UNIVERSIDADE ESTADUAL DA PARAÍBA

UEPB

JOÃO FERNANDES DE ARAÚJO NETO

ESTATÍSTICA DESCRITIVA E TESTE QUI-QUADRADO APLICADOS A ACIDENTES DE TRÂNSITO OCORRIDOS EM RODOVIAS FEDERAIS NA PARAÍBA EM 2012

CAMPINA GRANDE – PB

2014

JOÃO FERNANDES DE ARAÚJO NETO

ESTATÍSTICA DESCRITIVA E TESTE QUI-QUADRADO APLICADOS À ACIDENTES DE TRÂNSITO OCORRIDOS EM RODOVIAS FEDERAIS NA PARAÍBA EM 2012

Trabalho de conclusão de curso apresentado ao Curso de Bacharelado em Estatística da Universidade Estadual da Paraíba – UEPB, como requisito para obtenção de título de graduação em Estatística, sob a orientação do Professor Edwirde Luiz Silva.

CAMPINA GRANDE – PB

2014

A659e Araújo Neto, João Fernandes de

Estatística descritiva e teste qui-quadrado aplicados a acidentes de trânsito ocorridos em rodovias federais na Paraíba em 2012 [manuscrito] / João Fernandes de Araújo Neto. - 2014.

28 p.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2014.

"Orientação: Prof. Dr. Edwirde Luiz Silva, Departamento de Estatística".

1. Estatística descritiva. 2. Acidentes de trânsito. 3. Teste qui-quadrado. I. Título.

21. ed. CDD 519.5

JOÃO FERNANDES DE ARAÚJO NETO

ESTATÍSTICA DESCRITIVA E TESTE QUI-QUADRADO APLICADOS À ACIDENTES DE TRÂNSITO OCORRIDOS EM RODOVIAS FEDERAIS NA PARAÍBA EM 2012

Trabalho de conclusão de curso apresentado ao Curso de Bacharelado em Estatística da Universidade Estadual da Paraíba – UEPB, como requisito para obtenção de título de graduação em Estatística, sob a orientação do Professor Edwirde Luiz Silva.

APROVADO: 28 / 02 / 2014

Banca Examinadora

Edwirde Luiz Silva

Professor Edwirde Luiz Silva
Orientador

João Luna

João Gil de Luna
Examinador (a)

Tiago Almeida de Oliveira

Tiago Almeida de Oliveira
Examinador (a)

CAMPINA GRANDE – PB

2014

AGRADECIMENTOS

Agradeço primeiramente a Deus, que me deu força, determinação e perseverança para atingir meus objetivos.

Às mulheres de minha vida, Ednalva, Ana Beatriz e Ana Flávia, esposa e filhas, respectivamente, pela paciência, incentivo e apoio nos momentos difíceis.

Aos meus pais, Rivaldo e Fátima e meus irmãos, Edilma e Rivaldo Filho, pelo companheirismo e incentivo durante toda a jornada.

Aos meus demais familiares, que sempre estiveram ao meu lado.

Ao meu professor e orientador Edwirde, pela confiança, incentivo e ajuda, essenciais para o êxito deste projeto.

Aos meus demais professores, que souberam transmitir todo o conhecimento necessário para que eu pudesse chegar aqui.

À Polícia Rodoviária Federal, nas pessoas de seus dirigentes atuais, Luciana e Aurivan, pela concessão dos dados utilizados neste projeto e pelo apoio dispensado durante todo curso.

Enfim, obrigado a todos que de maneira direta ou indireta contribuíram para essa conquista.

RESUMO

Neste trabalho é apresentado um estudo estatística descritiva, com uso de tabelas, cálculo de média e desvio padrão e o teste de qui-quadrado, aplicados ao número de acidentes, cujo objetivo foi realizar um levantamento dos acidentes de trânsito ocorridos nas rodovias federais que passam na região de Campina Grande/PB, no ano de 2012, com dados reais fornecidos pela Polícia Rodoviária Federal. Este trabalho busca evidenciar dados dos acidentes, visando analisar maiores incidências de algumas características, como dia da semana de ocorrência, mês de ocorrência, classificação dos acidentes (gravidade), dentre outros, e se há relação entre alguns desses fatores. Esses dados podem ser utilizados em estudo de ações voltadas para redução de acidentes pela Polícia Rodoviária Federal.

Palavras-chaves: Estatística descritiva, teste qui-quadrado, acidentes.

ABSTRACT

This paper presents a study descriptive statistics, using tables, calculation of mean and standard deviation and chi-square test applied to the number of accidents, whose purpose was to conduct a survey of traffic accidents occurred on federal highways that pass area of Campina Grande / PB, in 2012, with real data provided by the Federal Highway Police. This paper seeks to show data of accidents in order to analyze larger impact of some features, such as day of the week of occurrence, month of occurrence, classification of accidents (gravity), among others, and if there is some relationship between these factors. These data can be used in the study of actions aimed at reducing accidents by the Federal Highway Police.

