 25.000 observações de 6 distribuições de probabilidade diferentes. Fazendo o produto de todas eles temos, como resultado, a Figura 2.8. Temos um cenário de conformidade com a lei de Benford, ou seja, o teorema descrito acima é satisfeito.

Tabela 2.5: Números pseudo aleatórios

Distribuição	Parâmetros	Amostra
Normal	(5,5 ; 1)	25.000
Uniforme	(1 ; 10)	25.000
Poisson	(7,5)	25.000
Beta	(2 ; 4)	25.000
Gama	(2,5 ; 5)	25.000
Weibull	(5 ; 1)	25.000

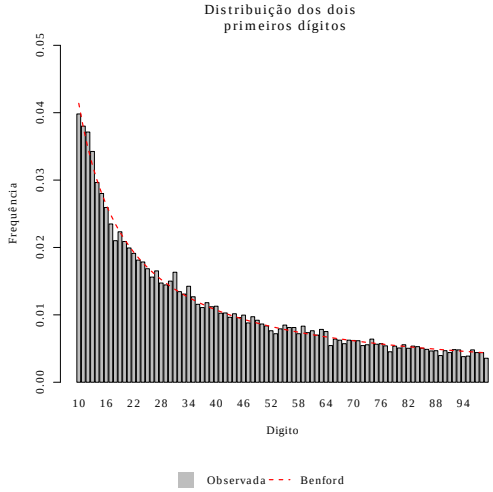


Figura 2.8: Produto das 6 distribuições

2.1.4 Medidas e Testes de Conformidade para Variáveis Aleatórias, segundo a Lei de Benford

Para as medidas e testes dessa Sub-seção considere F_O e F_E como sendo respectivamente as frequências absolutas observadas e esperadas. O mesmo comportamento segue para P_O e P_E referente as proporções observadas e esperadas pela Lei de Benford. Suponha ainda N como sendo o número de registros e K a quantidade de dígitos.

Teste Z

A estatística Z é usada para testar se existe diferença entre a proporção atual do dígito e a proporção esperada por Benford a partir de um nível de confiança adotado. Esse teste é realizado para cada um dos dígitos individualmente, permitindo com que inferências sejam feitas para o mesmo. Por não se tratar de um teste generalista, a não conformidade para um dos dígitos não implica necessariamente no mesmo resultado para todo o conjunto de dados. A Estatística do Teste é dada por:

$$Z = \frac{|PO_i - PE_i| - \left(\frac{1}{2N}\right)}{\sqrt{\frac{PE_i(1-PE_i)}{N}}} \tag{2.4}$$

O termo $1/(2N)$ é um termo de correção de continuidade, o mesmo somente é utilizado quando for menor que o primeiro termo do numerador ($|PO_i - PE_i|$).

A hipótese nula para esse teste designa que não há diferença estatística entre as proporções observadas e esperadas . Consequentemente, a hipótese alternativa sugere que existe diferença estatística entre os valores observados e esperados. Ao adotar um nível de significância de 5% e sabendo que pelo Teorema Central do Limite, Z terá distribuição aproximadamente normal padrão, o valor crítico desse teste é igual à 1,96, ou seja, para valores superiores a 1,96 a hipótese nula é rejeitada.

Teste Qui-Quadrado (χ^2)

O teste Qui-Quadrado de aderência é usado para verificar a adequabilidade de um conjunto de dados observados a um modelo probabilístico. A hipótese nula testada é de que as proporções observadas e esperadas são iguais para cada um dos dígitos. Uma vez que a hipótese nula não é rejeitada, temos evidências estatísticas suficientes para dizer que os dados estão conformes a Lei de Benford. A estatística é dada por:

$$\chi^2 = \sum_{i=1}^{K} \frac{(FO_i - FE_i)^2}{FE_i} \tag{2.5}$$

Os valores críticos do teste Qui-Quadrado vão variar de acordo com a distribuição de Benford testada. Uma vez que χ^2 tem distribuição Qui-Quadrado com $k-1$ graus de liberdade e considerando 5% de significância temos que, para o teste primeiro dígito o valor crítico será $\chi^2(8) = 15,51$. Para os testes do segundo e dois primeiros dígitos esses valores críticos serão respectivamente $\chi^2(9) = 16,92$ e $\chi^2(89) = 112,02$. Em casos que $\chi^2 > \chi^2_{k-1}$ deve-se rejeitar a hipótese nula, ou seja, os dados não estão conformes a lei de Benford.

Desvio Médio Absoluto (DMA)

A medida DMA (Desvio Médio Absoluto) para a lei de Benford é a média das distâncias entre a atual proporção dos dígitos e a proporção esperada. Ela nos dá uma

noção da variabilidade em um conjunto de dados, portanto, serve como uma medida de conformidade.

O calculo do DMA é dado pela expressão:

$$DMA = \sum_{i=1}^{K} \frac{|PO_i - PE_i|}{K} \tag{2.6}$$

A conclusão sobre conformidade ou não dos dados com a distribuição de Benford pela medida DMA, se dá pela comparação entre resultado obtido na medida com os valores da Tabela 2.6.

A Tabela 2.6 é o resultado do estudo do Drake e Nigrini (2000), no qual ele mapeia as regiões de conformidades e os valores críticos do teste DMA através de um estudo de simulações. Note que existe variação dos intervalos de conformidade dependendo da distribuição dos dígitos que será avaliada.

Tabela 2.6: Valores críticos e regiões de conformidade

Dígito	Intervalo	Conclusão
Primeiro Dígito	0,000 - 0,006	Conformidade
	0,006 - 0,012	Conformidade Aceitável
	0,012 - 0,015	Conformidade Marginalmente Aceitável
	acima de 0,015	Não conforme
Segundo Dígito	0,000 - 0,008	Conformidade
	0,008 - 0,010	Conformidade Aceitável
	0,010 - 0,012	Conformidade Marginalmente Aceitável
	acima de 0,012	Não conforme
Dois Primeiros Dígitos	0,0000 - 0,0012	Conformidade
	0,0012 - 0,0018	Conformidade Aceitável
	0,0018 - 0,0022	Conformidade Marginalmente Aceitável
	acima de 0,0022	Não conforme

2.2 Regras de Associação

As regras de associação ou *market basket analysis* (análise de cesta de mercado) é uma técnica de mineração de dados cujo principal objetivo é encontrar comportamentos dentro de um conjunto de dados e, a partir desses comportamentos, fazer inferências probabilísticas a respeito da ocorrência de eventos.

Levando a ideia de regras de associação para o panorama de uma compra no supermercado, cujo aplicação foi a propulsora dessa técnica, temos que, a partir de um conjunto de produtos que são comprados simultaneamente, pode se fazer inferências a cerca de outros produtos que também estarão presentes em suas compras. A partir de tais regras, por exemplo, se torna possível trabalhar estratégias de vendas como, por exemplo, colocar produtos com alta associação próximos para induzir clientes comprarem conjuntamente.

As possibilidades das regras são muitas. De volta ao exemplo do mercado, veja a variedade de produtos existentes e quantas regras podem ter que ser processadas. Por exemplo, a compra de leite e pão implica na do café. De acordo com Borgelt e Kruse (2002), faz-se necessário o desenvolvimento de algoritmos eficientes que restrinjam a busca a um subconjunto de todas as regras, sempre buscando a menor perda de informação possível.

2.2.1 Avaliação das regras

Nem todas as regras existentes são relevantes e muito menos implicam uma real dependência entre X e Y. Como forma de avaliar essas regras, foram propostas algumas medidas. Vale ressaltar que essas medidas também são propulsoras de diversos algoritmos para geração de regras. Nessa subseção ,serão apresentadas quatro medidas primordiais para essa técnica de mineração de dados.

Suporte

O suporte é a métrica mais implícita das regras de associação. A partir dessa medida, é possível observar quão frequentes um conjunto de itens são em determinada transação. O suporte também serve como uma forma de avaliar quais regras são mais importantes. Em cenários onde encontram-se suportes muito pequenos, não é possível obter informações suficientes a respeito da relação entre itens. Com isso, não se pode tirar nenhuma conclusão sobre essa regra. Dessa forma, muitos pesquisadores acabam definindo um limite mínimo para o suporte, com o intuito de não se obter regras fracas. Calcula-se o suporte de uma regra pela seguinte expressão:

$$Suporte(\{X\} \rightarrow \{Y\}) = \frac{frq(X,Y)}{N}. \tag{2.7}$$

Confiança

A confiança é a probabilidade condicional de Y ocorrer dado que X já ocorreu. Em conjunto com o suporte, a confiança é uma das formas mais comuns de mensurar a eficiência de uma regra. Calculamos a confiança por:

$$Confiança(\{X\} \rightarrow \{Y\}) = \frac{frq(X,Y)}{frq(X)} = \frac{Suporte(X \cap Y)}{Suporte(X)}. \tag{2.8}$$

Lift

O *lift* ou *interest*, como é chamado por alguns autores, é a medida mais utilizada para avaliar a dependência de uma regra de associação. Ela funciona como uma correlação. Porém, têm seu resultado variando entre zero e infinito. Uma vez que $lift(\{X\} \rightarrow \{Y\}) = 1$, dizemos que X é independente de Y. Caso $lift(\{X\} \rightarrow \{Y\}) > 1$, são positivamente dependentes e, consequentemente, para *lift* inferior a um temos uma dependência negativa entre X e Y. O fórmula para calcular o *lift* é dada por:

$$Lift(\{X\} \rightarrow \{Y\}) = \frac{Suporte(X \cap Y)}{Suporte(X) \times Suporte(Y)}. \tag{2.9}$$

Convicção

A convicção, assim como o *lift*, é uma medida que avalia a dependência. Um alto valor de convicção implica que o evento consequente tem uma alta dependência do evento antecedente. Note que o denominador pode ser igual a zero, no cenário de uma confiança perfeita. Quando isso ocorre, não se é possível calcular a convicção. Semelhante ao *lift*, se os itens são independentes, a convicção é 1. Sua formula é dada por:

$$Convição(\{X\} \rightarrow \{Y\}) = \frac{1 - Suporte(Y)}{1 - Confiança(\{X\} \rightarrow \{Y\})}. \tag{2.10}$$

2.2.2 Alogritmo *Apriori*

O algoritmo ***Apriori*** foi proposto por Agrawal, Imieliński e Swami (1993), tornando o mais conhecido para indução de regras de associação pela sua simplicidade. Além disso, Agrawal, Srikant et al. (1994) também provaram a eficiência de processamento do algoritmo, mesmo em banco de dados grandes e com diversas possibilidades de combinações.

O algoritmo consiste em duas etapas, no qual as regras devem satisfazer um limite mínimo das medidas de suporte e confiança, respectivamente. Ao não atingir esses limites mínimos, os itens acabam saindo do subconjunto que faz sucessivas buscas por regras, tornando assim o processo mais veloz. A escolha desses limites de suporte e confiança cabem ao pesquisador.

Outra grande vantagem do Apriori é o ele está disponível em diversas linguagens de programação, como é o caso do R project e Python.

2.2.3 Exemplo

Para os resultados que serão apresentado a seguir, considere um banco de dados fictício apresentado na Tabela 2.7. Suponha que esses dados são provenientes das compras em um mercado por 5 clientes.

Tabela 2.7: Itens adquiridos em um mercado, por cliente.

Cliente	Itens
1	{pão, leite}
2	{pão, fralda, cerveja, ovos}
3	{leite, fralda, cerveja, refrigerante}
4	{pão, leite, fralda, cerveja}
5	{pão, leite, fralda, refrigerante}

O primeiro passo para se obter regras de associação consiste na escolha de um algoritmo. No caso desse trabalho, o algoritmo escolhido foi o Apriori que se baseia em duas etapas: a primeira etapa fundamenta-se em um limite mínimo estipulado para o suporte. Neste exemplo, ilustramos o suporte mínimo como sendo 0,60. A Figura 2.9 ilustra essa parte do algoritmo.

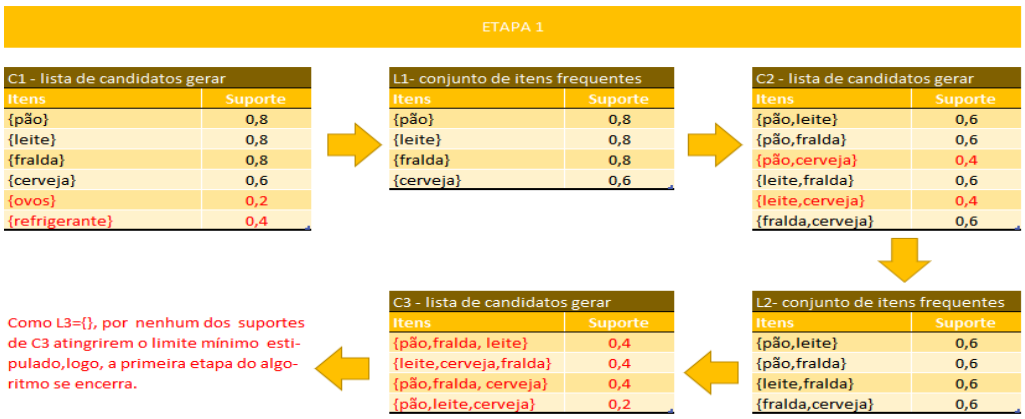


Figura 2.9: Primeira etapa algoritmo Apriori.

