

**TRABALHO DE CONCLUSÃO DE DISCIPLINA**

**Análise Exploratória de Dados**

Nacarle Figueiredo Delmiro

Vivian Lima de Freitas Melo

**FORTALEZA, BRASIL**

**2020**

1. **INTRODUÇÃO**

O presente trabalho tem por objetivo, com o apoio de ferramentas e modelos estatísticos - neste utilizamos as ferramentas Excel (Pacote Microsoft Office 2013) e o Pacote R, versão 3.6.3, realizar uma análise exploratória de cada questionamento enviado.

Para tal, iniciamos com a análise de cada questionamento, conforme explicitado em cada questão.

Buscamos então o entendimento e manipulação dos dados incluindo suas classificações categóricas, e demais informações necessárias para contribuição na análise de cada questionamento.

**PALAVRAS-CHAVE:** *Amostra, Processo de Amostragem, Per Capita, Estatística Descritiva, Boxplot, Histograma, Correlação, Teorema do Limite Central e Outlier.*

1. **DESENVOLVIMENTO**

2.1 Grande parte da estatística está baseada na amostra. O texto (A amostra com tendenciosidade introduzida) apresenta algumas características recorrentes e implicações em processos de amostragem.

1. (1 ponto) Discorra brevemente (5 a 10 linhas) quais os principais problemas apontados no texto sobre o processo de amostragem.

Houve uma necessidade em generalizar os resultados em virtude da própria natureza da amostragem, e a generalização dos resultados apresentou algumas discrepâncias em todas os exemplos citados.

Houve também uma má seleção da amostra, no momento em que vemos que a renda se baseou numa amostra composta por: (1) Cujos endereços são conhecidos e (2) Que responderam ao questionário.

Vimos que existiam os indivíduos que se desviaram do padrão, a exemplo de Yale, onde simplesmente tínhamos os vagabundos, os alcoólatras desempregados, os escritores e artistas miseráveis, que muito provavelmente não qualificaram os resultados.

Se alguns indivíduos selecionados não foram selecionados de forma aleatória, os resultados podem ter sido seriamente afetados.

Uma coisa é a teoria e a outra é a prática. Apenas em ambientes muito controlados é possível fazer com que as amostras sejam aleatórias. Além disso, quando temos universos compostos por grupos homogêneos (entre si) de pessoas, podemos aproveitar esse grupo para melhorar a qualidade da minha amostra (ou reduzir o tamanho dela).

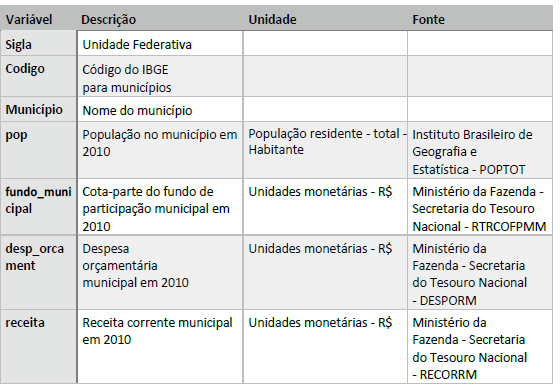
A tendência natural da amostra, em alguns casos, falsificou claramente a pesquisa.

b) (1 ponto) Atualmente estamos vivendo tempos difíceis de quarentena por conta da pandemia do COVID-19. Os dados disponibilizados pelos órgãos competentes podem apresentar os problemas descritos no texto? Argumente. (5 a 10 linhas).

Sabemos que a Pesquisa de Amostra Domiciliar (PAD) é uma das ferramentas utilizadas. Ela funciona como um censo. Com ela, é possível mapear todo o território por faixa etária, sexo e região administrativa, e isso nos ajuda a entender as questões demográficas: onde temos mais idosos, que estão no grupo de risco; qual o nível de acesso da população à internet, o que possibilita a ampliação de medidas como o teletrabalho e as aulas virtuais. Podemos ver também como está a questão do transporte público, quem utiliza, quem não. Ou seja, teoricamente, uma radiografia detalhada de todo o estado. Porém, não nos é garantido que exista um padrão de medição em todos os territórios que estão sendo medidos.

Pesquisadores já encontraram também, evidências de que pessoas com o novo coronavírus podem transmitir a doença sem que apresentem os sintomas, o que é chamado de caso assintomático. Entretanto, a carga viral é menor e o potencial de contágio também. O que nos dá a clareza de que a verdade é que não temos uma base unificada para sabermos onde a doença já chegou, o impacto econômico do vírus e a mortalidade em cada cidade. Ainda existe muita desinformação que circula sobre o assunto.

