Revisão Pré-Prova

Análise Exploratória

Considere a amostra abaixo:

ID_Cliente	Sexo	Meses_Última_Compra	Qtd_Compras_Último_Ano	Valor_Médio_Último_Ano
1	F	14,4	0	0
2	F	3,3	2	99
3	F	6,8	4	120
4	F	9,2	3	174
5	F	2,4	3	529
6	M	13,9	0	0
7	M	9,7	1	49
8	M	1,9	4	255
9	M	5,6	2	417
10	M	1,7	5	1221

1) Defina o tipo de cada variável:

Primeiramente precisamos identificar quais são as variáveis da amostra:

São elas: Sexo, Meses_Última_Compra, Qtd_Compras_Último_Ano, Valor_Médio_Último_Ano

Definindo:

Sexo: Qualitativa – Nominal

Meses_Última_Compra: Quantitativa – Contínua

Qtd_Compras_Último Ano: Quantitativa – Discreta

Valor_Médio_Último_Ano: Quantitativa – Contínua

2) Calcule as médias e os desvios padrões amostrais. Qual a diferença do cálculo entre o desvio padrão amostral e o populacional?

Vamos copiar primeiramente os valores para uma guia do Excel. A variável qualitativa Sexo não possui média ou desvio padrão, logo iremos calcular apenas das demais variáveis:

Cálculo das Médias (média(x)):

Meses_Última_Compra: 6,89

Qtd_Compras_Último Ano: 2,40

Valor_Médio_Último_Ano: 286,40

Cálculo dos Desvios Padrões Amostrais (desvpad.a(x))

Meses_Última_Compra: 4,78

Qtd_Compras_Último Ano: 1,71

Valor_Médio_Último_Ano: 372,11

A diferença entre os desvios padrões amostral e populacional se dá pelo fato de que o desvio padrão amostral é calculado pela raiz quadrada da variância amostral, que utiliza como denominador (n-1) e o populacional considera todo o tamanho da população já conhecido (N).

3) Podemos dizer que os homens compram mais que as mulheres? Conclua descritivamente utilizando a média.

Para resolver este exercício, podemos criar uma tabela dinâmica do cálculo das médias das métricas Qtd_Compras_Último_Ano e Valor_Médio_Último_Ano por sexo.

Rótulos de Linha	Média de Qtd_Compras_Último_Ano	Média de Valor_Médio_Último_Ano
F	2,4	184,4
M	2,4	388,4
Total Geral	**2,4**	**286,4**

Quando nos referimos a Qtd_Compras_Último_Ano, vemos que a média por sexo é de 2,4, logo não há diferença na quantidade de compras.

Porém quanto vemos pela variável Valor_Médio_Último_Ano, a média do valor gasto no último ano do sexo Masculino é maior do que do sexo Feminino, e logo podemos afirmar que os homens compram mais que as mulheres.

4) Nesta amostra, existe uma concentração maior de homens ou mulheres. Conclua utilizando a tabela de frequências.

Utilizando a tabela dinâmica do Excel, contando quantos usuários tem de cada sexo, temos o seguinte cenário:

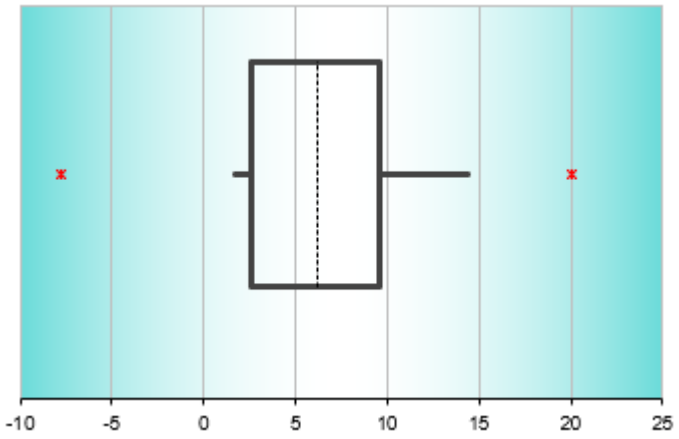
Rótulos de Linha	Contagem de ID_Cliente	Contagem de ID_Cliente2
F	5	50,00%
M	5	50,00%
Total Geral	**10**	**100,00%**

Com isso, podemos concluir que temos a mesma proporção de homens e mulheres, 50% de cada sexo.