Note que é gerado um subconjunto com todos os possíveis candidatos à virarem regras, e deles é verificado o suporte. Caso o suporte seja inferior ao valor mínimo determinado, ele sai do subconjunto de busca no próximo estágio. Na hipótese do suporte ser superior ao limite é gerado uma nova lista de candidatos, aumentando o número de itens avaliados. Na Figura 2.9 definimos a lista de candidatos por C e o subconjunto de busca por L. Esse processo ocorre até o momento em que o subconjunto de busca seja vazio.

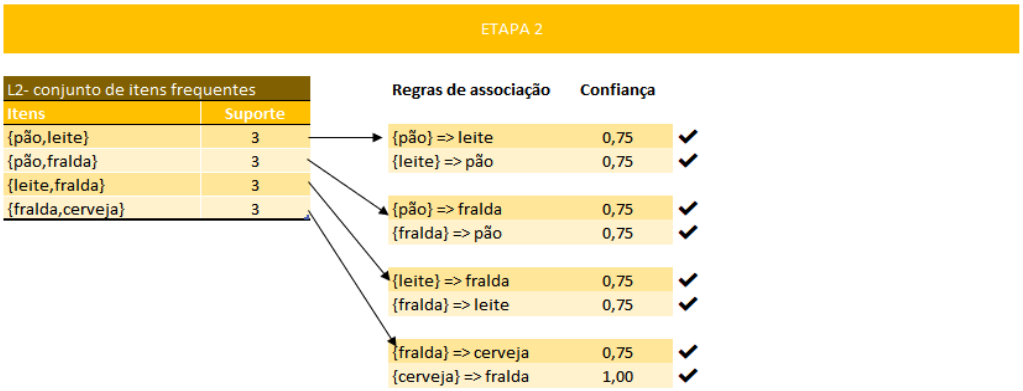


Figura 2.10: Segunda etapa algoritmo Apriori.

Na Figura 2.10, temos a representação da segunda etapa do Aprirori. Nesse estágio, a partir de um limite mínimo pré definido para a confiança, verificamos quais itens dos subconjuntos resultantes da primeira etapa realmente são regras. Observe que a ordem com que os itens aparecem na regra podem alterar o resultado da confiança. Com isso, adotando um limite mínimo de 0,50 para a confiança, temos 8 regras geradas.

Por fim, se têm as medidas de avaliação das regras. Para o uso do algoritmo Apriori necessitamos calcular o suporte e a confiança. Todavia, para melhor ilustração desse exemplo, também apresentaremos os seus cálculos. Tomando como exemplo a regra{pão}=>{leite}, temos:

$$Suporte(\{\text{pão}\} => \{leite\}) = \frac{\text{frq(pão, leite)}}{N} = \frac{3}{5} = 0,60$$

$$Confiança(\{\text{pão}\} => \{leite\}) = \frac{\text{frq(pão, leite)}}{\text{frq(pão)}} = \frac{3}{4} = 0,75$$

$$Lift(\{\text{pão}\} => \{leite\}) = \frac{Suporte(\text{pão} \cap \text{leite})}{Suporte(\text{pão}) \times Suporte(\text{leite})} = \frac{0,6}{0,8 \times 0,8} = 0,94$$

$$Convição(\{\text{pão}\} => \{leite\}) = \frac{1 - Suporte(\text{leite})}{1 - Confiança(\{\text{pão}\} => \{\text{leite}\})} = \frac{1 - 0,6}{1 - 0,75} = 1,6$$

A Tabela 2.8 apresenta o resultado da aplicação dessas medidas para todas as 8 regras resultantes do algoritmo Apriori.

Tabela 2.8: Regras de associação das compras do mercado.

Regras	Suporte	Confiança	*Lift*	Convicção
{pão} => {leite}	0,6	0,75	0,94	0,8
{leite} => {pão}	0,6	0,75	0,94	0,8
{pão} => {fralda}	0,6	0,75	0,94	0,8
{fralda} => {pão}	0,6	0,75	0,94	0,8
{leite} => {fralda}	0,6	0,75	0,94	0,8
{fralda} => {leite}	0,6	0,75	0,94	0,8
{fralda} => {cerveja}	0,6	0,75	1,25	1,6
{cerveja} => {fralda}	0,6	1,00	1,25	-

2.3 Auditoria

Com a grande disputa existente entre empresas do mercado, as grandes corporações buscam cada vez mais aumentar seu poder competitivo e, consequentemente, suas receitas e valor de mercado. É com isso que a figura do auditor ganha extrema importância. Essa profissão é responsável por garantir as confiabilidade das informações através de análises dos processos internos de uma empresa. Como consequência do trabalho de garantir a veracidade dos dados, esses auditores podem acabar se deparando com situações de fraudes.

Subdivide-se a auditoria em duas classes: externa ou interna, que são exercidas de maneiras bastante semelhantes porém, com objetivos específicos diferentes. Podemos rotular como auditoria externa, situações no qual se busca a confiabilidade das informações patrimoniais e demonstrativos de resultados, através de análises feitas em relatórios contábeis. Já a auditoria interna serve para avaliar a qualidade dos controles de uma corporação. Outra diferença bem clara entre elas é a continuidade dos trabalhos. A auditoria externa é feita por por terceiros e duram no máximo alguns meses do ano. No caso da interna a auditoria é feita por um funcionário da empresa e o trabalho é contínuo durante o ano todo. Esse trabalho tem como finalidade auxiliar na auditoria interna, visto que pretende-se que o acompanhamento seja feito de maneira contínua por meio de uma ferramenta de *Business Intelligence*.

Um dos objetivo da auditoria é identificar possíveis fraudes e por conta disso é bastante importante definir o termo e sua abrangência para que não seja confundida com erros operacionais. Para Longo, Jiménez e Marcos (2000) diferentemente de um erro,

uma fraude é um ato proposital que resulta em falsas declarações contábeis. Esses atos podem variar de apropriações indevidas, omissão de transações em documentos, aplicação inadequada de regras contábeis, falsificação e manipulação de registros dentre outros. Por conta da vasta variedade de fraudes, as mesmas podem ocorrem em diversos níveis e âmbitos corporativos. O impactando causado por uma fraudação pode ser em observado em uma ou mais áreas, logo, classifica-se então os tipos de fraudes pelos segmentos afetados. Dessa forma Pereira e Nascimento (2005) separa as fraudes em quatro grandes categorias são elas: contábil, financeira, controle interno e ética. O fato comum entre todas elas é o fato de ferirem as normas internas de uma entidade causando danos patrimoniais.

2.4 *Business Intelligence (BI)*

O século XXI é marcado pelos grandes fluxos de informações, gerados a partir dos avanços tecnológicos ao longo dos anos. O fácil acesso a internet e "*a computação em nuvem*"são alguns exemplos de tecnologias que vieram para facilitar a captura e o armazenamento dos dados.

Com esses avanços alguns problemas também começaram a surgir. Dentre eles, destaca-se a dificuldade com processamento dos dados para geração de informação. O dado bruto, em si, não auxilia a tomada de decisão. Assim, sem um processamento adequado esses dados são de pouca utilidade. É nesse contexto que ganha força ferramentas de BI.

O termo *Business Intelligence* apareceu, pela primeira vez, em meados do século passado no artigo *Business intelligence system*. Para Luhn (1958), BI era um sistema desenvolvido para automatizar e disseminar informações para diferentes consumidores e instituições com a finalidade de dar embasamento para que seja tomada a melhor escolha.

Com a alta procura por ferramentas de *business intelligence* diversos softwares foram desenvolvidos e possuem funcionalidade similares como: conexão a base de dados externas, visualização via tabelas e gráficos, medidas de análises e, em alguns casos, a possibilidade de tratamento dos dados. Contudo, cada software desse possui uma configuração, design e linguagem própria.

O BI *Self Service* surgiu como solução para não-programadores, que necessitavam de agilidade para gerar informação. Segundo Palmisano e Rosini (2003), as primeiras ferramentas desenvolvidas eram de difícil manipulação, além de carecer de um alto nível de conhecimento de programação. Vale ressaltar que, mesmo as ferramentas de BI *Self Service*, ainda precisam de conhecimentos de sintaxe para análises avançadas. Contudo,

o básico pode ser feito quase que por inteiro com cliques.

A empresa Gartner de consultoria e pesquisa realiza, anualmente, uma pesquisa elegendo os melhores softwares de BI oferecidos pelo mercado. Para avaliar cada uma das plataformas de BI quinze pontos críticos são levados em consideração, são eles: segurança, capacidade de gerenciamento, conectividade da fonte de dados, visualização de dados, nuvem, preparação dos dados, complexidade do modelo, catálogo de suplementos, informações automatizadas, análise avançada, geração de linguagem natural, análise incorporada, narração dos dados, criação de *dashboards*.

Na figura abaixo mostra um quadrante que classifica as ferramentas de BI em quatro grupos, são eles:

1. **Leader:** São os softwares mais completos do mercado. Esses possuem uma alta capacidade de se adaptar a diversas áreas de atuação e problemas. Além disso, tem suas limitações bem definidas. Outras vantagens desses produtos são: excelente preço, escala corporativa e fácil compreensão. Por todos esses fatores podemos rotular essas plataformas de BI como líderes do mercado.

2. **Challengers:** Como já diz o nome, são produtos novos que podem ou não obter sucesso no mercado. Muitos deles acabam esbarrando em limitações técnicas, por serem funcionais apenas a ambientes específicos.

3. **Visionaries:** Os visionários são os responsáveis por oferecer uma plataforma BI moderna, com aplicações profundas para as áreas que estão voltadas. Contudo, eles podem ter problemas para atender demandas mais funcionais ou atividades que necessitam baixa experiência do cliente.

4. **Niche players:** Os jogadores de nicho são produtos voltados para um segmento específico do mercado. Os mesmos podem ser orientados para o setor financeiro, logística, estoque, *etc*. Por serem desenvolvidos apenas para um objetivo, eles acabam se tornando limitados nos quesitos inovação e desempenho. Além disso, muitas vezes não conseguem competir com produtos mais flexíveis do mercado.

Na Figura 2.11, a *Microsoft* é atualmente a plataforma mais a direita do quadrante, segundo a pesquisa da Gartner (2020). A plataforma de *business intelligence* que representa a *Micorsoft* é o ***Power BI***. O gerente geral de plataforma, Arun Ulag, publicou no blog oficial da empresa o resultado da pesquisa da Gartner desse ano.

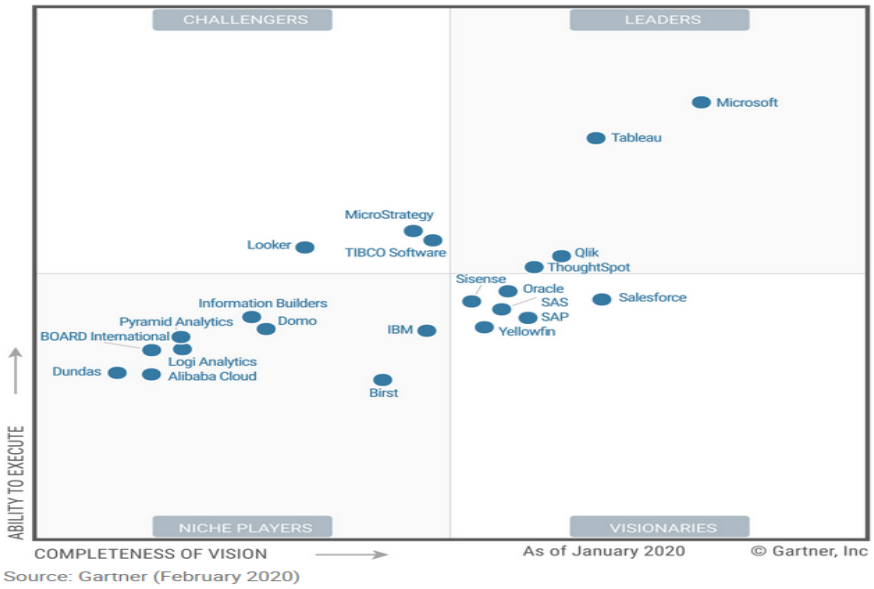


Figura 2.11: Quadrante mágico de softwares de BI da Gartner.

2.4.1 Power BI

O Power BI é o software de BI *self service* desenvolvido pela *Microsoft* que oferece todas as funcionalidades esperadas de uma plataforma classificada como líder, pela Gartner. Mesmo com seu lançamento em 24 de julho de 2015, em menos de 5 anos, já dominava o mercado. Trata-se de uma ferramenta completa, onde é possível se conectar a diversas fontes de dados e realizar análises. Segundo Lago e Alves (2018), outro diferencial do Power BI é a capacidade de tratamento dos dados, uma vez que nem todos os concorrentes de mercado tem essa mesma aptidão.

Segundo Lago e Alves (2018) o Power BI surgiu de suplementos e tecnologias que, inicialmente, foram desenvolvidas para o Excel. Os principais suplementos são: **Power Query Editor, Power View e Power Pivot**. Esses elevaram o patamar das planilhas eletrônicas por aumentar as possibilidades de tratamentos e análises dos dados. Tais funcionalidade levarem a *Microsoft* a estar no quadrante dos líderes da Gartner, na categoria plataformas de análises e *business intelligence*, pelo 13^o ano consecutivo, ou seja, bem antes da existência do Power BI.