2.2 A base de dados *“dados\_trabalho\_aed.xls”* apresenta as seguintes variáveis: Sigla, Código, Município, pop, fundo\_municipal, desp\_orcament e receita.



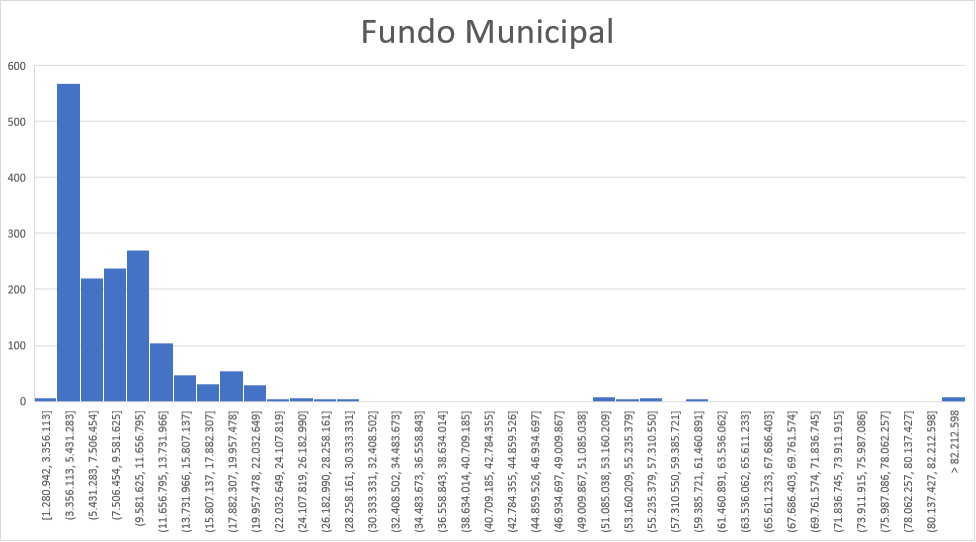
1. i. Importe os dados;

Conforme descrito anteriormente, os dados foram importados no Excel 2013.

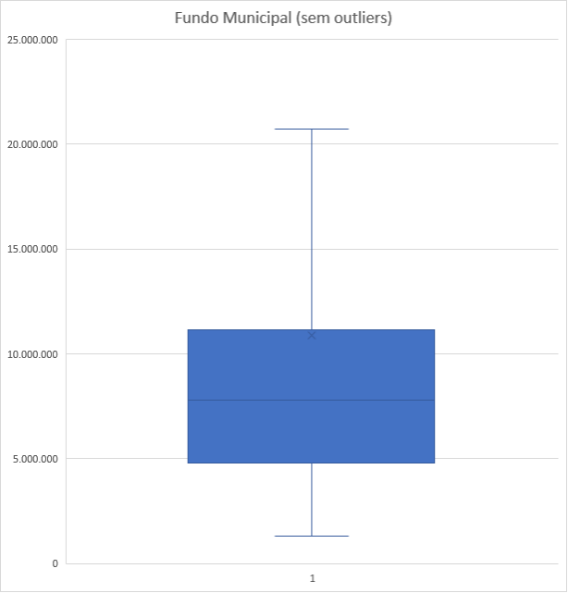
1. ii. Gere uma nova variável (**superávit**) por meio da diferença entre as variáveis **receita** e **desp\_orcament.**
2. iii. Calcule as variáveis fundo\_municipal per capita, desp\_orcament per capita e receita per capita.
3. (1 ponto) Elabore o histograma e o Boxplot para as variáveis fundo\_municipal, desp\_orcament e receita. É possível verificar se existe concentração? Na sequência, elabore os mesmos gráficos para as variáveis fundo\_municipal per capita, desp\_orcament per capita e receita per capita e compare os resultados com os resultados obtidos com as variáveis originais