5) Faça o boxplot para cada variável quantitativa. Alguma possui valores discrepantes: Compare as assimetrias.

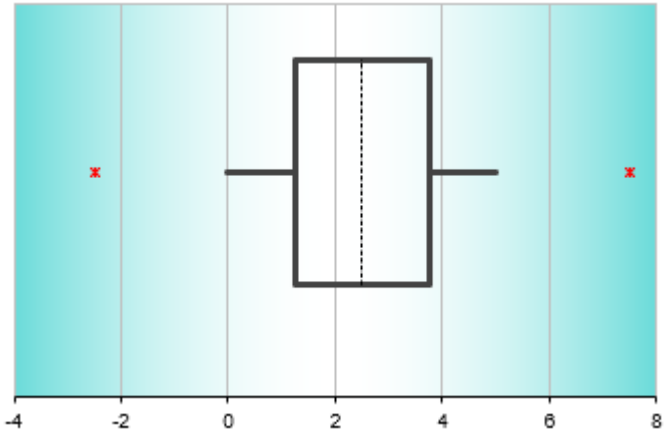
Abra o material da aula arquivo excel “Medidas Descritivas” para facilitar a resolução, cole cada conjunto de dados de variáveis, na coluna B

Meses_Última_Compra:



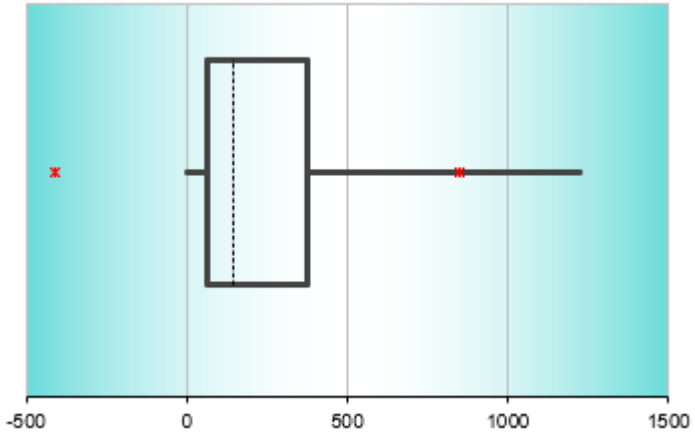
A média é maior que a mediana, logo o coeficiente de assimetria é positivo e podemos dizer que a distribuição é assimétrica a direita. Não há valores discrepantes.

Qtd_Compras_Último Ano:



A média é menor que a mediana, logo o coeficiente de assimetria é negativo e podemos dizer que a distribuição é assimétrica a esquerda. Não há valores discrepantes.

Valor_Médio_Último_Ano:



A média é maior que a mediana, logo o coeficiente de assimetria é positivo e podemos dizer que a distribuição é assimétrica a direita. Podemos identificar que há valores discrepantes, acima do limite superior do boxplot.

6) Qual o significado do 1º e 3º quartil de cada variável?

1º Quartil de cada variável significa que 25% dos dados são menores do que o valor de Q1

Meses_Última_Compra: 2,625

Qtd_Compras_Último Ano: 1,25

Valor_Médio_Último_Ano: 61,5

3º Quartil de cada variável significa que 75% dos dados são menores do que o valor de Q3