O Power BI têm três componentes, cada uma referente a um dos suplementos citados acima. O **Power Query Editor** é o ambiente no qual se faz a importação, conexão e tratamentos dos dados. É também no Query que se faz a integração do Power BI com os relatórios de outros sistemas (*DirectQuery*). A linguagem de programação desse componente é o **M**. Sua sintaxe é *case-sensitive*, ou seja, é sensível a letras maiúsculas e minúsculas e já possui mais de 700 funções disponíveis. É no Power Query que ocorre

a primeira etapa de um projeto de BI. É nessa fase que ocorre a padronização e filtros dos dados para uma melhor performance das medidas que pertencem a segunda etapa do processo.

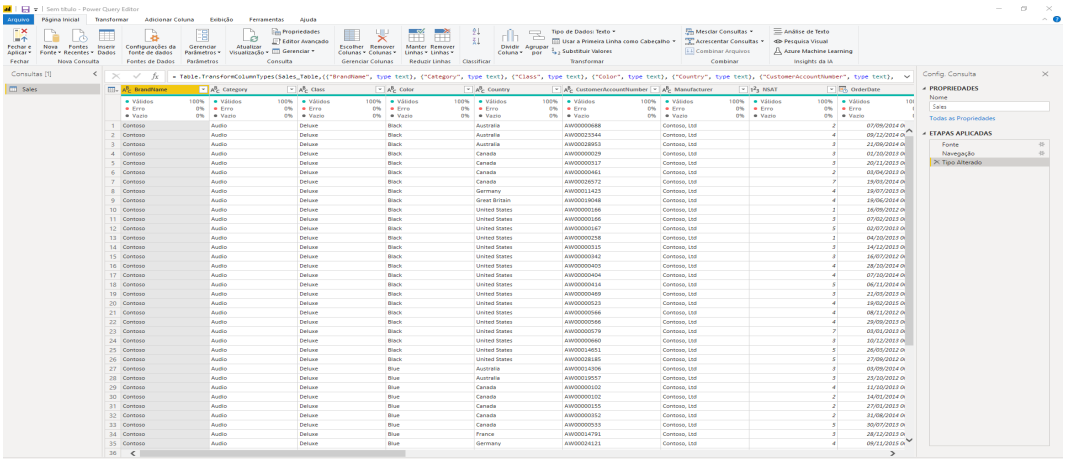


Figura 2.12: Editor de Power Query no Power BI.

A segunda etapa é denominada de modelagem. Nesse momento, são criados relacionamentos entre tabelas e são feitas medidas que processam as linhas da tabela para gerar alguma informação. No Excel, essa ferramenta se chama **Power Pivot**. A linguagem desse ambiente é o DAX e, segundo Ferrari e Russo (2015), é uma linguagem simples em comparação a outras linguagens. Sua sintaxe não é *case-sensitive*, sendo similar aos cálculos do Excel.

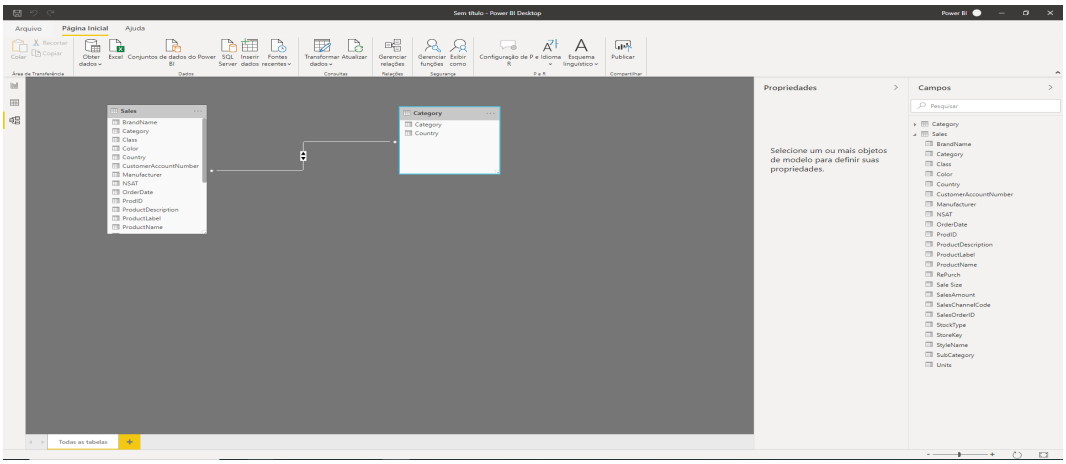


Figura 2.13: Relacionamento entre tabelas Power BI.

A terceira e última etapa é a da visualização dos dados. A Power BI disponibiliza um grande leque de opções de visualizações como: tabelas, gráficos, matrizes, mapas, cartões, além de integração com outros softwares como o R project e o Python, permitindo com que um novo visual seja criado. Ainda existe um espécie de loja, no qual se pode baixar

outros visuais que, em geral, são ferramentas de análises mais avançadas e especificas. Por exemplo, é possível baixar alguns tipos de gráficos de controle e de séries temporais.

Lago e Alves (2018) também dividem o Power BI em três plataformas: desktop, serviço e celulares. Cada uma possui uma particularidade, até por conta de estarem relacionadas aos planos pagos da *Microsoft*. O Power BI Desktop é a plataforma do desenvolvedor. Apenas nessa, é possível criar medidas e relacionar tabelas. O Power BI Serviço é uma espécie de nuvem. Seu objetivo é gerenciar os relatórios. Essa plataforma permite ao usuário criar visuais. Por fim, plataforma de celular possibilita acesso aos *dashboards* produzidos via mobile.

A *Microsoft* oferece três opções de planos para seus usuários, sendo eles: **Free**, **Pro** e **Premium**. Esses variam nos preços das mensalidades e na capacidade de realizar algumas funcionalidades. Maiores informações sobre o Microsoft Power BI podem ser obtidas em <https://powerbi.microsoft.com/pt-br/pricing/>.

CAPÍTULO 3

APLICAÇÕES

Neste Capítulo apresentaremos a aplicação da Lei de Benford a dois bancos de dados reais, sendo um deles referente a gastos e o outro referente às receitas operacionais, ambos de um hospital privado da cidade de João Pessoa. Por questões de confidencialidade tanto o nome do hospital quanto a dos fornecedores não serão expostos. Vale destacar que quando necessários para as aplicações os fornecedores serão apresentados por meio de identificadores numéricos. O primeiro conjunto de dados refere-se aos pagamentos efetuados pelas compras de medicamentos. O segundo conjunto de dados está ligado ao faturamento do setor de Unidade de Tratamento Intensivo (UTI). Adicionalmente será realizada a aplicação da técnica denominada regras de associação para o banco de dados dos pagamentos resultantes de compras de medicamentos.

Todos os resultados apresentados a seguir foram feitos nos softwares *R-project*[1] e *Power BI*[2]. No R, para a lei de Benford, o pacote *benford.analysis* foi tomado como base, criando a partir dele uma função que retorna os testes qui-quadrado, DMA e especialmente o teste Z, uma vez que o mesmo não está implementado no pacote original. As funções criadas estão disponibilizadas no Apêndice A. Para a aplicação da lei de Benford no Power BI foi necessário implementar todas as medidas na linguagem DAX. As regras de associação foram feitas no Power BI a partir de funções do R, por meio da integração entre os *softwares*.

[1] Team et al. (2013)
[2] *Microsoft corporation*

3.1 Medicamentos

A primeira análise será para a aquisição de medicamentos feitas por um hospital particular de João Pessoa referentes ao primeiro semestre de 2019. A compra de fármacos é uma das atividades mais essenciais para a gestão e manutenção do serviço assistencial, tendo em vista que são de suma importância para o tratamento de diversas enfermidades. Além disso por envolver um gasto muito grande de recursos financeiros as compras de medicações em geral têm um impacto grande nos custos de uma empresa e, como consequência, na sua saúde financeira. Segundo Sforsin et al. (2012), outro ponto a se destacar além do financeiro é a qualidade do produto comprado, dado que os hospitais têm que dar a assistência farmacêutica adequada a cada paciente. Portanto, é importante se ter uma boa gestão de compras de medicamentos no âmbito hospitalar.

Na Tabela 3.1, observa-se que, no período de seis meses, 1458 pagamentos de compras de medicamentos foram realizadas. Existe uma variação entre os meses em relação à quantidade de aquisições feitas. O mês de fevereiro teve os menores percentuais tanto na quantidade como no valor dos pagamentos do período, representando cerca de 13,8% e 9,38% do total. Para todos os meses, o coeficiente de assimetria foi positivo e os valores médios foram superiores à mediana, fato que segundo Wallace (2002) dá algum indício de que os dados analisados estão conforme à lei de Benford.

Tabela 3.1: Estatísticas descritivas das compras de medicamentos.

mês	n	valor pago	média	mediana	des.pad.	mín	máx	amplitude	assimetria
jan	242	972.072,15	4.016,83	859,80	8.132,67	10,79	45.922,02	45.911,23	3,18
fev	201	573.030,07	2.850,90	798,70	5.323,76	40,00	45.922,02	45.882,02	4,78
mar	234	1.100.144,57	4.701,47	1.046,48	9.816,33	4,80	55.473,52	55.468,72	3,16
abr	313	1.309.573,27	4.183,94	1.286,24	6.781,90	6,27	51.249,81	51.243,54	2,96
mai	243	1.007.796,58	4.147,31	1.422,35	6.770,29	6,27	40.874,40	40.868,13	2,93
jun	225	1.148.886,35	5.106,16	1.851,50	8.803,98	13,22	47.456,76	47.443,54	3,08
Total	1458	6.111.502.99	4.191,70	1.158,36	7.748,78	4,80	55.473,52	55.468,72	3,38

Primeiro Dígito

A Figura 3.1 apresenta as frequências relativas observadas do primeiro dígito referentes a compras de medicamentos feitas pelo hospital. Como as proporções observadas estão muito próximas das proporções de Benford, pode-se afirmar que o primeiro dígito está conforme à lei.

A Tabela 3.2 mostra os resultados dos testes Z, Qui-Quadrado e DMA do primeiro dígito dos pagamentos efetuados de compras de medicamentos. Para o teste Qui-Quadrado

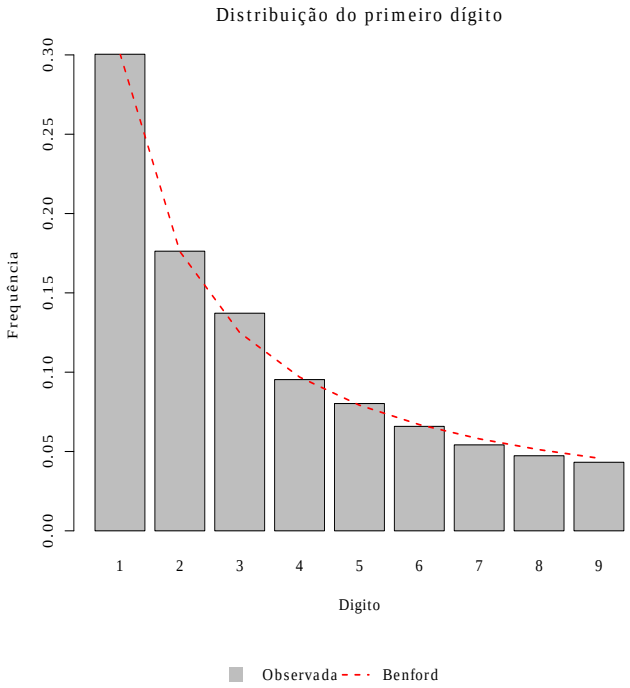


Figura 3.1: Distribuição do primeiro dígito para a compra de medicamentos.

com 5% de significância ($\chi^2(8) = 15,507$) a hipótese nula não é rejeitada, implicando que as frequências relativas observadas e esperadas são iguais, por consequência há evidências estatísticas suficientes para dizer que os primeiros dígitos dos dados estão conforme à Lei de Benford. No teste Z e no DMA a conclusão foi a mesma. Considerando o valor crítico de 1,96 ($\alpha = 0,05$) no teste Z, temos que para nenhum dos dígitos a hipótese nula foi violada implicando que as proporções observadas dos dígitos não diferem das frequências relativas esperadas. Por fim, o DMA calculado foi de 0,0030. Logo, pela classificação proposta por Drake e Nigrini (2000), os dados seguem a Lei de Benford.

Tabela 3.2: Teste do primeiro dígito da compra de medicamentos.

Dígito	Frequência Observada	Frequência Esperada	Proporção Observada	Benford	$\lvert P_O - P_E \rvert$	Teste Z	Teste χ^2	DMA
1	438	438,9017	0,3004	0,3010	0,0006	0,0229	0,0019	0,0001
2	257	256,7411	0,1763	0,1761	0,0002	0,0178	0,0003	0,0000
3	200	182,1607	0,1372	0,1249	0,0122	1,3734	1,7470	0,0014
4	139	141,2948	0,0953	0,0969	0,0016	0,1589	0,0373	0,0002
5	117	115,4463	0,0802	0,0792	0,0011	0,1022	0,0209	0,0001
6	96	97,6084	0,0658	0,0669	0,0011	0,1161	0,0265	0,0001
7	79	84,5523	0,0542	0,0580	0,0038	0,5661	0,3646	0,0004
8	69	74,5804	0,0473	0,0512	0,0038	0,6039	0,4175	0,0004
9	63	66,7144	0,0432	0,0458	0,0025	0,4029	0,2068	0,0003
Total	1.458	1.458,0000	1,0000	1,0000	0,0269	-	2,8228	0,0030

Segundo Dígito

Ao aplicar o teste Z para o segundo dígito, observou-se que existe uma não conformidade com a lei de Benford para os pagamentos cujo segundo dígito é 1, considerando 5% de significância. A Figura 3.2 evidencia essa não-congruência.