**Variável fundo\_municipal**

Histograma

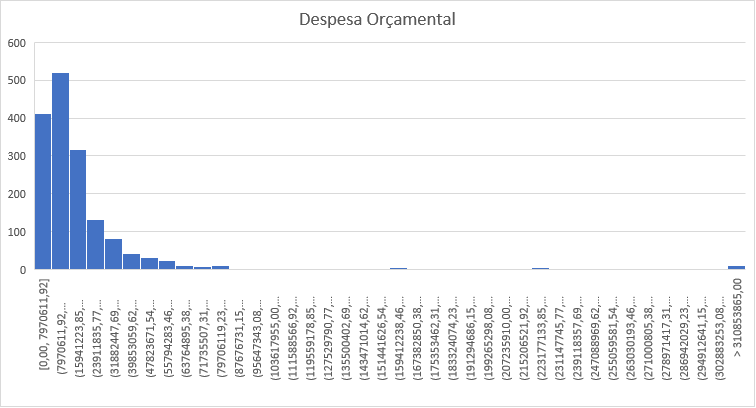


Boxplot

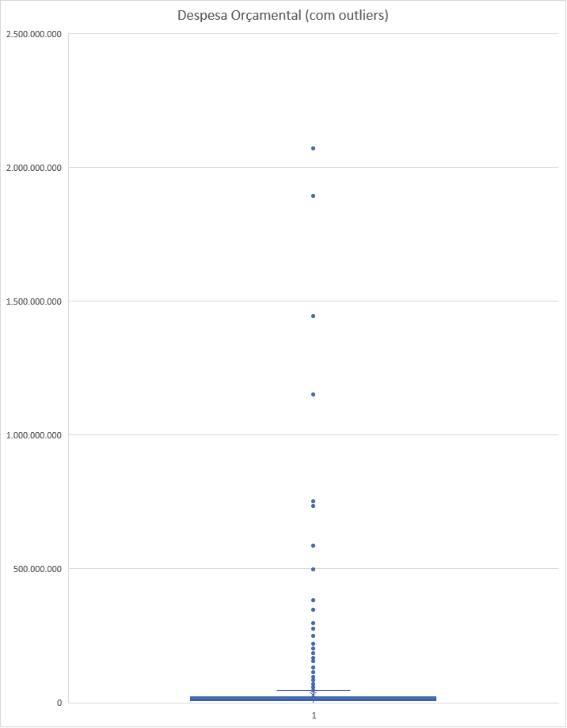
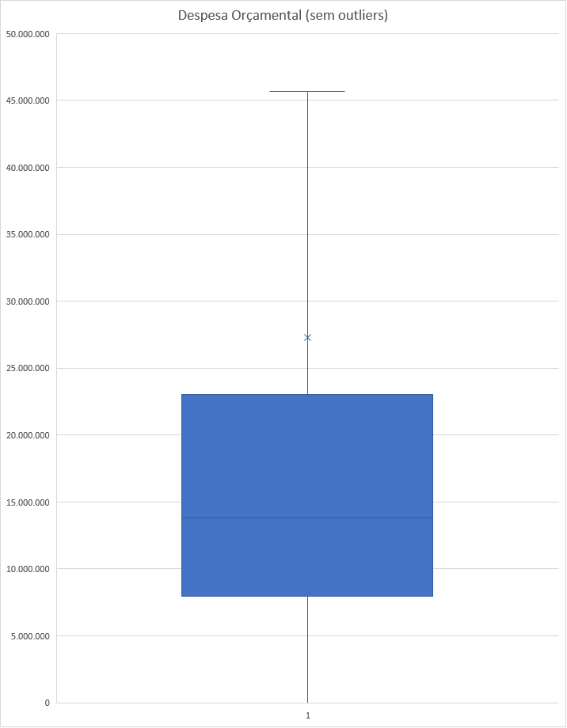


**Variável desp\_orcament**

Histograma

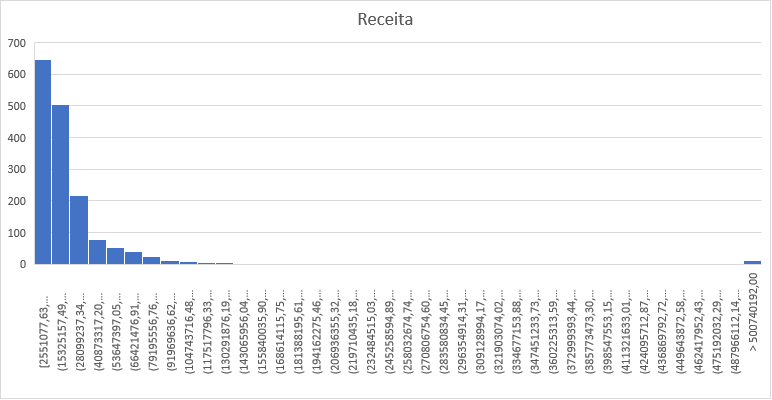


Boxplot

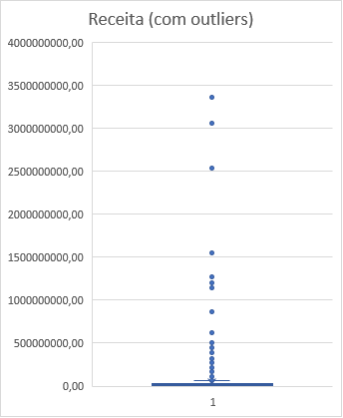
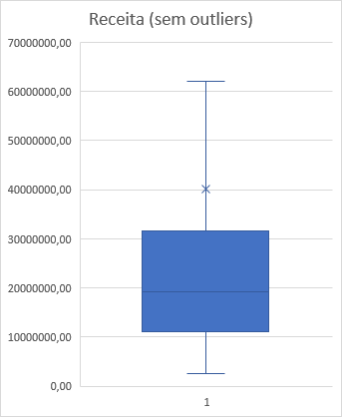