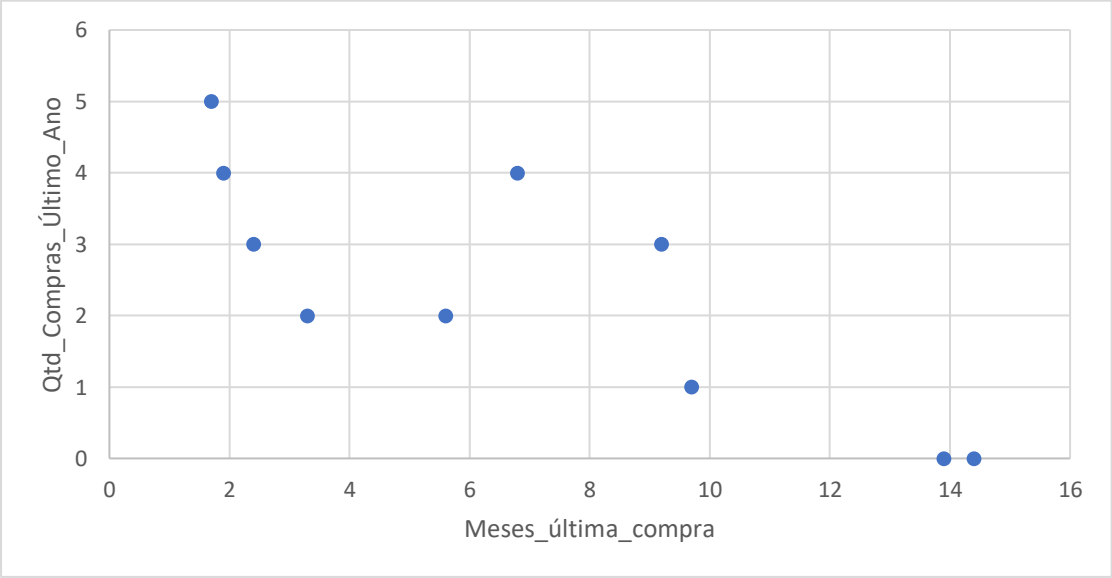
Meses_Última_Compra: 9,575

Qtd_Compras_Último Ano: 3,75

Valor_Médio_Último_Ano: 376,5

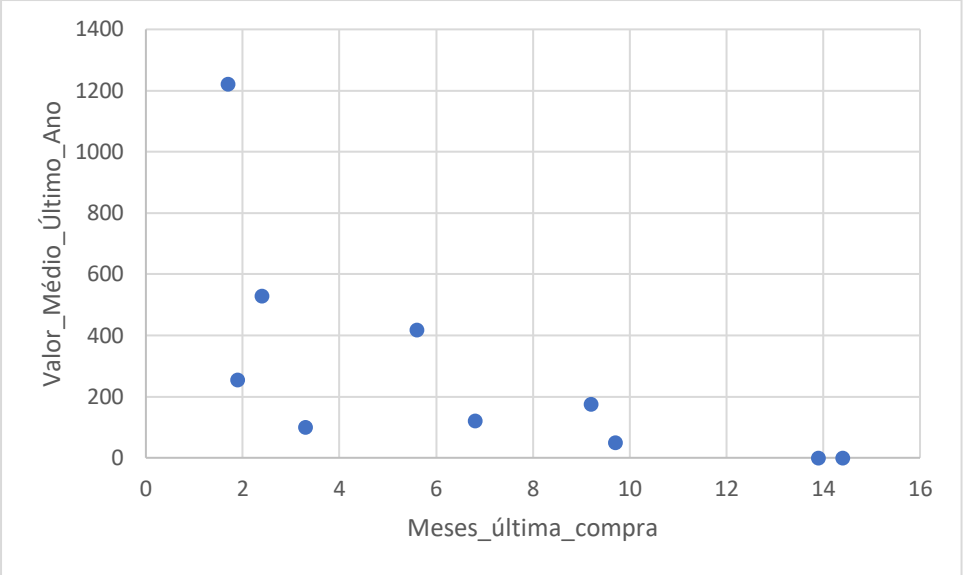
7) Faça gráficos de dispersão para a combinação das variáveis:

- Meses_Última_Compra x Qtd_Compras_Último_Ano



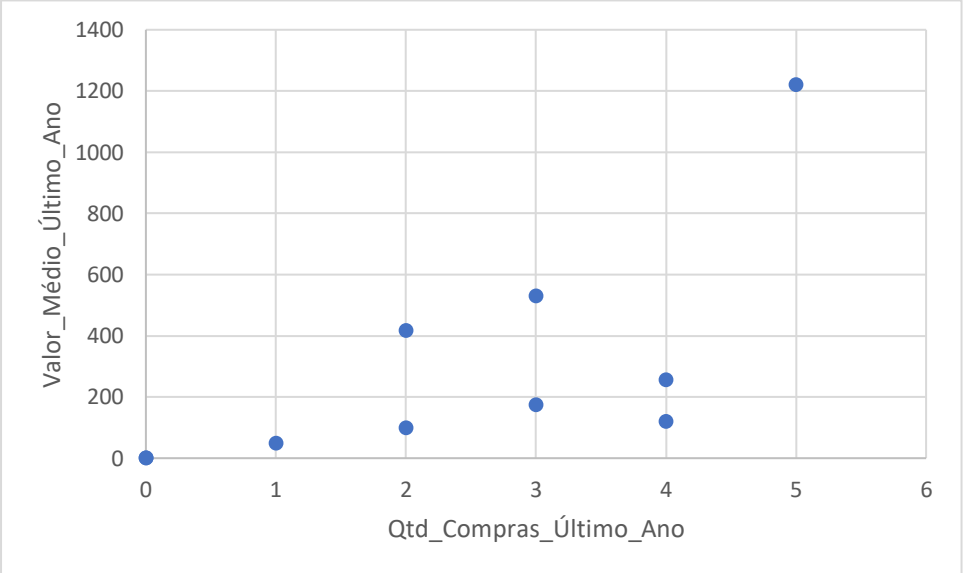
Percebe-se que quanto maior o tempo de meses da última compra menor a quantidade de compras no último ano.

- Meses_Última_Compra x Valor_Médio_Último_Ano



Percebe-se que quanto maior o tempo em meses da última compra, menor o valor médio do útimo ano.

- Qtd_Compras_Último_Ano x Valor_Médio_Último_Ano



Percebe-se que quanto mais compras feitas no último ano, maior o valor médio gasto no último ano.

8) Calcule o coeficiente de correção linear para cada combinação das variáveis. Interprete-os

- Meses_Última_Compra x Qtd_Compras_Último_Ano: -0,811678

Estas variáveis possuem uma correlação negativa entre elas, ou seja, a medida que uma aumenta, a outra diminui. E o contrário também ocorre.

- Meses_Última_Compra x Valor_Médio_Último_Ano: -0,64697

Estas variáveis possuem uma correlação negativa entre elas, ou seja, a medida que uma aumenta, a outra diminui. E o contrário também ocorre.

- Qtd_Compras_Último_Ano x Valor_Médio_Último_Ano: 0,683661

Estas variáveis possuem uma correlação positiva entre elas, ou seja, a medida que uma aumenta, a outra também aumenta, e o mesmo acontece ao contrário

Teste de Hipóteses, Estimação Intervalar e Distribuições de Probabilidades

1) **Uma prefeitura realizou uma pesquisa com 800 pessoas em diversos supermercados do município e verificou um ticket médio de R$ 550, com desvio padrão amostral de R$ 310.**
   Obtenha o intervalo com 90% de confiança para a média populacional.

O que temos de informações fornecidas:

Tamanho da amostra: n = 800

Ticket médio: x_barra: R$ 550

Desvio Padrão amostra: R$ 310

Intervalo de Confiança: 90%

Neste caso, como não temos o desvio padrão populacional conhecido, apenas o amostral, podemos utilizar o intervalo de confiança por t-student.

Para facilitar a resolução, vamos utilizar o material de apoio em Excel, chamado: “Teste_de_hipoteses_alunos”

Ao abrir o arquivo, vá na guia: IC_Média_t

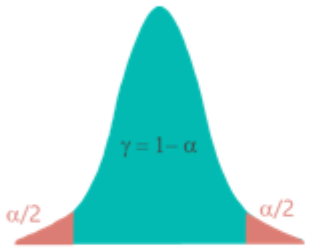
Preencha os valores da planilha, com os valores fornecidos pelo enunciado:

Teremos o seguinte cenário:

Margem de erro	
γ=	0,9
t=	1,647
n=	800
s=	310
x_barra=	550

Margem de erro (ME)=	18,05

Intervalo de confiaça (γ)
[$\bar{x}$ ± margem de erro]
[531,951 568,049]



$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

$\bar{x}$ é a média amostral.
γ é o coeficiente de confiança.
n é o tamanho da amostra.

- $t_{\alpha/2}$ fornece a área de $\alpha/2$ na cauda superior da **T-student.**
- s é o desvio padrão amostral:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Logo, o intervalo com 90% de confiança para o ticket médio será entre R$ 531,95 e R$ 568,05

2) Em 2016, uma pesquisa com 2000 pessoas mostrou que 54% estavam a favor da permanência do Reino Unido na União Europeia.
Qual o intervalo com 95% confiança para a população?