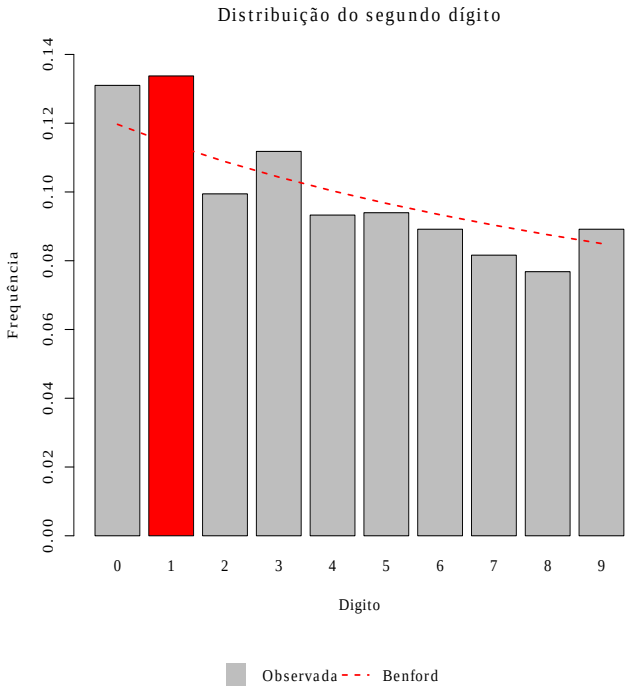


Figura 3.2: Distribuição do segundo dígito para a compra de medicamentos.

Na Tabela 3.3 é possível verificar os resultados dos testes Qui-Quadrado e Desvio Médio Absoluto. Apesar de ter um dígito não conforme, ao analisar o contexto global pode-se dizer que os dados estão de acordo com a lei, uma vez que a hipótese nula não é rejeitada para o teste Qui-Quadrado ($\chi^2(9) = 16,92 > 13,12 = \chi^2$) e o DMA é classificado como conformidade aceitável.

Ao fazer uma análise mais detalhada para o dígito 1, verificou-se que o número de pagamentos parcelados para esse dígito é proporcionalmente maior do que a média de todos os dígitos em um pouco mais de 2%. Outro possível motivo para alteração das frequências observadas se dá pelo fato de ser o dígito que possui a menor proporção de acréscimos sendo cerca de 9% menor que a proporção global.

Tabela 3.3: Teste do segundo dígito da compra de medicamentos.

Dígito	Frequência Observada	Frequência Esperada	Proporção Observada	Benford	$\lvert P_O - P_E \rvert$	Teste Z	Teste χ^2	DMA
0	191	174,4934	0,1310	0,1197	0,0113	1,2915	1,5615	0,0011
1	195	166,0516	0,1337	0,1139	0,0199	2,3453	5,0467	0,0020
2	145	158,6596	0,0995	0,1088	0,0094	1,1067	1,1760	0,0009
3	163	152,1131	0,1118	0,1043	0,0075	0,8899	0,7792	0,0007
4	136	146,2520	0,0933	0,1003	0,0070	0,8501	0,7186	0,0007
5	137	140,9594	0,0940	0,0967	0,0027	0,3066	0,1112	0,0003
6	130	136,1335	0,0892	0,0934	0,0042	0,5071	0,2763	0,0004
7	119	131,7303	0,0816	0,0904	0,0087	1,1173	1,2302	0,0009
8	112	127,6771	0,0768	0,0876	0,0108	1,4061	1,9249	0,0011
9	130	123,9300	0,0892	0,0850	0,0042	0,5231	0,2973	0,0004
Total	1.458	1.458,0000	1,0000	1,0000	0,0857	-	13,1220	0,0086

Dois Primeiros Dígitos

No teste Z dos dois primeiros dígitos, observa-se 6 pares de dígitos que ultrapassam o valor crítico adotado de 1,96: 10, 14, 24, 34, 45 e 51. A Figura 3.3 tem a representação gráfica das proporções observadas em comparação com as esperadas. As barras em vermelho são os dígitos que rejeitaram a hipótese nula pelo teste Z, considerando 5% de significância.

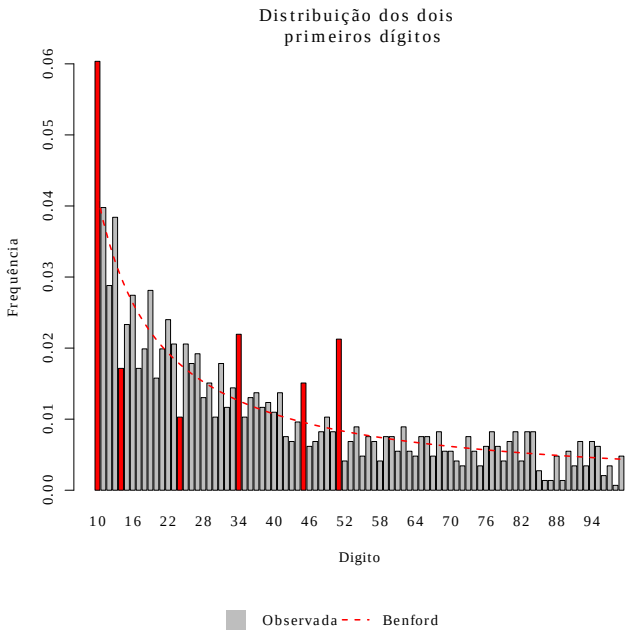


Figura 3.3: Distribuição dos dois primeiros dígitos para a compra de medicamentos.

A estatística Qui-Quadrado para os dois primeiros dígitos foi de 154,30. Com 5% de significância e 89 graus de liberdade, o valor crítico é de 112,022. Uma vez que a

estatística obtida é superior ao valor crítico, rejeita-se a hipótese nula de que os dados estão de acordo com a lei de Benford.

Por fim, no teste DMA, a estatística observada foi de 0,0027, excedendo em aproximadamente 0,0005 o limiar máximo proposto por Drake e Nigrini (2000), caracterizando um cenário de não conformidade com Benford, apesar da proximidade com esse limite.

Tabela 3.4: Teste dos dois primeiros dígitos da compra de medicamentos.

Dígito	Frequência Observada	Frequência Esperada	Proporção Observada	Benford	$\lvert P_O - P_E \rvert$	Teste Z	Teste χ^2	DMA
10	88	60,3505	0,0604	0,0414	0,0190	3,5694	12,6675	0,0002
11	58	55,0957	0,0398	0,0378	0,0020	0,3302	0,1531	0,0000
12	42	50,6832	0,0288	0,0348	0,0060	1,1700	1,4876	0,0001
13	56	46,9253	0,0384	0,0322	0,0062	1,2724	1,7549	0,0001
14	25	43,6864	0,0171	0,0300	0,0128	2,7937	7,9929	0,0001
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
22	35	28,1469	0,0240	0,0193	0,0047	1,2092	1,6686	0,0001
23	30	26,9488	0,0206	0,0185	0,0021	0,4960	0,3455	0,0000
24	15	25,8485	0,0103	0,0177	0,0074	2,0537	4,5531	0,0001
25	30	24,8346	0,0206	0,0170	0,0035	0,9443	1,0744	0,0000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
33	21	18,9029	0,0144	0,0130	0,0014	0,3697	0,2326	0,0000
34	32	18,3549	0,0219	0,0126	0,0094	3,0877	10,1437	0,0001
35	15	17,8378	0,0103	0,0122	0,0019	0,5570	0,4515	0,0000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
45	22	13,9171	0,0151	0,0095	0,0055	2,0424	4,6945	0,0001
46	9	13,6178	0,0062	0,0093	0,0032	1,1211	1,5659	0,0000
47	10	13,3310	0,0069	0,0091	0,0023	0,7789	0,8323	0,0000
48	12	13,0562	0,0082	0,0090	0,0007	0,1546	0,0854	0,0000
49	15	12,7924	0,0103	0,0088	0,0015	0,4795	0,3810	0,0000
50	12	12,5391	0,0082	0,0086	0,0004	0,0111	0,0232	0,0000
51	31	12,2956	0,0213	0,0084	0,0128	5,2137	28,4539	0,0001
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
97	5	6,4944	0,0034	0,0045	0,0010	0,3911	0,3439	0,0000
98	1	6,4285	0,0007	0,0044	0,0037	1,9481	4,5841	0,0000
99	7	6,3639	0,0048	0,0044	0,0004	0,0541	0,0636	0,0000
Total	1458	1458,0000	1,0000	1,0000	0,2466	-	151,1156	0,0027

Deste modo, levando-se em consideração os resultados dos três testes de conformidade apresentados na Tabela 3.4, há evidências de uma não aleatoriedade das compras de medicamentos realizadas no primeiro semestre de 2019. Todavia, a diferença no teste DMA foi muito pequena. Além disso, segundo Nigrini (2012), se até 6 dos 90 pares de dígitos mostrarem não-conformidade, em geral ainda se tem um cenário de aleatoriedade.

A Tabela 3.5 apesenta uma análise sintética para os dígitos não conformes, com a finalidade de detectar as possíveis causas das disparidades entre as probabilidades de

Benford e as proporções observadas. Constatou-se que para os pares 34 e 51 o número de pagamentos resultantes de parcelamentos de compras de medicamentos foram bem elevados, aumentando mais de 13%. Para o par 24 ocorre o contrário: o número dos pagamentos parcelados foi 14% inferior à média. Em relação aos prazos com que os pagamentos foram efetuados destacam-se os pares 45 e 51 sendo respectivamente a menor proporção de pagamentos dentro do prazo e a maior. Para o par 14, destaca-se o fornecedor 39, cujo percentual em relação a frequência do par de dígitos foi de 32%.

Tabela 3.5: Detalhamento dos dígitos não conformes.

	Parcelado		Prazo		Acréscimo	
Dígito	sim	não	sim	não	sim	não
10	59,09%	40,91%	22,73%	77,27%	78,86%	21,14%
14	52,00%	48,00%	28,00%	72,00%	68,00%	32,00%
24	40,00%	60,00%	13,33%	86,67%	73,33%	26,67%
34	71,88%	28,12%	21,88%	78,12%	81,25%	18,75%
45	22,73%	77,27%	31,82%	68,18%	95,45%	4,55%
51	67,74%	32,26%	9,68%	90,32%	96,77%	3,23%
Geral	53,98%	46,02%	26,13%	73,87%	78,12%	21,88%

3.2 Unidade de Tratamento Intensivo (UTI)

Nesta subseção será feita a aplicação dos testes dos dígitos para os dados do faturamento do primeiro semestre de 2019 do setor da UTI. A Unidade de Tratamento Intensivo é responsável pela captação de uma boa parcela do faturamento de um hospital, por ser um setor de alto custo. O monitoramento e a assistência ao paciente ocorrem 24 horas. Além disso, por tratar de quadros graves de enfermidade, espera-se que os gastos com medicações e exames sejam altos.

Tabela 3.6: Estatísticas descritivas do faturamento da UTI.

n	valor pago	média	mediana	des.pad.	mín	máx	amplitude	assimetria
535	19.415.040,32	10.116,84	36.289,79	78.999,16	7,78	742.232,82	741.925,04	4,81

A Tabela 3.6 mostra que a média das receitas por atendimento é bem superior a sua mediana caracterizando assim a assimetria à esquerda, que como já dito antes é uma característica que da indícios de conformidade. Outro ponto a se destacar é a amplitude do faturamento. Vale ressaltar que o valor mínimo foi baixo devido à data de registro dos faturamentos não estarem fechadas, com isso novos itens podem ter entrado e pela data de corte não foram contabilizados.

Primeiro dígito

Ao fazer a análise dos dígitos por meio do teste Z, rejeitamos a hipótese nula para todos os dígitos, uma vez que nenhuma estatística de teste foi superior ao valor crítico adotado, 1,96 ($\alpha = 5\%$), o que permite afirmar que, estatisticamente, as proporções observadas e esperadas não são iguais.

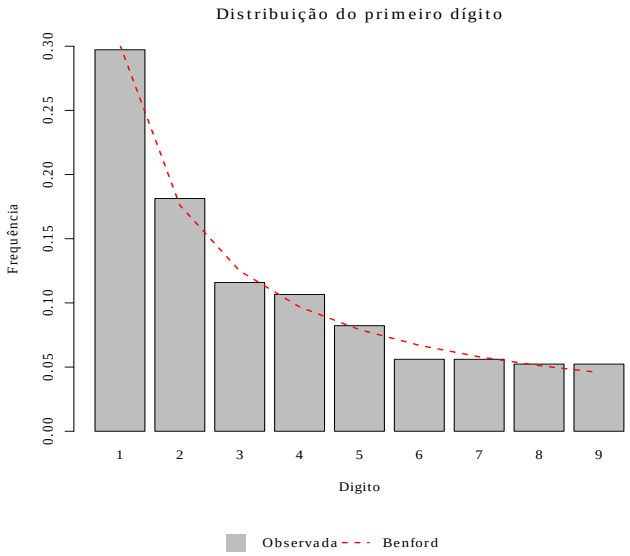


Figura 3.4: Distribuição do primeiro dígito para o faturamento da UTI.