Keywords: Descriptive statistics, chi-square, accidents.

SUMÁRIO

1 INTRODUÇÃO

A estatística descritiva é o ramo da estatística que coleta, resume e apresenta dados, para uma visualização clara, possibilitando analisar os dados e identificar características importantes. Assim, utiliza-se técnicas descritivas (tabelas e gráficos), tanto para variáveis numéricas, não utilizadas neste trabalho, quanto para variáveis categóricas (BARBETTA, 2012).

A importância da estatística nos dias de hoje é muito grande, pois ela auxilia no nascer, crescimento e manutenção das instituições, ajudando no monitoramento dos processos, para que se tomem decisões de correção e adaptação desses processos (BARBETTA, 2012).

A maioria das analises estatísticas são feitas usando uma biblioteca de programas estatísticos, escritos a priori. Apresentam-se algumas saídas de vários pacotes estatísticos em todo trabalho. Não se discute a facilidade de uso dos pacotes com relação à entrada e a edição de dados ou ao uso dos comandos. (MONTGOMERY, 2003).

Nas rodovias federais de todo país, ocorreram, no ano de 2012, 184.503 acidentes, gerando um total de 8.661 mortos e 104.385 feridos. Além das perdas emotivas, esses acidentes geram um custo financeiro muito grande. No geral, as rodovias federais têm uma boa conservação, estando, em sua maioria, em bom estado, não sendo fator determinante para a ocorrência de acidentes.

Estudos preliminares indicam que a imprudência é a principal causa de acidentes no país. Podem-se incluir em imprudência vários comportamentos dos condutores e pedestres, como por exemplo: ingestão de bebidas alcoólicas, excesso de velocidade, desrespeito à sinalização e outras.

Um estudo do número de acidentes é de grande importância, em virtude dos inúmeros casos diários que ocorrem. Muitas vidas são ceifadas e muitos feridos em acidentes ocupam leitos nos hospitais, causando grandes prejuízos morais, sentimentais e econômicos, o que requer um estudo mais aprofundado, na busca de se evitar tais acidentes.

Nesse estudo, serão utilizadas técnicas de análise descritiva, com uso de tabelas, cálculo de medidas de posição, de dispersão, e por fim, será utilizado o teste de qui-quadrado de Pearson, para estudo de dados de acidentes de trânsito ocorridos nas rodovias federais que passam na região de Campina Grande – PB, ocorridos no ano de 2012.

As medidas de posição e de dispersão podem resumir os dados, fornecendo informações importantes, de modo a se tomar decisões com mais segurança.

Já o teste de qui-quadrado de Pearson possibilita comparar os dados e definir se eles diferem estatisticamente ou não, ajudando também na tomada de decisões.

A finalidade deste trabalho é mostrar a descrição do tipo de acidentes, encontrar a média mensal de acidentes, seu desvio padrão e variância, relacionar os dias da semana com a classificação de acidentes e verificar se existe diferenciação estatística entre a quantidade de acidentes mês a mês, relativos a acidentes ocorridos nas rodovias federais da região de Campina Grande/PB. Será que existe alguma relação entre as variáveis apresentadas?

Este estudo verificará se existe alguma diferença na quantidade de acidentes envolvendo veículos em relação aos dias da semana e entre os meses do ano. Se houver diferenças na quantidade de acidentes nestas rodovias, pode significar que durante os dias da semana ou mês de ocorrência, deve haver uma diferenciação entre a abordagem de atuação policial para prevenir acidentes.

2. OBJETIVO

a) Mostrar de forma descritiva a ocorrência de acidentes por dia da semana;
b) Relacionar a ocorrência de acidentes por dia da semana com a classificação de acidentes;
c) Verificar se existe relação entre o dia da semana e a classificação de acidentes;
d) Verificar se as quantidades de acidentes ocorreram de maneira igual nos meses do ano.

3. FUNDAMENTAÇÃO TEÓRICA

3.1 Estatística Descritiva

3.1.1 Distribuição de Frequência

De início, deve-se construir uma distribuição de frequências, que é a organização dos dados de acordo com as ocorrências dos diferentes resultados, conforme exemplo abaixo.

Tabela 1. Distribuição de frequências da Variável X

Variável X	**Frequência**	**Percentagem**
X_1	x_1	p_1
X_2	x_2	p_2
. . .	. . .	. . .
X_n	x_n	p_n
Total	Σx_n	100,00

Onde a primeira coluna mostra as categorias da variável X. A segunda coluna mostra as frequências observadas e a terceira coluna representa uma medida relativa das frequências observadas, que são muito importantes para comparação da distribuição.

Pode-se ter também uma tabela de representação conforme a Tabela 2, onde são listados dados conjuntos das variáveis X e Y.

Tabela 2. Distribuição de frequências conjuntas das variáveis X e Y.