**Variável Receita**

Histograma

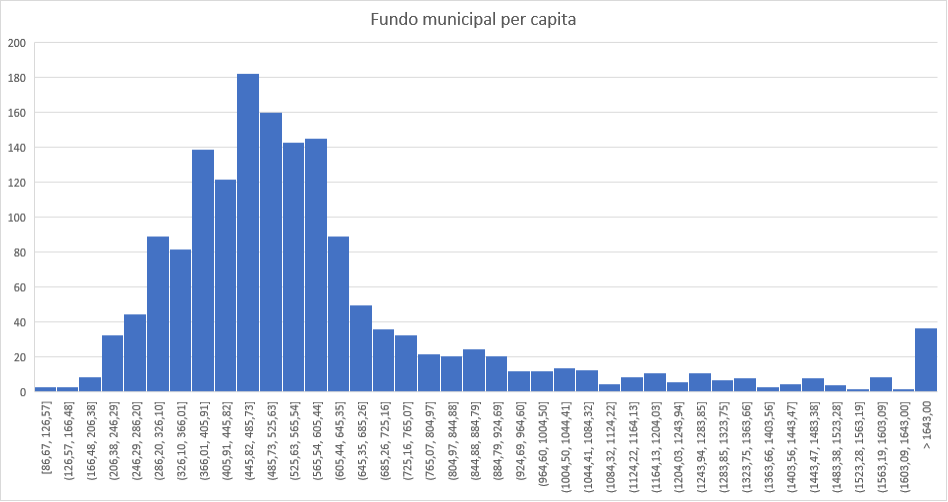


Boxplot

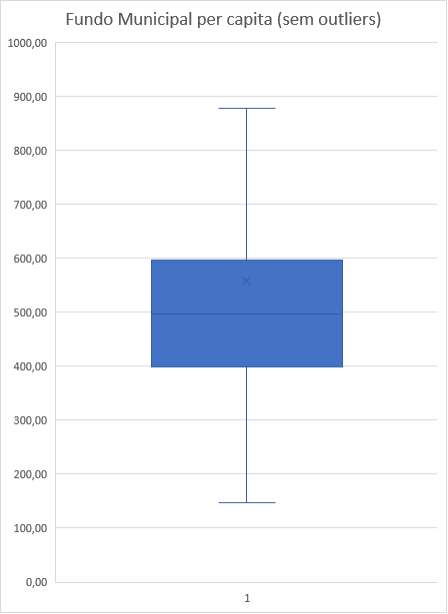
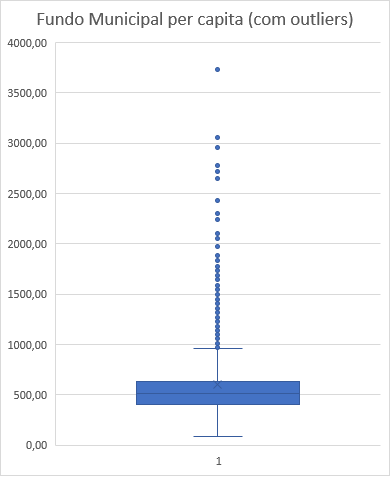