O que temos de informações fornecidas:

Pessoas que participaram da pesquisa: n = 2000

Favor da permanência do Reino Unido na União Europeia: p = 0,54 ou 54%

Confiança: 95%

Para facilitar a resolução, vamos utilizar o material de apoio em Excel, chamado: “Teste_de_hipoteses_alunos”

Ao abrir o arquivo, vá na guia: IC_Prop_Z (pois estamos tratando de intervalo de confiança de proporção)

Preencha os valores da planilha, com os valores fornecidos pelo enunciado:

Teremos o seguinte cenário:

Margem de erro	
γ=	0,95
z=	1,960
n=	2000
var(p)=	0,2484
p=	0,54

Margem de erro (ME)=	0,02

Intervalo de confiaça (γ)
[$\bar{p}$ ± margem de erro]
[0,5182 0,5618]

Sendo $Z \sim N(0,1)$:

$$\left(\bar{p} - Z_\gamma \sqrt{\frac{\bar{p}(1-\bar{p})}{n}};\ \bar{p} + Z_\gamma \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\right)$$

- O $\bar{p}$ é a proporção amostral
- n é o tamanho da amostra

→ Margem de Erro $Z_\gamma \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$

Logo, o intervalo com 95% de confiança para a proporção de pessoas a favor da permanência do Reino Unido na União Europeia está entre 51,82% e 56,18%.

3) Em uma amostra com 300 brasileiros, foi obtido um peso médio de 70 kg, enquanto uma amostra com 400 americanos resultou em uma média de 82 kg por pessoa. Suponha que o peso do corpo humano possui uma variância (populacional) de 64 kg.
Com 95% de confiança, é possível concluir que brasileiros e americanos possuem pesos médios diferentes?

O que temos de informações fornecidas:

Brasileiros: n1 = 300

Peso médio Brasileiros: x_barra1 = 70kg

Americanos: n2 = 400

Peso médio Americanos: x_barra2 = 82kg

Variância peso corpo humano populacional = 64kg (logo implica que o desvio padrão populacional é de 8kg) – Este ponto é muito importante, pois há problemas que temos direto o desvio padrão, e outros nos é fornecido a variância. Prestar muita atenção, pois se utilizar a variância o resultado pode ser outro.

Confiança: 95%

Para facilitar a resolução, vamos utilizar o material de apoio em Excel, chamado: “Teste_de_hipoteses_alunos”

Ao abrir o arquivo, vá na guia: IC_Média_Z (pois temos a variância populacional conhecida)

Vamos primeiro calcular o intervalo de confiança dos Brasileiros:

Margem de erro	
γ=	0,95
z=	1,960
n=	300
σ=	8
x_barra=	70

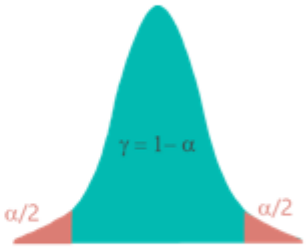
Margem de erro (ME)=	0,91

Intervalo de confiaça (γ)

[$\bar{x}$ ± margem de erro]

[69,095 70,905]

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$



- $z_{\alpha/2}$ fornece a área de $\alpha/2$ na cauda superior da **Normal.**
- σ é o desvio padrão populacional.

$\bar{x}$ é a média amostral.
γ é o coeficiente de confiança.
n é o tamanho da amostra.

Agora dos Americanos:

Margem de erro	
γ=	0,95
z=	1,960
n=	400
σ=	8
x_barra=	82

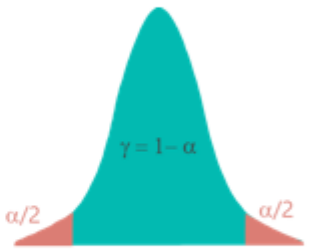
Margem de erro (ME)=	0,78

Intervalo de confiaça (γ)

[$\bar{x}$ ± margem de erro]

[81,216 82,784]

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$



- $z_{\alpha/2}$ fornece a área de $\alpha/2$ na cauda superior da **Normal.**
- σ é o desvio padrão populacional.

$\bar{x}$ é a média amostral.
γ é o coeficiente de confiança.
n é o tamanho da amostra.

Comparando os dois intervalos de confiança, podemos concluir que brasileiros e americanos possuem pesos médios diferentes com 95% de confiança, pois não há intersecção entre eles.