Variável X	$\mathbf{Y_1}$	$\mathbf{Y_2}$	**. . .**	$\mathbf{Y_m}$	**Total**
X_1	x_1y_1 (p_{11})	x_1y_2 (p_{11})	. . .	x_1y_m (p_{1m})	Σx_1y_m (Σp_{1m})
X_2	x_2y_1 (p_{21})	x_2y_2 (p_{21})	. . .	x_2y_m (p_{2m})	Σx_2y_m (Σp_{2m})
. . .	. . .	. . .	. . .	. . .	. . .
X_n	x_ny_1 (p_{n1})	x_ny_2 (p_{n1})	. . .	x_ny_m (p_{nm})	Σx_ny_m (Σp_{nm})
Total	Σx_ny_1 (100)	Σx_ny_2 (100)	. . .	Σx_ny_m (100)	$\Sigma x_n\Sigma y_m$ (100)

Nota: Números entre parênteses correspondem à percentagem em relação ao total da coluna.

Este tipo de tabela pode ser chamada de Tabela de Contingência ou de Dupla Entrada, devido a seu formato. Assim distribuídos, os dados fornecem muitas informações sobre as características dos dados, pois as percentagens fornecem uma visualização mais fácil de entendimento dos dados. (BARBETTA, 2012).

Os totais das linhas e das colunas também podem fornecer informações importantes, pois mostram informações sobre as variáveis se isolarmos uma das características da outra variável.

3.1.2 Medidas de Posição

Quando se tem um conjunto de dados conforme descrição abaixo, onde se tem uma série de valores quantitativos observados, extrai-se algumas informações que podem resumir esse conjunto de dados, a exemplo de medidas de posição e de dispersão.

Existem várias medidas de posição que podem resumir um conjunto de dados. Dentre as quais, pode-se citar a média aritmética, a mediana, a moda e outras. Nesse trabalho, vamos utilizar a média aritmética, a mais usada e conhecida delas, e o gráfico de boxplot.

Pode-se definir a média aritmética, que representa uma posição central ou um valor típico dos dados, como sendo a somatória dos valores dos dados dividida pela quantidade de valores observados, ou seja:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$, onde n é o número de observações. (1)

O boxplot é uma representação gráfica baseada no resumo de cinco números de um conjunto de dados, que corresponde a:

X_{menor}, Q_1, Mediana, Q_3 e X_{maior}.

Onde X_{menor} é o menor valor não outlier (valor atípico) do conjunto de dados, X_{maior} é o maior valor não outlier, Q_1 e Q_3 são, respectivamente, o primeiro e o terceiro quartis e a Mediana do conjunto de dados, que representa o valor central dos dados, ou seja, divide os dados ao meio. (BUSSAB, MORETTIN, 2002).

O gráfico boxplot pode ser representado pela figura 1, abaixo:

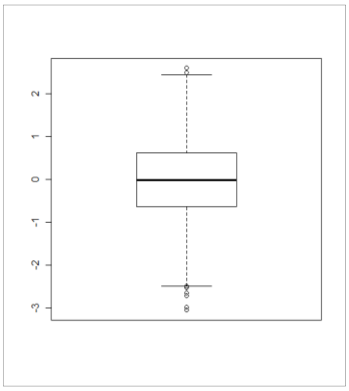


Figura 1: Diagrama esquemático do boxplot.

Para confeccionar o boxplot, é necessário calcular os valores resumo Q_1, Mediana e Q_3. Assim, tem-se que definir os quartis e a mediana.

O primeiro quartil, Q_1, representa o valor do conjunto de dados que delimita os 25% valores inferiores desse conjunto. É expresso por:

$Q1 = \frac{(n+1)}{4}$, valor na ordem de classificação dos dados (2)

A mediana é expressada por:

A mediana, $Med = \frac{n+1}{2}$, valor na ordem de classificação dos dados. (3)

O terceiro quartil, Q_3, representa o valor que delimita os 25% maiores valores dos dados. É expresso por:

$Q3 = \frac{3(n+1)}{4}$, valor na ordem de classificação dos dados (4)

onde n é o número de dados do conjunto.

Os valores de X_{maior} e X_{menor} são, respectivamente, os menores e os maiores valores do conjunto de dados que não são considerados outlier. Para se definir os valores limites de outlier, usa-se as expressões abaixo:

$L_{Oi} = Q_1 – 1,5(Q_3-Q_1)$ (5)

$L_{Os} = Q_3 + 1,5(Q_3-Q_1)$ (6)

Quando o conjunto de dados possui valores que estejam fora dos limites L_{Oi} e L_{Os}, esses valores são representados no boxplot por pontos isolados.

As medidas de dispersão, como seu próprio nome diz, representam a dispersão dos dados em relação à média aritmética. Nesse estudo, serão utilizados o desvio padrão e a variância.