Para o teste Qui-Quadrado com 5% de significância, a hipótese nula também não é rejeitada, dado que $\chi^2(8) = 15,507 > \chi^2 = 2,534$. Com isso pode-se assumir que os dados estão conforme à Lei de Benford. Por fim o DMA obtido foi de 0,0057, indicando um cenário de conformidade.

Tabela 3.7: Teste do primeiro dígito das receitas do setor de UTI

Dígito	Frequência Observada	Frequência Esperada	Proporção Observada	Benford	$\lvert P_O - P_E \rvert$	Teste Z	Teste χ^2	DMA
1	159	161,0510	0,2972	0,3010	0,0038	0,1462	0,0261	0,0004
2	97	94,2088	0,1813	0,1761	0,0052	0,2601	0,0827	0,0006
3	62	66,8422	0,1159	0,1249	0,0091	0,5678	0,3508	0,0010
4	57	51,8469	0,1065	0,0969	0,0096	0,6800	0,5122	0,0011
5	44	42,3620	0,0822	0,0792	0,0031	0,1822	0,0633	0,0003
6	30	35,8165	0,0561	0,0669	0,0109	0,9197	0,9446	0,0012
7	30	31,0257	0,0561	0,0580	0,0019	0,0972	0,0339	0,0002
8	28	27,3666	0,0523	0,0512	0,0012	0,0262	0,0147	0,0001
9	28	24,4803	0,0523	0,0458	0,0066	0,6248	0,5061	0,0007
Total	535	535,0000	1,0000	1,0000	0,0513	-	2,5343	0,0057

Segundo Dígito

No teste de segundo dígito para UTI ao avaliar a adequação pelo teste Z com 5% de significância, a hipótese nula não foi rejeitada para nenhum dos dígitos. A Figura 3.5 mostra que as proporções esperadas e observadas não apresentam grande discrepância entre si.

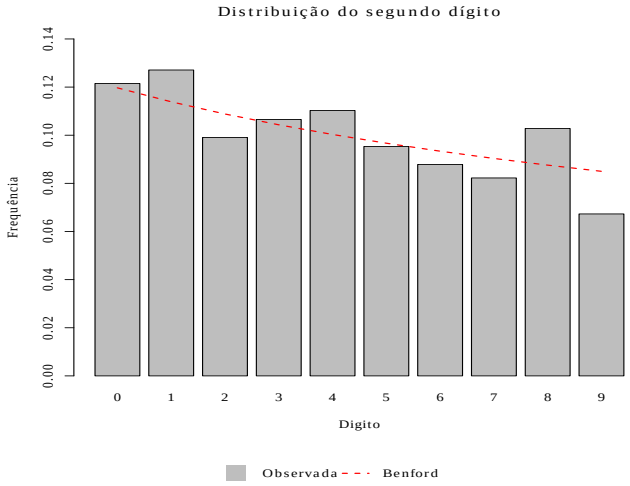


Figura 3.5: Distribuição do segundo para para o faturamento da UTI.

Pela Tabela 3.8 observa-se que o DMA para o teste do segundo dígito foi de 0,0085. Comparando com o valor proposto na Tabela 2.6, pode-se classificar a conformidade como aceitável. A estatística encontrada no teste Qui-Quadrado foi de $\chi^2 = 5,94$, portanto não se deve rejeitar a hipótese nula de que os dados estão conforme à lei de Benford. Com os resultados obtidos nos três testes pode-se afirmar que os dados seguem a distribuição proposta por Benford.

Tabela 3.8: Teste do segundo dígito das receitas do setor de UTI.

Dígito	Frequência Observada	Frequência Esperada	Proporção Observada	Benford	$\lvert P_O - P_E \rvert$	Teste Z	Teste χ^2	DMA
0	65	64,0288	0,1215	0,1197	0,0018	0,0628	0,0147	0,0002
1	68	60,9312	0,1271	0,1139	0,0132	0,8940	0,8201	0,0013
2	53	58,2187	0,0991	0,1088	0,0098	0,6551	0,4678	0,0010
3	57	55,8166	0,1065	0,1043	0,0022	0,0967	0,0251	0,0002
4	59	53,6659	0,1103	0,1003	0,0100	0,6957	0,5302	0,0010
5	51	51,7238	0,0953	0,0967	0,0014	0,0327	0,0101	0,0001
6	47	49,9530	0,0879	0,0934	0,0055	0,3645	0,1746	0,0006
7	44	48,3373	0,0822	0,0904	0,0081	0,5787	0,3892	0,0008
8	55	46,8500	0,1028	0,0876	0,0152	1,1701	1,4178	0,0015
9	36	45,4750	0,0673	0,0850	0,0177	1,3914	1,9742	0,0018
Total	535	535,0000	1,0000	1,0000	0,0849	-	5,8237	0,0085

Dois primeiros dígitos

A Figura 3.6 apresenta as frequências relativas de Benford e as observadas, sendo respectivamente as linhas tracejadas e as barras. Nota-se, que existe apenas uma coluna vermelha, representando o dígito cuja hipótese do teste Z foi rejeitada. O resultado dos outros testes de conformidade estão apresentados na Tabela 3.9.

Para o teste qui-quadrado, a hipótese nula de que as proporções são iguais não é rejeitada, indicando uma boa conformidade das receitas dos atendimentos hospitalares com a lei de Benford. Por outro lado, o teste DMA foi de 0,0032, logo, pela Tabela 2.6 os dados podem ser classificados como não conformes.

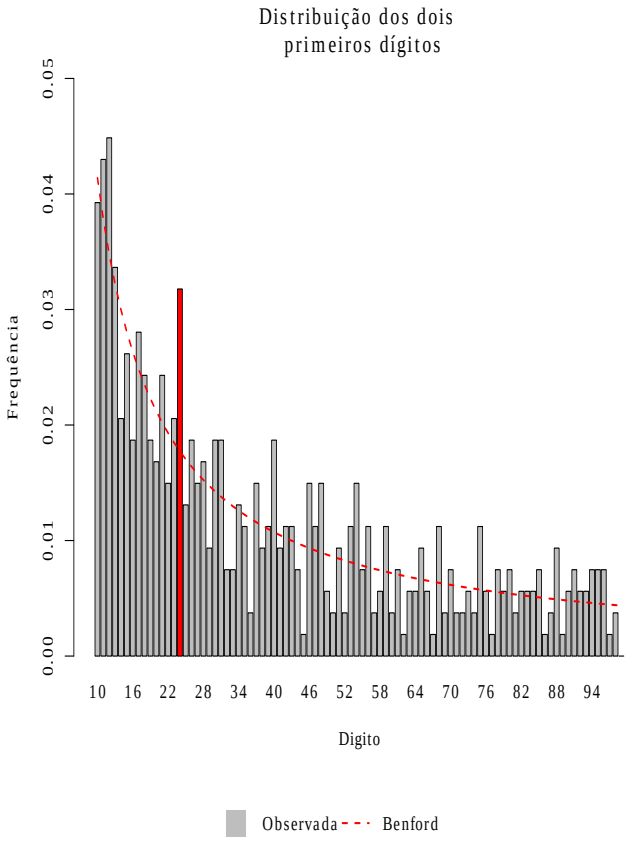


Figura 3.6: Distribuição dos dois primeiros dígitos para o faturamento da UTI.

Uma vez que a hipótese nula foi rejeitada apenas para um par de dígitos pelos teste Z e Qui-quadrado, pode-se dizer que a distribuição dos dígitos das receitas operacionais da UTI estão conforme à lei de Benford.

Tabela 3.9: Teste dos dois primeiros dígitos da receitas do setor de UTI.

Dígito	Frequência Observada	Frequência Esperada	Proporção Observada	Benford	$\lvert P_O - P_E \rvert$	Teste Z	Teste χ^2	DMA
10	21	22,1451	0,0393	0,0414	0,0021	0,1400	0,0592	0,0000
11	23	20,2169	0,0430	0,0378	0,0052	0,5177	0,3831	0,0001
12	24	18,5977	0,0449	0,0348	0,0101	1,1570	1,5693	0,0001
13	18	17,2188	0,0336	0,0322	0,0015	0,0689	0,0354	0,0000
14	11	16,0303	0,0206	0,0300	0,0094	1,1489	1,5785	0,0001
15	14	14,9954	0,0262	0,0280	0,0019	0,1298	0,0661	0,0000
16	10	14,0860	0,0187	0,0263	0,0076	0,9683	1,1852	0,0001
17	15	13,2806	0,0280	0,0248	0,0032	0,3388	0,2226	0,0000
18	13	12,5624	0,0243	0,0235	0,0008	0,1249	0,0152	0,0000
19	10	11,9179	0,0187	0,0223	0,0036	0,4154	0,3086	0,0000
20	9	11,3363	0,0168	0,0212	0,0044	0,5513	0,4815	0,0000
21	13	10,8088	0,0243	0,0202	0,0041	0,5197	0,4442	0,0000
22	8	10,3283	0,0150	0,0193	0,0044	0,5745	0,5248	0,0000
23	11	9,8886	0,0206	0,0185	0,0021	0,1962	0,1249	0,0000
24	17	9,4849	0,0318	0,0177	0,0140	2,2983	5,9544	0,0002
25	7	9,1128	0,0131	0,0170	0,0039	0,5389	0,4899	0,0000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Total	535	535,0000	1,0000	1,0000	0,2837	-	71,6755	0,0032

3.3 Aplicação no Power BI

Esta Seção demonstra a aplicação da lei de Benford juntamente com as regras de associação no software Power BI. Mesclando essas técnicas espera-se encontrar comportamentos para os dígitos não conformes e assim detectar possíveis fraudes. O banco de dados que será usado para essas aplicações será o dos pagamentos de medicamentos, ou seja, os resultados obtidos aqui são similares aos apresentados na subseção anterior com a exceção das regras de associação.

3.3.1 Importação e preparação dos dados

Antes da importação dos dados das compras de medicamentos foram criadas 4 tabelas na linguagem M que serão cruciais para aplicação dos métodos e para desfrutar de todas as funcionalidades do Power BI. A primeira das tabelas criadas foi a Calendário, cujo principal objetivo é auxiliar no relacionamento entre diferentes tabelas de um mesmo projeto, além de ser também um facilitador quando se deseja trabalhar com várias datas diferentes na mesma tabela. Com a tabela Calendário é possível alterar facilmente os contextos da datas com o intermédio de medidas em DAX. Por esses motivos, a tabela Calendário é considerada como uma boa prática para ferramentas de BI (LAGO; ALVES,

2018). Essa tabela percorre todas as possíveis datas entre a menor e a maior data do projeto, incluindo informações adicionais que poderão ser exploradas como: mês, dia da semana, ano, trimestre, dias úteis dentre outros.

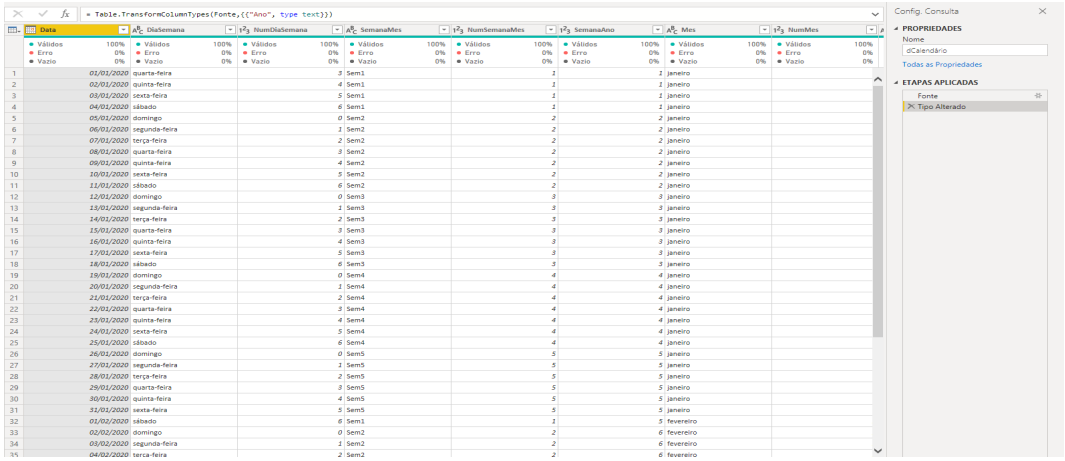


Figura 3.7: Exemplo de tabela calendário.

Em seguida foi criada uma tabela com as probabilidades esperadas de Benford para o primeiro, segundo e dois primeiros dígitos. Essas serão utilizadas para relacionamento entre tabelas, concedendo também a possibilidade de contar com os recursos de *drill down, drill up* e *drill through*. Os *drill's* do Power BI são usados como uma espécie de detalhamento para dados com graus hierárquicos. A Figura 3.8 demonstra a construção do código no PowerQuery Editor para os valores esperados de Benford para o primeiro dígito. O resultado é uma tabela com duas colunas, uma com os dígitos e outra com a probabilidade esperada.