**Variável fundo\_municipal per capita**

Histograma

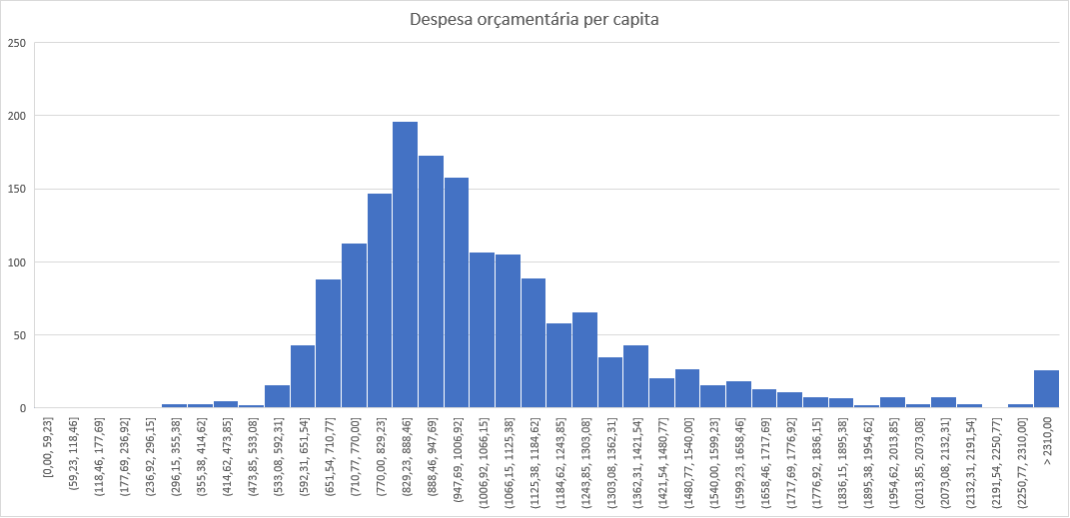


Boxplot

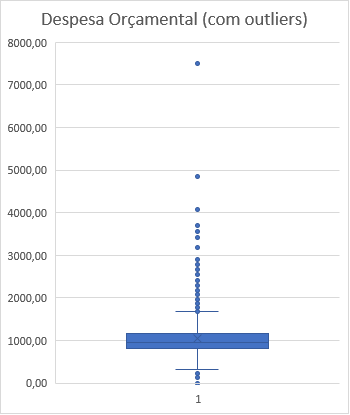
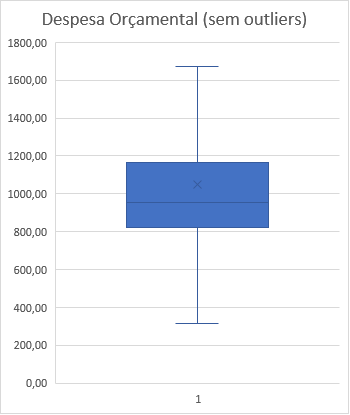