No caso do desvio padrão amostral, que é uma medida de dispersão dos dados, primeiramente calcula-se a variância, que é um valor que representa os “desvios médios quadráticos” em relação à média (BARBETTA, 2012). Então, a variância amostral pode ser expressa por:

$$S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1} \text{, onde n é o número de observações.} \quad (7)$$

O desvio padrão pode ser definido como a raiz quadrada positiva da variância, sendo expressado por:

$$S = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}} \quad (8)$$

A média, a variância e o desvio padrão são as principais medidas representativas de uma população ou amostra de dados, podendo fornecer informações importantes sobre essa população ou amostra.

3.2 Teste de hipótese qui-quadrado

Em estatística, um resultado com probabilidade de ocorrência igual ou inferior a 5% é considerado como tendo pouca probabilidade de acontecer. Com base nisso, estabeleceu-se que, com base em um único conjunto de dados pode-se chegar a uma conclusão a respeito da casualidade ou não dos desvios entre os valores observados e esperados. Em outras palavras, economizando tempo e trabalho, pode-se concluir se os desvios entre os valores observados e esperados devem ter alguma significação ou se eles podem ser considerados como ocorridos casualmente.

Os testes não paramétricos são particularmente úteis para decisões sobre dados oriundos de pesquisas da área de ciências humanas. Para aplicá-los, não é necessário admitir hipóteses sobre distribuições de probabilidades da população da qual tenham sido extraídas amostras para análises. As provas não paramétricas são prioritariamente adaptáveis aos estudos que envolvem variáveis com níveis de mensuração nominal e ordinal, bem como à investigação de pequenas amostras, como apresentada neste trabalho (MARTINS, 2011).

As provas não paramétricas são também denominadas provas livres de distribuição, pois para aplicá-las não é necessário fazer suposições quanto ao modelo de distribuição de probabilidade da população. Esses testes são recomendados para análises de resultados de experimentos com dados emparelhados – do tipo antes e depois – , para verificar se variáveis são independentes ou relacionadas, e também para o tratamento estatístico de dados oriundos de tabelas de com dupla entrada. Em outras palavras, os testes não paramétricos são boas opções para situação em que ocorre violação dos pressupostos básicos necessários para a aplicação de um teste paramétrico.

3.2.1 Nível de Significância

O risco expresso que o pesquisador incorre em rejeitar uma hipótese verdadeira é denominado nível de significância e, geralmente, simbolizado pela letra grega alfa (α). Esse nível, portanto, deve ser sempre estabelecido antes da análise dos dados e é usualmente fixado em 5%, valor este considerado neste trabalho.

3.2.2 Graus de Liberdade

A situação de falta de independência entre os valores observados e os esperados leva á introdução do conceito de grau de liberdade do qui-quadrado, χ^2, que pode ser expresso como o número de informação da amostra que são necessárias aos cálculos dos valores esperados nessas classes.

As figuras 2 e 3 mostram dois modelos de densidade da função qui-quadrado com 8 gl e 50 gl. O algoritmo para gerar tais figuras se encontram no apêndice 1.

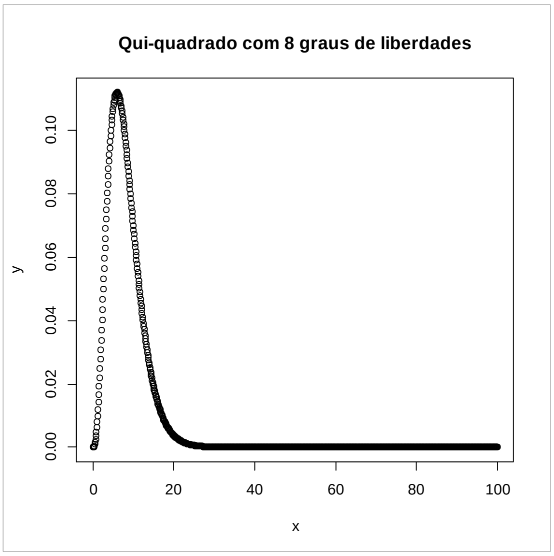


Figura 2. Distribuição qui-quadrada com 8 graus de liberdade.

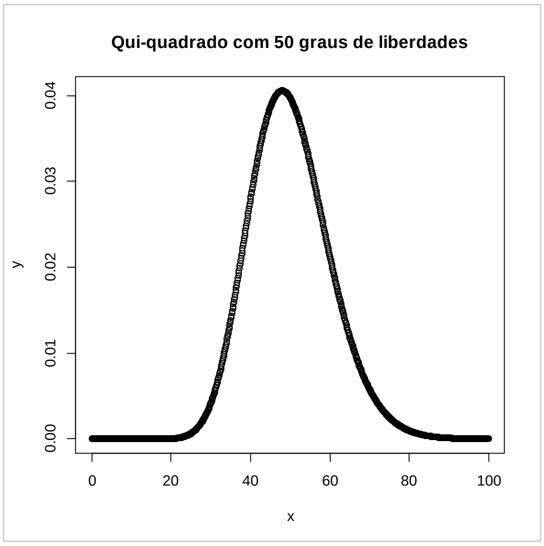


Figura 3. Distribuição qui-quadrada com 50 graus de liberdade.