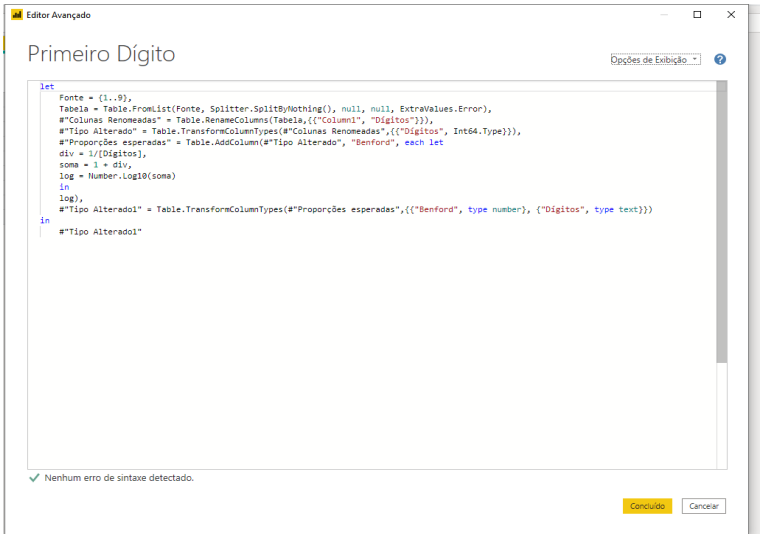


Figura 3.8: Probabilidade esperada do primeiro dígito em M.

Posteriormente foi feita a importação dos dados usando arquivos Excel (*xlsx*) e a definição dos tipos das variáveis. Para a aplicação das regras de associação viu-se a necessidade

de transformar algumas das colunas numéricas em dicotômicas. Ao dicotomizar as variáveis ACRESCIMO e DESCONTO classificamos como 1 todos os pagamentos cujo valor não fosse zero, caso contrario 0. Criou-se ainda a variável PARCELAMENTOS com base no valor da nota fiscal e no valor pago. Caso os valores apresentem diferenças pode-se concluir que o pagamento foi efetuado em momentos diferentes. Além disso, os valores das parcelas podem ser diferentes, dependendo da negociação feita com o fornecedor.

Na Figura 3.9, o pagamento efetuado no dia 18 de fevereiro custou alguns centavos a mais de que as outras duas parcelas. Destaca-se o fato de que essa diferença não foi proveniente de descontos, acréscimos ou atraso na data do pagamentos, implicando que essa diferença ocorre por conta de negociação com o fornecedor. Outro ponto de ressalva é que para esse caso ao realizar o teste dos dois primeiros dígitos os pagamentos serão contabilizados nas frequências dos pares 53 e 54.

DOCUMENTO	VALOR_PAGAMENTO	DT_PAGAMENTO	ACRESCIMO	DESCONTO	PRAZO
2176328	539,95	03/01/2019	0	0	1
2176328	539,95	04/02/2019	0	0	1
2176328	540,10	18/02/2019	0	0	1
Total	**1.620,00**				

Figura 3.9: Exemplo de parcelas com valores diferentes.

3.3.2 Aplicação dos testes dos dígitos

Para a aplicação dos testes dos dígitos, criou-se um relacionamento entre a tabela de medicamentos e as tabelas criadas com as frequências esperadas de Benford. Porém, antes foi necessário gerar três colunas calculadas na sintaxe DAX, extraindo os dígitos do valor dos pagamentos, com cada uma das colunas representando um dos testes de Benford. Antes da extração dos dígitos, utilizamos do teorema da invariância à escala e multiplicamos por 100 a variável com os valores dos pagamentos. O intuito foi auxiliar a remoção dos dígitos com apenas uma parte inteira e duas decimais. Para que seja feita a ligação entre essas tabelas é necessário que elas tenham colunas com valores semelhantes.

A Figura 3.10 apresenta um *dashboard* com as estatísticas descritivas das compras de medicamentos. Por ser uma ferramenta dinâmica ganha-se performance no tempo necessário para se gerar informação. Com alguns filtros é possível alterar facilmente contextos, proporcionando um maior aprofundamento nas análises.

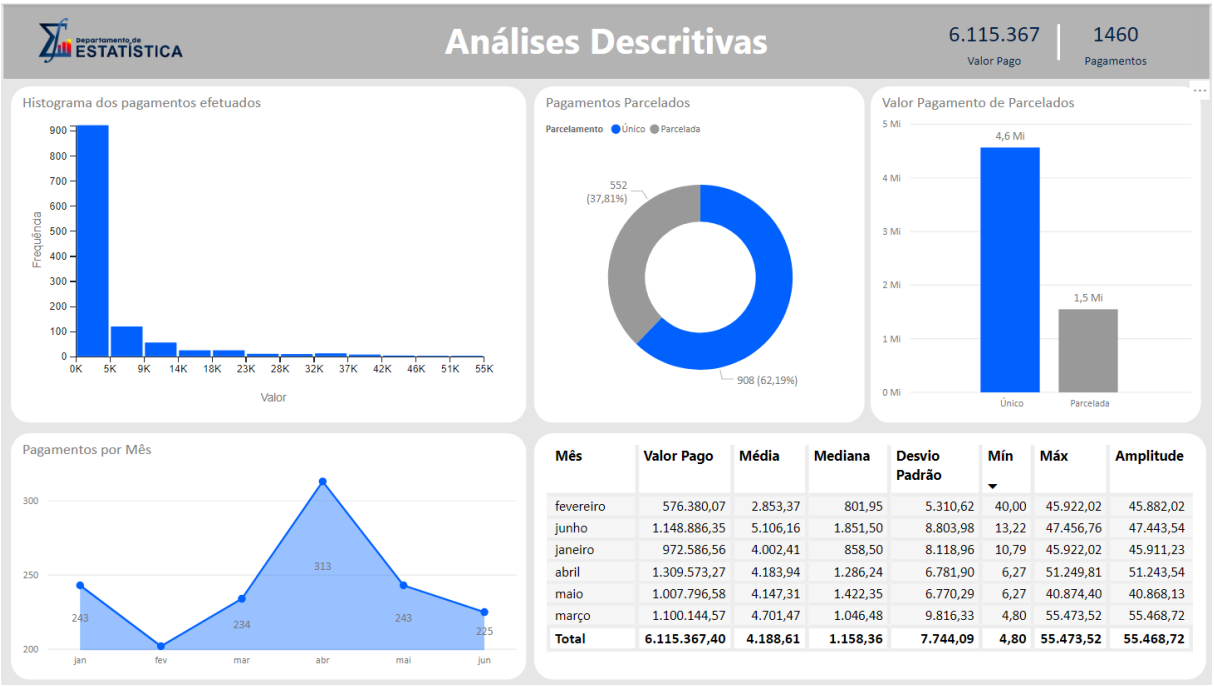
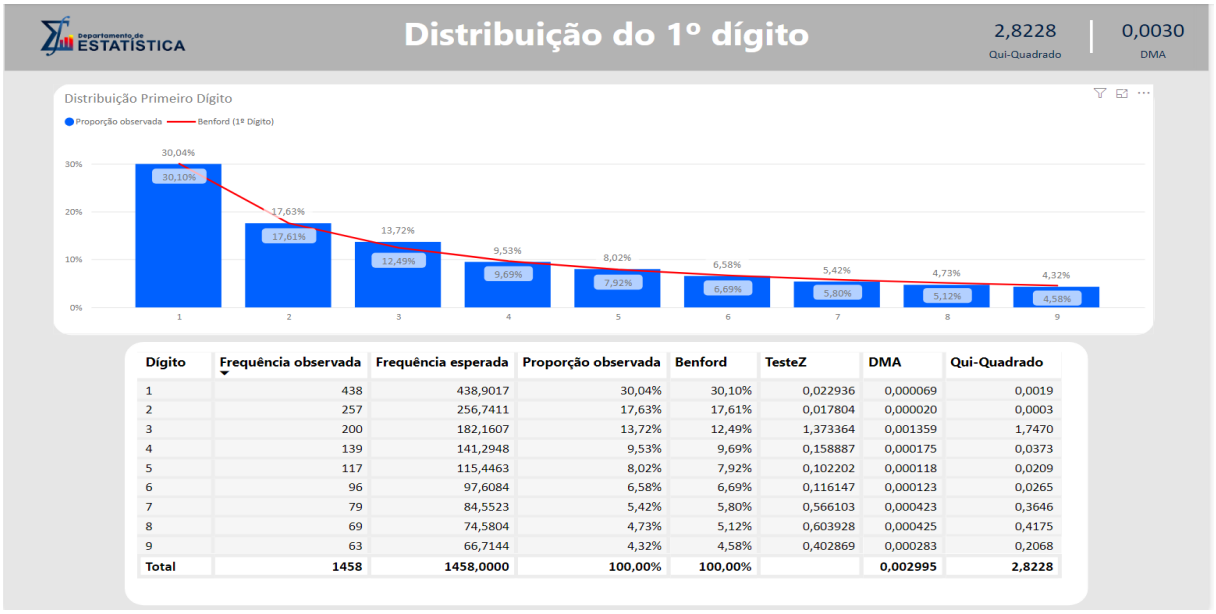


Figura 3.10: *Dashboard* das estatísticas descritivas.

3.3.3 Distribuição dos dígitos

Serão aplicados nessa subseção os testes para os dígitos no Power BI. Todos os resultados apresentados são resultados de medidas implementadas em DAX. A Figura 3.11 representa os testes de conformidade para o primeiro dígito. As barras em azul do gráfico representam as frequências relativas observadas para cada um dos 9 dígitos e a linha em vermelho representa a proporção esperada por Benford. Foi feita uma formatação condicional fazendo com que ocorra uma alteração das cores dos dígitos não conformes pelo teste Z, para diferentes níveis de confiança.

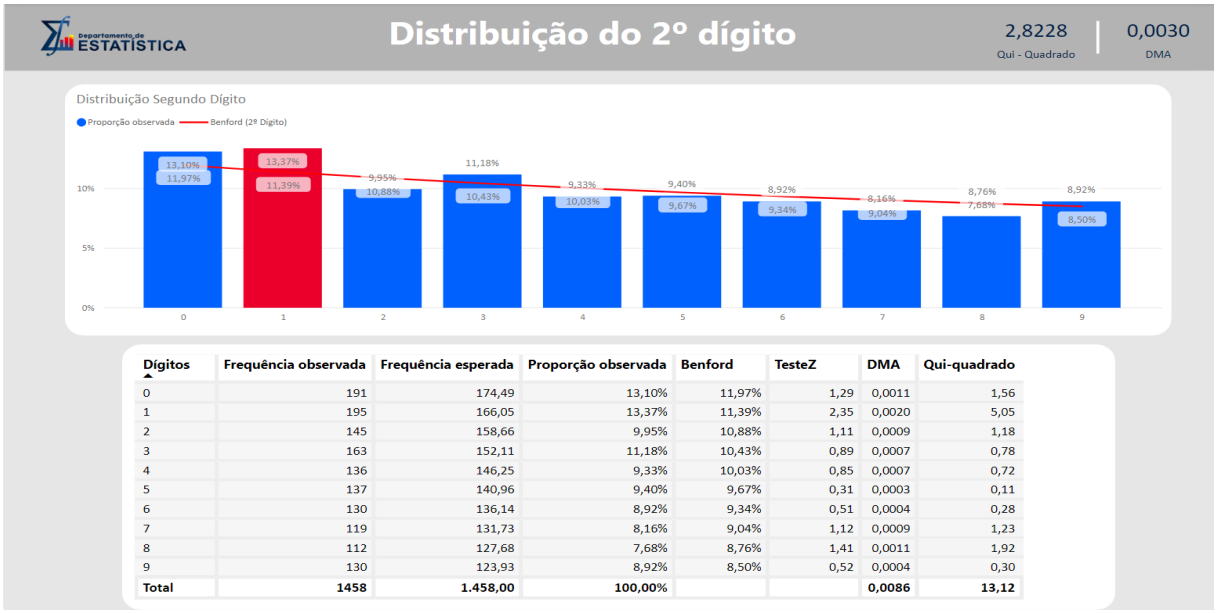
Nota-se pelo *Dashboard* do primeiro dígito, dado na Figura 3.11 que, para os três testes de conformidade aplicados, encontrou-se um cenário de compatibilidade entre as frequências observadas e as frequências esperadas por Benford. O teste Z não expôs anomalias dos dígitos, considerando uma significância de 5%, ou seja, a hipótese nula de que os dígitos seguem a lei de Benford não é violada. Adicionalmente, os testes Qui-quadrado e DMA, que podem ser considerados testes globais de adequação à Lei de Benford, também tiveram suas conclusões sinalizando conformidade com Benford.