**Variável desp\_orcament per capita**

Histograma

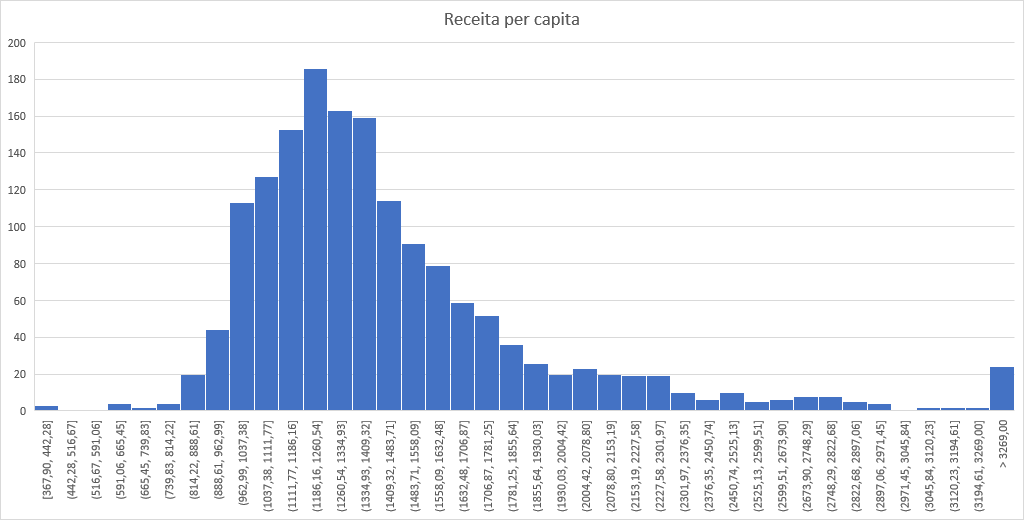


Boxplot

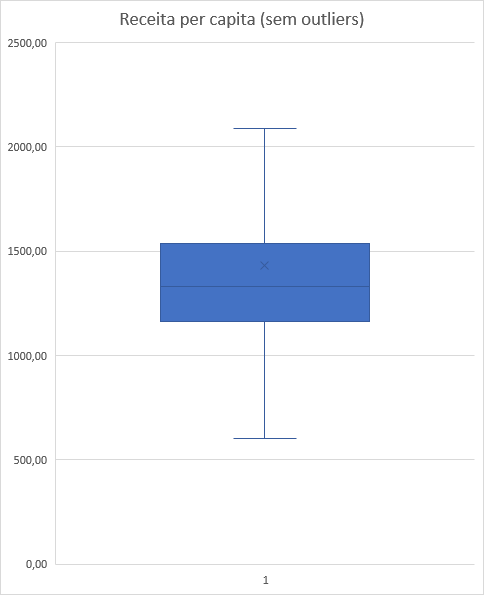
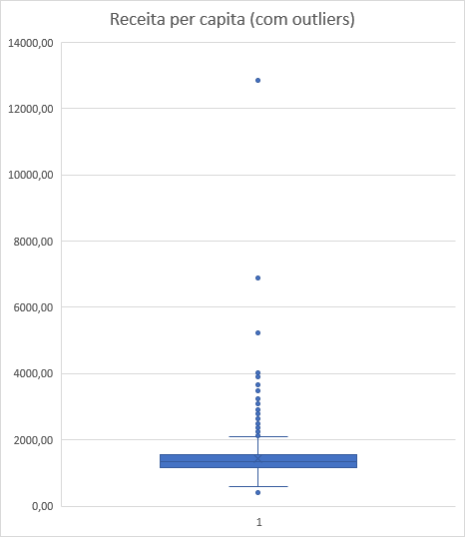


**Variável receita per capita**

Histograma



Boxplot



Análise das Variáveis Originais.

A partir destes histogramas, nota-se que, o conjunto de dados apresenta uma forte concentração à esquerda, distribuição assimétrica negativa, com uma diminuição brusca na frequência até valores que se tornam muito distantes (*outliers*).

Através dos Boxplot, podemos verificar que nestas distribuições existem diversos *outliers* com valores superiores aos limites superiores (LTS), porém, não foram verificados outliers abaixo dos seus limites inferiores (LTI).

Análise das Variáveis Originais versus Per Capita.

Ao analisarmos os Histogramas das novas variáveis Per Capita, exibidos graficamente em barras previamente tabuladas, conseguimos verificar que a concentração mesmo ainda estando assimétrica, deslocou-se para o centro, diminuindo esta assimetria.

Através da análise dos Boxplot das novas variáveis Per Capita, tais quais mostram a mediana, primeiro e terceiro quartis de uma distribuição (pontos 50%, 25% e 75% na distribuição acumulada), podemos perceber que um número menor de *Outliers* aumentou a proporção dos limites superiores (LTS) e inferiores (LTI) no Boxsplot facilitando sua visualização.

1. b) (2 pontos) Calcule as estatísticas descritivas (média, quartis, desvio padrão, coeficiente de variação, assimetria e curtose) para as variáveis fundo\_municipal, desp\_orcament e receita. Obtenha as mesmas estatísticas para as variáveis fundo\_municipal per capita, desp\_orcament per capita e receita per capita. Elabore uma análise comparando os resultados com os resultados obtidos com as variáveis originais.

Dados Vide Anexo 01.

Fundo Municipal Versus Per Capita

No histograma do Fundo Municipal, a média é mais influenciada pelos *outliers* do que no Fundo municipal per capita, pois, a distância deles em relação a mediana é bem maior, tornando a média bem maior que a mediana.

No Fundo Municipal, A média é bem maior do que a mediana com assimetria negativa bem maior (14,25) do que a do Fundo Municipal per capita (2,96). Isso se deve a presença de *outliers* mais distantes da mediana do que no Fundo Municipal per capita.

Também devido a presença mais acentuada de *outliers*, o coeficiente de variação no Fundo municipal é bem maior (2,19) do que o do fundo municipal per capita (0,58).

O grau de assimetria negativa também é muito superior (14,25) no Fundo Municipal do que no Fundo Municipal per capita (2,97)

Os altos coeficientes de curtoses classificam as distribuições como leptocúrticas.

Despesa Orçamentária Versus Per Capita

No histograma Despesa Orçamentaria, a assimetria negativa é bem mais evidente, logo, conclui-se que a média é menor que a mediana. No segundo histograma, Despesa Per Capita, vimos que a forma é mais influenciada pela média, aproximando as caudas de uma normal, porém, ainda caracterizando bastante desvio.

Na Despesa Orçamentaria, a média é bem maior do que a mediana com assimetria negativa bem maior que a do que a da Despesa Per Capita, isso se deve a presença de *outliers* mais distantes da mediana do que a anteriormente citada.

Também devido à presença mais acentuada de *outliers*, o coeficiente de variação na Despesa Orçamentaria é bem maior que a da segunda amostra Per Capita.

Conforme explicitado no item "media", o grau de assimetria negativa também é muito superior na Despesa Orçamentaria do que na Despesa Per capita.