Observa-se que quanto mais aumenta o grau de liberdade mais a direita se localiza a "onda".

3.3 Teste de Aderência

Quando se deseja realizar um teste estatístico para verificar se há adequação de ajustamento entre as frequências observadas e as frequências esperadas (distribuição uniforme). Isto é, se as discrepâncias ($F_{oi} - F_{ei}$), i = 1, 2 , ... , k são devidas ao acaso, ou se de fato existe diferença significativa entre as frequências.

O Contraste de hipótese da qui-quadrada de bondade de ajuste pode ser utilizado para esse estudo. Este tipo de contraste verifica se os dados se ajustam razoavelmente a uma família de distribuição, no caso, a distribuição uniforme (ÁVILA, 2006).

Seja X = (X_1, ...X_n) uma massa aleatória de uma variável aleatória X com distribuição Fo. Pretende-se contrastar:

H_0: As duas variáveis são independentes.

H_1: Existe associação entra as duas variáveis.

O ocorrido da distribuição populacional divide-se em k conjuntos mutuamente excludentes e exaustivos, $A_1, A_2, \ldots, A_k$ ($A_i \cap A_j = \emptyset$, $i \neq j$, $U^k_{i=1} A_k = \Omega$). Seja p_i a probabilidade associada a cada A_i sob H_o

$$p_i = P_{Ho}[X \in A_i],\ i = 1, \ldots, k \tag{9}$$

Tem-se que $\sum_{i=1}^{k} pi = 1$. Seja O_i a variável aleatória que conta o número de observações da amostra que estão no subconjunto A_i, i =1,...,k. Estas variáveis são conhecidas como frequências observadas. Tem-se que $\sum_{i=1}^{k} O = n$. O número de observações em cada subconjunto A_i que poderia esperar sob H_o expressa-se $E_i = np_i$, frequência esperadas. Obviamente a variável aleatória ($O_i, \ldots, O_{k-1}$) tem uma distribuição multinominal de parâmetros n e ($p_i, \ldots, p_{k-1}$). Observa-se que a variável aleatória $O_k = n - \sum_{i=1}^{k} O_i$ e a probabilidade $p_k = 1 - \sum_{i=1}^{k-1} p_i$.

A ideia então consiste em substituir o contraste não paramétrico proposto anteriormente pelo contraste de hipótese nula $H_o = p_1 = p_1', \ldots, p_k = p_k'$ versus a alternativa de que alguma das igualdades não seja certa. Estes contrastes se realizará com base nos valores ($O_1, \ldots, O_k$).

Este problema foi elaborado por Pearson, que introduz como medida de discrepância entre frequências observadas e esperadas a soma das diferenças:

$$D = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i = \sum (O_i - np_i)^2 / np_i \tag{10}$$

A intuição de Pearson ao eleger esta medida e, em particular, os mencionados fatores de ponderação, não apenas permite determinar a distribuição no amostra de D senão que proporcionar valores muito próximo aos que obtêm ao utilizar o método da razão de verossimilhança para resolver este problema. O seguinte resultado mostra esta distribuição. (ÁVILA, 2006).

Teorema 1. Seja (X_1, X2, …, Xk-1) um vetor aleatório com distribuição multinominal M(n, p_1, …, p_{k-1}). Então:

$$Y_k = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \sim \chi^2_{k-1} \quad (11)$$

Demonstração

Para demonstrar este resultado tem-se que considerar o caso $k = 2$. Em tal caso:

$$Y_2 = \sum_{i=1}^{2} \left[\frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} \right]$$

$$= \frac{(X_1 - np_1)^2}{np_1} + \frac{[(n - X_1) - n \cdot (1 - p_1)]^2}{n \cdot (1 - p_1)}$$

$$= (X_1 - np_1)^2 \cdot \left[\frac{1}{np_1} + \frac{1}{n \cdot (1 - p_1)} \right] = \frac{(X_1 - np_1)^2}{np_1 \cdot (1 - p_1)}$$

E, posto que $X_1 \sim B(n, p_1)$, mediante o teorema do limite central, obtém que, efetivamente $Y_2 \sim \chi^2_{k-1}$

O teorema 1 permite contrastar a hipótese nula H_0: $p_1' = p_2' = \ldots = p_k'$ ao utilizar

$$D = \sum_{i=1}^{k} \left[\frac{(O_i - np_i)^2}{np_i} \right] \sim \chi^2_{k-1} \quad (12)$$

De forma que, para n suficiente grande, rejeita-se H_0 a um nível de significância α para valores grandes da estatística D.