Dígito	Frequência observada	Frequência esperada	Proporção observada	Benford	TesteZ	DMA	Qui-Quadrado
1	438	438,9017	30,04%	30,10%	0,022936	0,000069	0,0019
2	257	256,7411	17,63%	17,61%	0,017804	0,000020	0,0003
3	200	182,1607	13,72%	12,49%	1,373364	0,001359	1,7470
4	139	141,2948	9,53%	9,69%	0,158887	0,000175	0,0373
5	117	115,4463	8,02%	7,92%	0,102202	0,000118	0,0209
6	96	97,6084	6,58%	6,69%	0,116147	0,000123	0,0265
7	79	84,5523	5,42%	5,80%	0,566103	0,000423	0,3646
8	69	74,5804	4,73%	5,12%	0,603928	0,000425	0,4175
9	63	66,7144	4,32%	4,58%	0,402869	0,000283	0,2068
Total	**1458**	**1458,0000**	**100,00%**	**100,00%**		**0,002995**	**2,8228**

Figura 3.11: *Dashboard* distribuição do primeiro dígito.

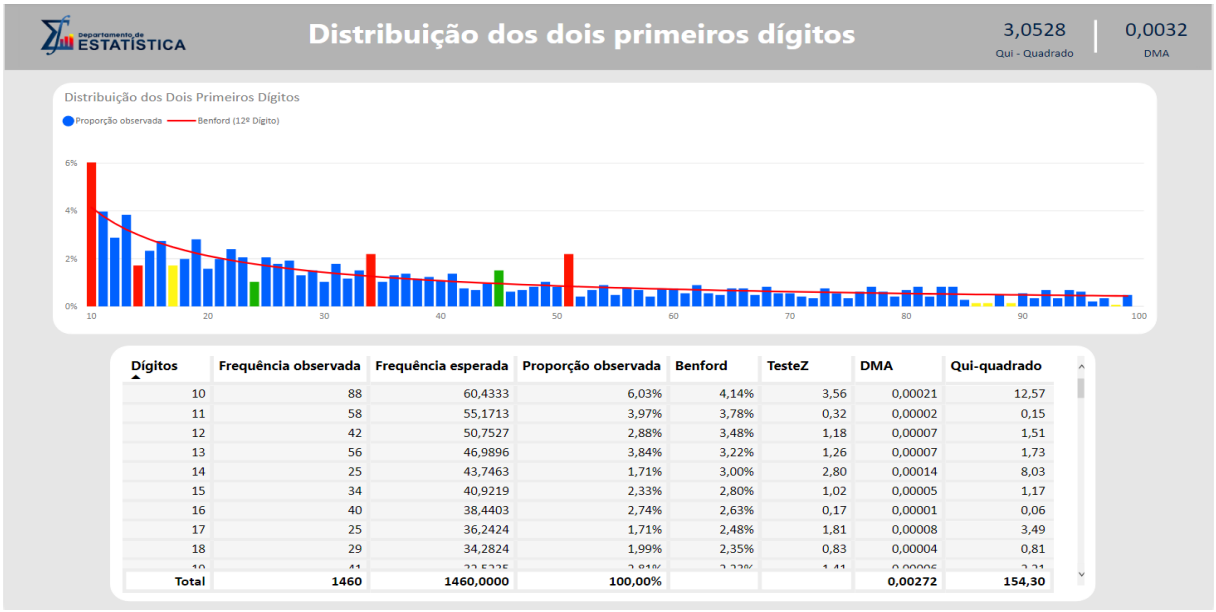
Na Figura 3.12, que se refere ao teste do segundo dígito para os pagamentos efetuados, percebe-se que para o digito 1 apresenta uma distorção em relação as proporções esperadas segundo a lei de Benford. Assim como no *Dashboard* do primeiro dígito criou-se uma formatação condicional do teste Z, implicando na alteração da cor da barra do dígito cuja hipótese nula de conformidade fosse rejeitada. Apesar de termos um dígito não conforme pelo teste Z, de uma maneira geral há conformidade com a Lei de Benford, pois a estatística do teste é inferior a $\chi^2(9) = 16,92$, fornecendo evidências suficientes para a não rejeição da hipótese nula. Além disso ao comparar o DMA com o valor tabelado, a conformidade pode ser considerada aceitável.



Dígitos	Frequência observada	Frequência esperada	Proporção observada	Benford	TesteZ	DMA	Qui-quadrado
0	191	174,49	13,10%	11,97%	1,29	0,0011	1,56
1	195	166,05	13,37%	11,39%	2,35	0,0020	5,05
2	145	158,66	9,95%	10,88%	1,11	0,0009	1,18
3	163	152,11	11,18%	10,43%	0,89	0,0007	0,78
4	136	146,25	9,33%	10,03%	0,85	0,0007	0,72
5	137	140,96	9,40%	9,67%	0,31	0,0003	0,11
6	130	136,14	8,92%	9,34%	0,51	0,0004	0,28
7	119	131,73	8,16%	9,04%	1,12	0,0009	1,23
8	112	127,68	7,68%	8,76%	1,41	0,0011	1,92
9	130	123,93	8,92%	8,50%	0,52	0,0004	0,30
Total	**1458**	**1.458,00**	**100,00%**			**0,0086**	**13,12**

Figura 3.12: *Dashboard* distribuição do segundo dígito.

A Figura 3.13 apresenta os testes para os dois primeiros dígitos no Power BI. No gráfico, os dígitos estão classificados com cores vermelhas, verdes, amarelas e azuis. Cada cor representa o resultado do teste Z considerando diferentes níveis de confiança. Rejeitamos o teste Z considerando 99% ,95% e 90% de confiança para as cores vermelho, verde e amarela respectivamente. Consequentemente quando azul, não se rejeita a hipótese nula de que seguem a distribuição de Benford. Com a estatística Qui-Quadrado de 154,30 e considerando 5% de significância, a hipótese nula é rejeitada uma vez que estatística obtida é superior ao valor crítico de 112,022. O DMA observado foi de 0,0027, que excede o limiar máximo proposto por Drake e Nigrini (2000).



Dígitos	Frequência observada	Frequência esperada	Proporção observada	Benford	TesteZ	DMA	Qui-quadrado
10	88	60,4333	6,03%	4,14%	3,56	0,00021	12,57
11	58	55,1713	3,97%	3,78%	0,32	0,00002	0,15
12	42	50,7527	2,88%	3,48%	1,18	0,00007	1,51
13	56	46,9896	3,84%	3,22%	1,26	0,00007	1,73
14	25	43,7463	1,71%	3,00%	2,80	0,00014	8,03
15	34	40,9219	2,33%	2,80%	1,02	0,00005	1,17
16	40	38,4403	2,74%	2,63%	0,17	0,00001	0,06
17	25	36,2424	1,71%	2,48%	1,81	0,00008	3,49
18	29	34,2824	1,99%	2,35%	0,83	0,00004	0,81
Total	1460	1460,0000	100,00%			0,00272	154,30

Figura 3.13: *Dashboard* distribuição dos dois primeiros dígitos.

3.3.4 Regras de Associação

Para a utilização do algoritmo Apriori definimos o suporte igual a 0,001 e confiança de 0,40. Esses valores foram escolhidos com base nas frequências dos dígitos. Uma vez que a maior frequência entre os dígitos foi de 88 pagamentos que representa apenas 6% do banco de dados e que sabemos pela Lei de Benford que essa frequência tende a decrescer. Ao aplicar a técnica para os pares de dígitos não conformes pelo teste Z (10, 14, 24, 34, 45 e 51), 9 regras foram obtidas. Escolheu-se a medida *lift* para a ordenação das melhores regras por dígitos, por ser uma medida que testa a dependência de X e Y a partir das frequências que esses eventos ocorrem simultaneamente.

Tabela 3.10: Regras de associação para os dígitos não conformes.

Dígito	Regras	n	Suporte	Confiança	Lift	Convicção
10	CD_FORNECEDOR=871,PARCELADOS=Parcelado	2	0,0014	1,0000	16,5682	-
	CD_FORNECEDOR=2482,PRAZO=0	2	0,0014	0,6667	11,0455	2,8189
	CD_FORNECEDOR=5032	2	0,0014	0,4000	6,6273	1,5661
14	CD_FORNECEDOR=17777776	2	0,0014	1,0000	58,3200	-
45	CD_FORNECEDOR=690,PRAZO=1,PARCELADOS=Parcelado	2	0,0014	0,5000	33,1364	1,9698
	CD_FORNECEDOR=15555554,NM_SETOR=ALMOXARIFADO DA CAF,PRAZO=0,PARCELADOS=Único	4	0,0027	0,4000	26,5091	1,6415
51	CD_FORNECEDOR=413,PARCELADOS=Parcelado	2	0,0014	1,0000	47,0323	-
	CD_FORNECEDOR=409,ACRESCIMO=0	2	0,0014	0,6667	31,3548	2,9362
	CD_FORNECEDOR=5266,PRAZO=1	2	0,0014	0,5000	23,5161	1,9575

Nota-se na Tabela 3.10 que o par 10 obteve três regras de associação. Com um *lift* de 16,57 e uma confiança de 1 pode-se dizer que a ocorrência dos eventos do fornecedor **871** em conjunto com pagamentos parcelados implicam em uma dependência positiva quando os primeiros dígitos são 10, além disso também tem-se que 100% das vezes que esses eventos ocorreram ao mesmo tempo foi para esses dígitos. Vale destacar que para esse caso não é possível calcular a convicção porque a confiança é perfeita, ou seja, é igual a 1, o que faz o denominador da Equação 2.10 ser 0. A segunda regra para o par 10 sugere que os pagamentos efetuados dentro do prazo para o fornecedor **2482** tem uma dependência positiva com o dígito. Diferentemente da regra anterior temos a medida da convicção, que confirma uma relação de dependência positiva entre esses eventos e os dígitos. A terceira e ultima regra do par 10 diz que os pagamentos de medicamentos do fornecedor **5032** têm relação dependente e positiva para os dígitos. Dos 5 pagamentos feitos a essa empresa dois deles tem seus dois primeiros dígitos 10.

Obteve-se apenas uma regra para o par 14, indicando que todas os pagamentos para a empresa **17777776** tinham 14 como os dois primeiros dígitos, fazendo com que o *lift* dessa regra seja o mais alto. Por representar uma quantidade bastante pequena de pagamentos, essa regra pode ser considerada frágil e insuficiente para explicar a não conformidade com Benford. Esses pagamentos têm características de uma compra ocasional de medicamentos, fato confirmado pela não recorrência de compras para esse fornecedor.

Os dígitos 24 e 34 obtiveram regras. Em ambos os casos, os limites mínimos de suporte e confiança definidos não foram satisfeitos para o funcionamento do algoritmo Apriori. Para os dígitos 45 encontrou-se duas regras. A primeira delas refere-se aos pagamentos para o fornecedor **690** dentro do prazo e parcelados, implicando em uma dependência positiva com os dígitos 45, fato confirmado pelo *lift* e convicção sendo superiores a 1. A segunda regra desses dígitos foi a que apresentou a maior frequência dentre os comportamentos encontrados. Essa regra é originada pelo pagamentos efetuados pelo fornecedor **15555554** com destino ao almoxarifado do Centro de Abastecimento Farmacêutico (CAF) além disso seu valor não é fruto de um parcelamento. Pelas medidas de convicção e *lift*

temos que elas implicam em uma dependência positiva com os dígitos 45. Vale destacar que a confiança dessa regra é igual ao limite mínimo estipulado.

Por fim para os dígitos 51, cujo teste Z ficou mais distante da confiança adotada, três regras foram encontradas. Dentre elas destaca-se os pagamentos parcelados ao fornecedor **413** por ter o maior *lift*, além de ter a confiança de 100%. As outras duas regras são pagamentos dos fornecedores **409** e **5266** com algum valor de acréscimo na valor da nota fiscal e pagamento efetuado com atraso respectivamente. Essas últimas duas regras tem suas métricas com resultados parecidos. Em ambos os casos, convicção e *lift* foram maiores que um e suas frequências observadas foram iguais a dois.

A Figura 3.14 é uma exemplificação de como as regras de associação são apresentadas no software Power BI. Os tamanhos dos pontos variam de acordo com o suporte e a cor com o *lift*. Não existe ainda uma saída de dados em forma de tabela para o Power BI, portanto os resultados apresentados na Tabela 3.10 foram todos extraídos do software R. Os pacotes usados para gerar todos os resultados dessa Seção foram: *arules* e *arulesViz*.