Os altos coeficientes de curtoses classificam as distribuições como leptocúrticas.

Receita Orçamentária Versus Per Capita

No histograma Receita Orçamentaria, a assimetria negativa é bem mais evidente, logo, conclui-se que a média é menor que a mediana. No segundo histograma, Receita Per Capita, vimos que a forma é mais influenciada pela média, aproximando as caudas de uma normal, porém, ainda caracterizando bastante desvio.

Na Receita Orçamentaria, a média é bem maior do que a mediana com assimetria negativa bem maior que a do que a da Receita Per Capita, isso se deve a presença de *outliers* mais distantes da mediana do que a anteriormente citada.

Também devido à presença mais acentuada de *outliers*, o coeficiente de variação na Receita Orçamentaria é bem maior que a da segunda amostra Per Capita.

Conforme explicitado no item "media", o grau de assimetria negativa também é muito superior na Receita Orçamentaria do que na Receita Per capita.

Os altos coeficientes de curtoses classificam as distribuições como leptocúrticas.

3) (2,5 pontos) Experimento: Distribuição amostral da média – Teorema do Limite Central.

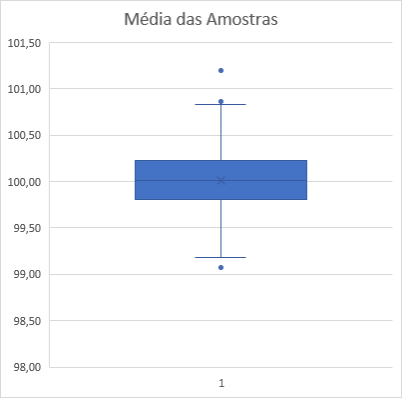
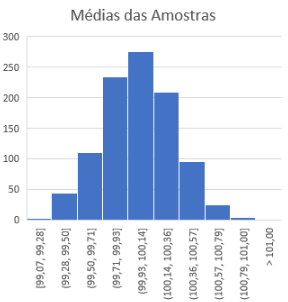
a) Simule 1000 amostras de tamanho *n*=1000 relativa a uma variável X com média 𝜇 e desvio padrão 𝜎 de uma distribuição de probabilidades qualquer (exceto normal).

As amostras foram geradas usando o Plugin Action Stat no Excel 2013. Vide Anexo 02, arquivo Amostras.xlsx.

b) Calcule a média amostral para cada amostra.

Vide Anexo 02, arquivo Medias e Observações.xlsx, coluna A.

c) Faça um histograma dos valores das medias amostrais.



d) Examine os dados de forma gráfica: histograma, quanto ao padrão geral, centro e dispersão, *box-plot* quanto a presença de *outliers* ou outros desvios.

Tendência Central, Variabilidade e Forma

De acordo com o coeficiente de assimetria de Pearson de quartil, os valores de simetria estão entre -1>x>1, logo, verificamos que o Histograma é simétrico, tipo distribuição normal (sino): A frequência é mais alta no centro e decresce gradualmente para as caudas de maneira simétrica. A média e a mediana localizam-se no centro do histograma (ponto de pico). Nesta tabela de frequência, a classe de maior frequência, chamada de classe modal (pico), corresponde à 99,93 - 100,14.

Em termos de Dispersão, o Intervalo Interquartílico é 0,42, a Amplitude é de 2,13, Variância é de 0,09 e Desvio Padrão de 0,30 exibindo uma variabilidade de dados baixa.

A distância do segundo quartil para o primeiro é de 0,21, este valor se repete para a distância do terceiro quartil para o segundo. Já a distância do terceiro quartil para o quarto é 0,98, neste momento, conseguimos visualizar dois *outliers* ultrapassando o limite superior (LTS), e a distância entre o valor mínimo e o primeiro quartil é de 0,73, exibindo um *Outlier* abaixo do Limite Inferior (LTI).

e) Verifique se os dados seguem o padrão de um modelo Normal. Seguem abaixo alguns critérios para tal verificação.

i. Há simetria no histograma e forma de sino?

Sim, o histograma é simétrico e em forma de sino.

ii. A média, a mediana e a moda são próximas?

Sim, e encontram-se na mesma classe.

iii. A amplitude não ultrapassa 6 vezes o desvio padrão?

Não ultrapassa, pois, o desvio padrão é de 0,30 e seis vezes este valor é igual a 1,80 e a Amplitude é de 2,13.

iv. A proporção de *outliers* é bem pequena e não ultrapassa 0,5%?

A proporção é bem pequena com apenas três *outliers* em mil elementos, ou seja, 3/997 =0,003 e não ultrapassa 0,5%.