4 MATERIAIS E MÉTODOS

Este estudo analisa variáveis qualitativas, que são variáveis cujos resultados são observados na forma de categorias. Tais dados podem ser organizados e estudados na forma de tabelas e/ou gráficos.

Contatos preliminares com a Polícia Rodoviária Federal, serviram de norte para a escolha das variáveis a serem estudadas, para ajudar na tomada de decisões administrativas visando prevenir a ocorrência de acidentes.

Serão estudados a quantidade de acidentes por dia da semana, a classificação dos acidentes (sem vítimas, com feridos, com mortos e ignorado), quantidade de acidentes por mês de ocorrência.

Na região de Campina Grande/PB, existem quatros rodovias federais, nas quais os acidentes ocorrem, a citar: BR 104, BR 110, BR 230 e BR 412, rodovias essas que são atendidas pela Polícia Rodoviária Federal.

5 APLICAÇÕES E DISCUSSÕES

Em nosso estudo, utilizaremos dados de acidentes de trânsito ocorridos nas rodovias federais que passam na região de Campina Grande – PB, no ano de 2012, com dados fornecidos pela Polícia Rodoviária Federal, por intermédio da Delegacia PRF de Campina Grande – PB. Algumas variáveis, devido a sua importância, serão estudadas, a citar:

a) Dia da semana dos acidentes;

b) Classificação do acidente: com feridos, com mortos, sem vítimas e ignorado;

c) Mês de ocorrência do acidente;

Inicialmente analisaremos os dados fornecidos pela Tabela 3, que mostra a quantidade de acidentes de acordo com o dia da semana.

Tabela 3. Frequência do número de acidentes em relação ao dia da semana.

Dia	Frequência	Percentual	Freq. Acumulada
Domingo	**155**	**19,1**	19,1
Segunda	131	16,2	35,3
Terça	86	10,6	45,9
Quarta	79	9,7	55,6
Quinta	100	12,3	67,9
Sexta	125	15,4	83,4
Sábado	**135**	**16,6**	100,0

Pode-se verificar que nos finais de semana há uma maior incidência de acidentes, como se observa na tabela 3. Isso se deve ao fato de que nos finais de semana o número de condutores imprudentes aumenta. O domingo apresentou a maior incidência de acidentes, 155 no total, representando 19,1% dos acidentes. Já a quarta feira, com 9,7% dos acidentes, foi o dia com menor incidência.

Tabela 4. Classificação dos acidentes versus dia da semana

		Classificação de acidentes				Total
		Sem vítimas	Com feridos	Com mortos	Ignorado	
Dia	Domingo	62	**75**	**18**	0	155
	Segunda	68	54	9	0	131
	Terça	52	28	5	1	86
	Quarta	37	34	8	0	79
	Quinta	58	38	4	0	100
	Sexta	59	51	**13**	2	125
	Sábado	68	**55**	7	5	135
Total		404	335	64	8	811

Na tabela 4 observa-se que sábado, domingo e segunda apresentam os maiores percentuais de acidentes. Os acidentes com mortos se observam mais no domingo e sexta feira, enquanto que os acidentes com feridos ocorrem mais nos domingos e sábados.

Nota-se que, dos acidentes em que houve mortos, 28,1% ocorreram no dia de domingo, contra apenas 6,2% na quinta-feira.

Verifica-se também que as maiores quantidades de acidentes ocorreram nos fins de semana, com mais incidência nos domingos e sábados, respectivamente. Pode-se observar também que ocorrem muitos acidentes na sexta e segunda feiras.

Tais incidências podem ser justificadas por conta do aumento de ingestão de bebidas alcoólicas nos finais de semana, que acarretam, por consequência, aumento na imprudência dos motoristas.

Dar-se-á continuidade ao estudo aplicando-se as teorias de média aritmética e variância (desvio padrão) nos dados relativos ao quantitativo de acidentes classificados segundo o mês de ocorrência, de acordo com os dados da Tabela 5:

Tabela 5: Quantidades de acidentes classificados por mês de ocorrência.

Classificação de Acidentes	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez	Total
Freq. Observada	62	56	59	72	63	93	73	65	52	80	60	76	811

Temos que a média aritmética é dada por:

$\bar{X} = \frac{\sum X}{n}$, onde n é o número de observações.

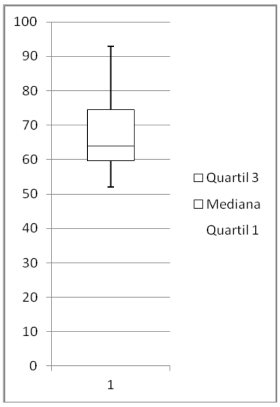
Então, tem-se que:

$\overline{X}$ = 811/12 = 67,58

Assim, pode-se afirmar que a média mensal de acidentes nas rodovias federais da região de Campina Grande – PB é 67,58. Comparando esse valor com os dados da Tabela 6, verificamos que o mês de junho foi o mês em que ocorreu a quantidade mais discrepante do valor médio. Tal discrepância pode ser justificada pela ocorrência de período festivo na região durante todo o mês e também pela incidência de chuvas no período, que aumenta o risco de acidentes.