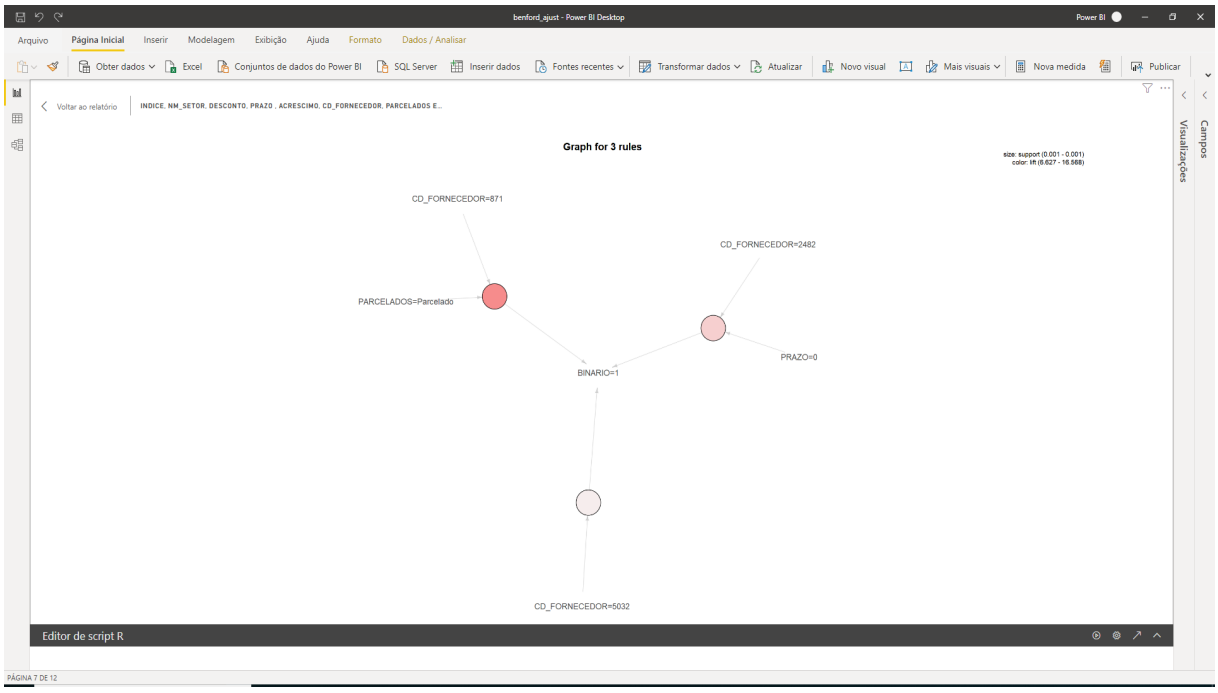


Figura 3.14: Gráfico das regras de associação no Power BI.

CAPÍTULO 4

CONSIDERAÇÕES FINAIS

4.1 Conclusão

Este trabalho apresentou a aplicação da lei de Benford e regras de associação no software Power BI como ferramentas de auxílio para auditorias internas. As aplicações foram feitas em dois conjuntos de dados reais que tem impacto direto na saúde financeira de uma empresa, por envolver milhões reais.

Ao aplicar a lei de Benford no faturamento da UTI e fazer os testes dos dígitos encontramos um cenário de total conformidade com as frequências esperadas. Apenas o par de dígitos 24 apresentou uma distorção significativa com Benford, considerando 95% de confiança. Um estudo foi feito sobre esse dígito, contudo nenhum indício da causa dessa pequena alteração foi encontrada.

Para o banco de dados dos pagamentos pertinentes a compras de medicamentos algumas disparidades entre as frequências esperadas e observadas foram encontradas. Ao avaliar a distribuição do segundo dígito pelo teste Z detectamos que apenas 1 dos 10 dígitos não se adéquam à lei. Todavia essa diferença não foi suficiente para alterar o cenário de conformidade pelos testes globais Qui-Quadrado e DMA. No teste dos dois primeiros dígitos seis casos se apresentaram como não conformes segundo o teste Z. Adicionalmente os resultados dos testes Qui-Quadrado e DMA apontaram um cenário de não conformidade. Viu-se que os valores encontrados provenientes das medidas de conformidade beiraram os limites aceitáveis de adequação com Benford. Além disso, ao fazer um estudo mais profundo para os dígitos que mais destoaram de sua frequência esperada, não se encontrou

nenhum indício de fraude. A técnica de mineração de dados conhecida como regras de associação também foi aplicada para os dígitos não conformes, com o intuito de encontrar algum evento fraudulento. Poucas regras foram encontradas, além disso todas elas tinham características de pagamentos ocasionais. Com isso, pode-se dizer que ambos os bancos de dados apresentaram resultados similares, não apresentando indícios de fraude.

Neste trabalho mostrou-se a possibilidade de uso de técnicas de auditoria em diferentes *softwares*, especificamente o Power BI e o R project. A aplicação no Power BI se mostrou bastante promissora por se tratar de uma ferramenta *self-service*, com uma maior dinâmica para construção de visuais e mudanças de contextos. Portanto, diante dos objetivos da auditoria interna o Power BI é uma boa ferramenta de auxílio para o controle da veracidade das informações contábeis e patrimoniais de uma empresa.

4.2 Trabalhos futuros

- Disseminação da aplicação desses métodos para outros setores e bancos de dados;
- Conexão direta em uma *Data Warehouse*(DW) para acompanhamento contínuo desses resultados via Power BI;
- Implementação de outras técnicas de auditoria no Power BI;
- Publicação de um artigo com esse trabalho.

REFERÊNCIAS BIBLIOGRÁFICAS

ABDALA, V. *PF faz operação contra fraudes na compra de medicamentos*. 2020. Disponível em: <https://agenciabrasil.ebc.com.br/geral/noticia/2020-01/pf-faz-operacao-contra-fraudes-na-compra-de-medicamentos>.

ADHIKARI, A.; SARKAR, B. Distribution of most significant digit in certain functions whose arguments are random variables. *Sankhyā: The Indian Journal of Statistics, Series B*, JSTOR, p. 47–58, 1968.

AGRAWAL, R.; IMIELIŃSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. In: *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*. [S.l.: s.n.], 1993. p. 207–216.

AGRAWAL, R.; SRIKANT, R. et al. Fast algorithms for mining association rules. In: *Proc. 20th int. conf. very large data bases, VLDB*. [S.l.: s.n.], 1994. v. 1215, p. 487–499.

BENFORD, F. The law of anomalous numbers. *Proceedings of the American philosophical society*, JSTOR, p. 551–572, 1938.

BORGELT, C.; KRUSE, R. Induction of association rules: Apriori implementation. In: SPRINGER. *Compstat*. [S.l.], 2002. p. 395–400.

DRAKE, P. D.; NIGRINI, M. J. Computer assisted analytical procedures using benford's law. *Journal of Accounting Education*, Elsevier, v. 18, n. 2, p. 127–146, 2000.

FERRARI, A.; RUSSO, M. *The Definitive Guide to DAX: Business Intelligence with Microsoft Excel, SQL Server Analysis Services, and Power BI*. [S.l.]: Microsoft Press, 2015.

GARTNER. *Magic Quadrant for Analytics and Business Intelligence Platforms*. 2020. Disponível em: <https://www.gartner.com/doc/reprints?id=1-3TXXSLV&ct=170221&st=sb>.

JAMAIN, A. Benford's law. *Unpublished Dissertation Report, Department of Mathematics, Imperial College, London*, 2001.

LAGO, K.; ALVES, L. *Dominando o Power BI*. [S.l.: s.n.], 2018. v. 1.

LEE, W.; STOLFO, S. Data mining approaches for intrusion detection. 1998.

LONGO, C. G.; JIMÉNEZ, A.; MARCOS, S. *Manual de Auditoria E Revisão de Demonstrações Financeiras: Novas Normas Brasileiras E Internacionais de Auditoria* . [S.l.]: Editora Atlas SA, 2000.

LUHN, H. P. A business intelligence system. *IBM Journal of research and development*, IBM, v. 2, n. 4, p. 314–319, 1958.

NEWCOMB, S. Note on the frequency of use of the different digits in natural numbers. *American Journal of mathematics*, v. 4, n. 1, p. 39–40, 1881.

NIGRINI, M. J. *Benford's Law: Applications for forensic accounting, auditing, and fraud detection*. [S.l.]: John Wiley & Sons, 2012. v. 586.

PALMISANO, Â.; ROSINI, A. M. *Administração de sistemas de informação e a gestão do conhecimento*. [S.l.]: Cengage Learning Editores, 2003.

PEREIRA, A. C.; NASCIMENTO, W. S. do. Um estudo sobre a atuação da auditoria interna na detecção de fraudes nas empresas do setor privado no estado de são paulo. *Revista Brasileira de Gestão de Negócios-RBGN*, Fundação Escola de Comércio Álvares Penteado, v. 7, n. 19, p. 46–56, 2005.

PINKHAM, R. S. On the distribution of first significant digits. *The Annals of Mathematical Statistics*, JSTOR, v. 32, n. 4, p. 1223–1230, 1961.

PORTELA, O. T.; SCHMIDT, A. S. Proposta de metodologia de avaliação e diagnóstico de gestão hospitalar. *Acta Paulista de Enfermagem*, Universidade Federal de São Paulo, v. 21, p. 198–202, 2008.

SFORSIN, A. C. P. et al. Gestão de compras em farmácia hospitalar. *Pharmacia Brasileira*, v. 16, n. 85, p. 1–30, 2012.

SILVA, C. V. S.; RALHA, C. G. Detecção de cartéis em licitações públicas com agentes de mineração de dados. Faculdade Cenecista de Campo Largo-Paraná-Brasil, 2011.

TEAM, R. C. et al. *R: A language and environment for statistical computing*. [S.l.]: Vienna, Austria, 2013.

WALLACE, W. A. Assessing the quality of data used for benchmarking and decision-making. *The Journal of Government Financial Management*, Association of Government Accountants, v. 51, n. 3, p. 16, 2002.

APÊNDICE A

IMPLEMENTAÇÃO COMPUTACIONAL - R

Uma vez que o pacote *benford.analysis* utilizado para análise da lei de Benford do software R não possui todos os testes usados nesse trabalho, criou-se uma nova função adaptando alguns desses comandos já existentes.

```
rm(list = ls())

# Pacotes ------------------------------------------------------------------
library(benford.analysis)


# DISTRIBUICAO PRIMEIRO E DOIS PRIMEIROS DIGITOS --------------------------

result_table <- function(x, digits = 2) {
  dados_bfd <-
    benford(data = x, number.of.digits = digits)$bfd[, c(1, 2, 4, 6, 7)]
  digitos <- length(dados_bfd$digits)
  N <- sum(dados_bfd$data.dist.freq)

  # Estatística Z
  po <- dados_bfd$data.dist # proporção observada
  pe <- dados_bfd$benford.dist# proporção esperada
  corr <- 1 / (2 * N) # correcao de continuidade
```

```
  num <-
    ifelse(abs(po - pe) > corr, abs(po - pe) - corr, abs(po - pe)) # numerador
  den <- sqrt((pe * (1 - pe)) / N) # denominador
  estz <- num / den
  estz
  nrejz <- sum(estz > 1.96)

  # Teste Qui-Qudrado
  FO <- dados_bfd$data.dist.freq
  FE <- dados_bfd$benford.dist.freq

  qui <- (FO - FE) ^ 2 / FE
  estqui <- sum(qui)

  # DMA
  DMA <- abs(po - pe) / digitos
  estDMA <- sum(DMA)

  # Consolidação dos resultados
  tabela <- cbind(dados_bfd, estz, qui, DMA)

  # Retorno
  return(list(
    tabela = tabela,
    DMA = estDMA,
    Qui = estqui,
    nrejz = nrejz
  ))

}
```

```
# DISTRIBUICAO DO SEGUNDO DIGITO  -------------------------
result_2digit <- function(x) {
  # Probabilidade segundo dígito
  prob <- c(0.11968,
            0.11389,
            0.10882,
            0.10433,
            0.10031,
            0.09668,
            0.09337,
            0.09035,
            0.08757,
            0.085
  )


  N <- length(x)
  dados_bfd <- as.data.frame(table(x))
  dados_bfd <- cbind(dados_bfd,N*prob,dados_bfd$Freq/N,prob)
  colnames(dados_bfd) <- c("digito","data.dist.freq","benford.dist.freq","data.dist
  digitos <- length(dados_bfd$digito)
  # Estatística Z
  po <- dados_bfd$data.dist # proporção observada
  pe <- dados_bfd$benford.dist# proporção esperada
  corr <- 1 / (2 * N) # correcao de continuidade

  num <-
    ifelse(abs(po - pe) > corr, abs(po - pe) - corr, abs(po - pe)) # numerador
  den <- sqrt((pe * (1 - pe)) / N) # denominador
  estz <- num / den
  estz
  nrejz <- sum(estz > 1.96)

  # Teste Qui-Qudrado
  FO <- dados_bfd$data.dist.freq
  FE <- dados_bfd$benford.dist.freq

  qui <- (FO - FE) ^ 2 / FE
```

```
  estqui <- sum(qui)

  # MAD
  MAD <- abs(po - pe) / digitos
  estMAD <- sum(MAD)

  # Consolidação dos resultados
  tabela <- cbind(dados_bfd, estz, qui, MAD)

  # Retorno
  return(list(
    tabela = tabela,
    MAD = estMAD,
    Qui = estqui,
    nrejz = nrejz
  ))

}




# GRAFICOS  ---------------------------------------------
dev.off()
layout(rbind(1,2), heights=c(7,1))  # put legend on bottom 1/8th of the chart

bp <-
  barplot(
    results_d1$data.dist ,
    col = ifelse(results_d1$estz > 1.96, "red", "grey"),
    names.arg = 1:9,
    ylab = "Frequência",
    xlab = "Digito",
    main = "Distribuição do primeiro dígito",
    ylim = c(0,0.30)
  )
lines(x = bp, results_d1$benford.dist, type="l", col = "red", lty=2, lwd=2)

par(mar=c(0, 0, 0, 0))
```

```
# c(bottom, left, top, right)
plot.new()
legend(
  "top",
  legend = c("Observada", "Benford"),
  xpd = TRUE,
  horiz = TRUE,
  pch = c(15, NA),
  col = c("grey", "red"),
  lty = c(0, 2),
  lwd = 2,
  title = NA,
  cex = 1,
  bty = 'n',
  pt.cex = c(3,1)
)
```