4) **Uma empresa vende fatias de muçarela em embalagens de 300g cada, com um desvio padrão (populacional) de 1g, conforme processo automatizado.**
Uma amostra de 50 embalagens foi pesada, resultando em uma média de 299g.
Pode-se concluir que o processo está fora do padrão a um nível de 99% de confiança?

O que temos de informações fornecidas:

Peso médio das embalagens das fatias de muçarela: 300g

Desvio padrão populacional: 1g

Amostra: 50 embalagens

Média amostral: 299g

Nível de confiança: 99%

Este problema se trata de teste de hipóteses, queremos saber se estão produzindo mais ou menos quantidade de muçarela durante o processo, logo temos as seguintes hipóteses:

H0: $\mu = 300g$

Há: $\mu \neq 300g$

Trata-se de um teste bilateral, pois o processo estará desajustado se produzir menos ou mais do que especificado na embalagem.

Para facilitar a resolução, vamos utilizar o material de apoio em Excel, chamado: “Teste_de_hipoteses_alunos”

Ao abrir o arquivo, vá na guia: TH_Média_z (pois temos a variância populacional conhecida)

Iremos trabalhar com a parte BILATERAL no arquivo

Para um nível de confiança de 99%, nosso valor alfa é 1%

Temos o seguinte cenário:

Teste de Hipótese: Bilateral

α=	0,01
γ=	0,99
z_α/2=	2,576
n=	50
σ=	1
μ=	300

x_barra=	299
Z=	- 7,07
p-valor=	0,0000

2 * (1-DIST.NORMP.N(z; VERDADEIRO))
INV.NORMP(0,975)= 1,96

$H_0: \mu = \mu_0$
$H_a: \mu \neq \mu_0$

Teste Bilateral (Cauda Inferior e Superior)

A estatística do teste é -7,07, o valor de z associado a 99% (bilateral) é 2,576, logo rejeito H0, se t<= -2,576 ou t>= 2,576. Com isso rejeitamos H0, portanto com 99% de confiança podemos dizer que o processo está fora do padrão.

Pelo p-valor temos que, o p-valor do teste é menor que 0,0000 e o alfa do teste é 0,1, como p-valor do teste é menor que o alfa, rejeitamos H0 e podemos dizer com 99% de confiança que o processo está fora do padrão.

5) Suponha que, em determinada região, 70% dos domicílios possuíam internet banda larga no ano passado.
Este ano, foi realizada uma pesquisa com uma amostra de 1000 domicílios e constatou-se que 750 possuem internet banda larga.
É possível afirmar estatisticamente que houve um aumento deste percentual, com 90% de confiança?

O que temos de informações fornecidas:

Domicílios que possuíam internet banda larga ano passado: 70%

Domicílios este ano: n = 1000

Domicílios com internet banda larga este ano: 75% (750/1000)

Confiança 90%

Formulando o teste de hipóteses:

H0: μ = 70%

Há: μ > 70%

Para facilitar a resolução, vamos utilizar o material de apoio em Excel, chamado: “Teste_de_hipoteses_alunos”

Ao abrir o arquivo, vá na guia: TH_Prop_Z (pois não temos a variância populacional conhecida)

Iremos trabalhar com a parte UNILATERAL (p>p_0) no arquivo

Temos o seguinte cenário:

Teste de Hipótese: Unilateral (H_a: p > p_0)

α=	0,1
γ=	0,9
z_α=	1,28
n=	1000
var(p_0)=	0,21
p_0=	0,7

p_barra=	0,75
Z=	3,45
p-valor=	0,0003

1-DIST.NORMP.N(z; VERDADEIRO)
INV.NORMP(0,95)= 1,645

Teste Unilateral

$H_0: p = p_0$
$H_a: p > p_0$

Temos que nosso valor de Z associado a 90% de confiança é 1,28, logo rejeito H0 se z>= 1,28. Temos que o z calculado é 3,45, logo rejeitamos H0 e podemos dizer com 90% de confiança que a proporção de domicílios com internet banda larga aumentou.

Pelo p-valor temos que, o p-valor do teste é 0,0003 e o alfa do teste é 0,1, como p-valor do teste é menor que o alfa, rejeitamos H0 e podemos dizer com 90% de confiança que a proporção de domicílios com internet banda larga aumentou.