Outra forma de analisar os dados é por meio do gráfico boxplot. Para tanto, utilizando-se os dados da tabela 5, tem-se que:

Limite inferior outlier	37,62
Menor valor não outlier	52
Quartil 1	59,75
Mediana	64
Quartil 3	74,5
Maior valor não outlier	93
Limite superior outlier	96,62



Ao se observar o boxplot, verifica-se que não houve nenhum valor outlier, ou seja, fora dos limites estatísticos, bem como se verifica uma não uniformidade entre os dados, pois a diferença entre o primeiro quartil e a mediana é bem menor que a diferença entre a mediana e o terceiro quartil.

Ao se comparar a média, a mediana e os quartis, verifica-se que a média aponta um valor mais central entre os quartis, podendo ser considerada como uma medida melhor de centralidade.

Para se ter uma indicação mais precisa da real diferença entre as quantidades mensais de acidentes, pode-se calcular a variância e o desvio padrão dos dados. Assim, tem-se que:

$S^2 = \dfrac{\sum\left(X - \overline{X}\right)^2}{n-1}$, onde n é o número de observações.

S² = [(62 - 67,58)² + (56 - 67,58)² + ... + (76 – 67,58)²] / 11 = 135,17

Em consequência, tem-se que o desvio padrão será calculado por:

$$S=\sqrt{\frac{\sum\left(X-\overline{X}\right)^2}{n-1}} \Rightarrow S=\sqrt{S^2} \Rightarrow S=\sqrt{135,} \Rightarrow S=11,6$$

O desvio padrão indica que existe uma variação padrão entre os dados de acidentes mensais de 11,62 acidentes. Isso pode indicar uma diferenciação muito grande na ocorrência de acidentes mensalmente.

Aplica-se o teste qui-quadrado de Pearson nos dados das tabelas 6 e 7, para verificar se existe relação entre a classificação dos acidentes e o dia da semana.

Tabela 6. Classificação dos acidentes versus dia da semana (valores observados).

		Classificação de acidentes				Total
		Sem vítimas	Com feridos	Com mortos	Ignorado	
Dia	Domingo	62	**75**	**18**	0	155
	Segunda	68	54	9	0	131
	Terça	52	28	5	1	86
	Quarta	37	34	8	0	79
	Quinta	58	38	4	0	100
	Sexta	59	51	**13**	2	125
	Sábado	68	**55**	7	5	135
Total		404	335	64	8	811

Tabela 7. Classificação dos acidentes versus dia da semana (valores esperados).

		Classificação de acidentes				Total
		Sem vítimas	Com feridos	Com mortos	Ignorado	
Dia	Domingo	77,21	64,03	12,23	1,53	155
	Segunda	65,26	54,11	10,34	1,29	131
	Terça	42,84	35,52	6,79	0,85	86
	Quarta	39,35	32,63	6,23	0,78	79
	Quinta	49,82	41,31	7,89	0,99	100
	Sexta	62,27	51,63	9,86	1,23	125
	Sábado	67,25	55,76	10,65	1,33	135
Total		404	335	64	8	811

Para tanto, teremos as seguintes hipóteses a serem analisadas:

H_o : As quantidades de acidentes nos dias da semana são iguais.

H_1 : As quantidades de acidentes nos dias da semana são diferentes.

Para um nível de confiança de 95%.

$\chi^2 = \sum\left[\frac{(f_0 - f_e)^2}{f_e}\right]$, com *k* graus de liberdade. onde f_e é a frequência esperada e k = (l-1).(c-1).

$\chi^2 = [(62 – 77,21)^2/77,21 + (75 – 64,03)^2/64,03 + . . . + (5 – 1,33)^2/1,33]$

$\chi^2 = 33,78$ com gl=18.

Com esse valor de χ^2 , com gl=18, tem-se que o valor $p = 0,013 < 0,05$.

Esses valores indicam que se deve rejeitar H_o, em favor de H_1, ou seja, existe relação entre a classificação de acidentes (gravidade) com os dias da semana. Isso pode conduzir a uma análise mais aprofundada desses dados, podendo direcionar as ações de prevenção de acidentes para atuações diferentes em relação aos dias da semana.

Por fim, será utilizado o teste de qui-quadrado de Pearson nos dados da tabela 6, para aferir se há diferença entre a quantidade de acidentes em relação ao mês de ocorrência. Nesse caso, não há uma comparação entre duas categorias, mas sim, se verifica se há diferença dentro da própria categoria. Os cálculos são feitos de maneira semelhante, apenas o cálculo do grau de liberdade é que é diferente, sendo definido como o número de observações menos um, ou seja, k – 1.