4) Os dados a seguir correspondem à variável renda familiar e gasto com alimentação (em unidades monetárias) para uma amostra de 25 famílias.

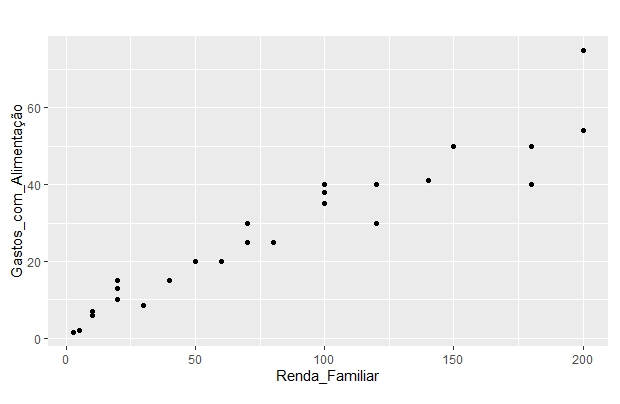
|  |  |
| --- | --- |
| Renda Familiar (X) | Gasto com Alimentação (Y) |
| 3 | 1,5 |
| 5 | 2,0 |
| 10 | 6,0 |
| 10 | 7,0 |
| 20 | 10,0 |
| 20 | 13,0 |
| 20 | 15,0 |
| 30 | 8,5 |
| 40 | 15,0 |
| 50 | 20,0 |
| 60 | 20,0 |
| 70 | 25,0 |
| 70 | 30,0 |
| 80 | 25,0 |
| 100 | 40,0 |
| 100 | 35,0 |
| 100 | 38,0 |
| 120 | 30,0 |
| 120 | 40,0 |
| 140 | 41,0 |
| 150 | 50,0 |
| 180 | 40,0 |
| 180 | 50,0 |
| 200 | 75,0 |
| 200 | 54,0 |

1. a) (0,5 ponto) Calcular o coeficiente de correlação entre essas variáveis. O que se pode concluir a partir deste resultado?

O Coeficiente de Correlação é de 0.9504443. A correlação é positiva, pois as duas variáveis movem-se juntas, e o resultado desta se aproxima de um. É uma correlação forte, pois as duas variáveis se movem em direções similares.

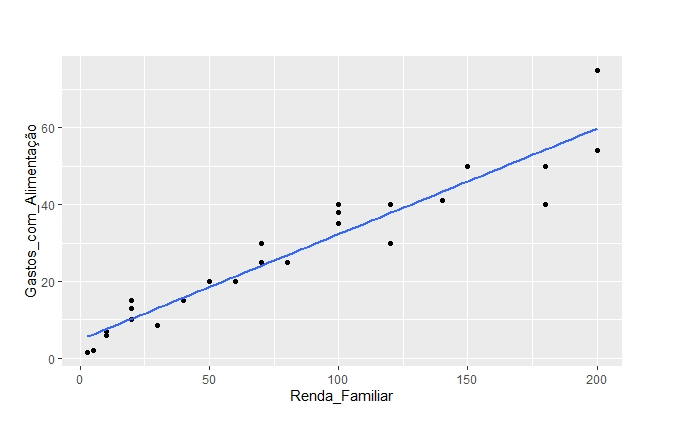
Vide Anexo 03.

1. b) (0,5 ponto) Construa o diagrama de dispersão da variável gasto com alimentação (Y) em função da renda familiar (X).



1. c) (1 ponto) Obtenha um modelo linear do gasto com alimentação em função da renda familiar. O que se pode concluir sobre o valor do coeficiente angular da reta encontrada?

Interpretando a reta de regressão estimada da média de todos os números da variável Gasto de alimentação (Y) em função da média da Renda familiar (X), o resultado é positivo, o que representa a linha crescente e inclinada para a direita. Considerando que o ponto (83.12, 27.64), divide o gráfico em quatro quadrantes, e a multiplicação dos quadrantes 1 e 3 resulta em um valor positivo, podemos concluir que a correlação é alta e direta.



1. d) (0,5 ponto) Obtenha e interprete o coeficiente de determinação do modelo obtido.

O coeficiente de determinação explica o grau de ajuste do modelo, ou seja, o percentual de variação de Y que é explicado pela variabilidade de X. Para cada unidade de variação em renda familiar, é esperado um aumento de 0,9033444 em Y.

**ANEXOS**

Todos os anexos necessários para o bom e completo entendimento desta análise estão listados abaixo:

• **Anexo 01** – Segunda Questão, letra B;

• **Anexo 02** – Arquivos Terceira Questão;

• **Anexo 03** – Quarta Questão, letra A.