Nesses termos, tem-se as seguintes hipóteses a ser analisadas:

H_o : As quantidades de acidentes ocorreram de maneira igual nos meses do ano.

H_1 : As quantidades de acidentes ocorreram de maneira diferente nos meses do ano.

Para um nível de confiança de 95%.

Portanto tem-se:

$\chi^2 =$, $\sum\left[\frac{(f_0 - f_e)^2}{f_e}\right]$ com *k-1* graus de liberdade, onde f_e é média aritmética e k é o nº de observações.

$\chi^2 = [(62 – 67{,}58)^2/67{,}58 + (56 – 67{,}58)^2/67{,}58 + . . . + (76 – 67{,}58)^2/67{,}58]$

$\chi^2 = 22{,}001$ com gl = 11.

Com esse valor de χ^2 , com gl=11, tem-se que o valor $p = 0{,}0243 < 0{,}05$.

Assim, rejeita-se H_o em favor de H_1, e se pode afirmar que os acidentes ocorrem de maneira diferente nos meses do ano, ou seja, há diferenciação estatística na quantidade de acidentes nos diferentes meses do ano.

6 CONCLUSÃO

De acordo com os resultados da pesquisa, teste de qui-quadrado, pode-se verificar que o número de acidentes nas rodovias são diferentes a cada dia da semana, em relação à classificação dos acidentes (gravidade). Existe relação entre o número acidentes classificados de acordo com a gravidade e o dia. Em outras palavras, o teste poderia concluir que os "os dias da semana tem alguma relação com quantidade de acidentes classificados pela gravidade, provavelmente nos finais de semana".

Quanto ao dia da semana e a classificação de acidentes mostrou-se que no sábado, domingo e segunda (muitos pela madrugada) apresentam maiores percentuais de acidentes. Os acidentes com mortos ocorrem mais no domingo e sexta feira, enquanto que os acidentes com feridos ocorrem mais no domingo, sábado e segunda.

A análise do gráfico de boxplot indica que os dados mensais de acidentes se mostraram como sendo uma distribuição assimétrica à direita, apesar de não ter apresentado nenhum dado atípico (outlier).

Feita a análise em relação ao número de acidentes nos meses do ano, o teste qui-quadrado provou que realmente existe diferença na ocorrência de acidentes nos diferentes meses do ano.

Os simples resultados desta análise podem ser considerados como norteadores para as autoridades, diga-se Polícia Rodoviária Federal, no sentido de planejamento de suas operações visando o combate aos atos ilícitos que resultem na ocorrência de acidentes nas rodovias, assim como serve para direcionar as estratégias no sentido de conhecer os dias, os tipos de acidentes etc, tornando possível aprimorar a fiscalização com intuito de reduzir os percentuais de acidentes nas rodovias federais.

Não se pretende exaurir o assunto e sim dar um direcionamento a futuras pesquisas voltadas para conhecimento das causas de acidentes nas rodovias federais brasileiras.

7 REFERÊNCIAS

ÁVILA, M.J.M. Estatística Matemática. Espanha: Grupo editorial universitário, 2006.

BARBETTA, Pedro Alberto, Estatística Aplicada às Ciências Sociais, 8ª ed, Florianópolis/SC, Editora UFSC, 2012.

BUSSAB, Wilton de O., MORETTIN, Pedro A., Estatística Básica, 5ª ed, São Paulo/SP, Editora Saraiva, 2002.

COSTA NETO, Pedro Luiz de oliveira, 1939 – Estatística – São Paulo, Edgard Blucher, 1977.

DANCEY, Christine P. Estatística sem matemática para psicólogo. Porto Alegre: Armed, 2006.

MARTINS, G.A; DOMINGUES, O. Estatística Geral e Aplicada. 4ª Ed revisada e Ampliada. São Paulo: Editora Atlas, 2011.

MONTGOMERY, D.C; RUNGER, G.C. Estatística aplicada à probabilidade para engenheiros. 2. Ed. Rio de Janeiro: LTC, 2003.

PEREIRA, Julio Cesar R. Análise de Dados Qualitativos: Estratégias metodológicas para as Ciências da Saúde. Humanas e Sociais. 3. ed. São Paulo: Editora da Universidade de São Paulo, 2001.

STEPHAN, Levine, BERENSON, Krehbiel, Estatística – Teoria e Aplicações, 5ª ed, Rio de Janeiro/RJ, Editora LTC, 2008.

APÊNDICE

Gerando números aleatórios qui-quadrado:

```
x<-100*seq(0,1, length=1000)    # sequencia de 1000 entre 1 e 100
y<-dchisq(x,8)   # densidade da qui-quadrada com 8 graus de liberdade
y<-dchisq(x,50)   # densidade da qui-quadrada com 50 graus de liberdade
plot(x, y, main=”Qui-quadrado com 8 graus de liberdade”)
plot(x, y, main=”Qui-quadrado com 50 graus de liberdade”)
```