**REVISÃO**

**ANÁLISE EXPLORATÓRIA DE DADOS**

1. Na Tabela abaixo está apresentada a inadimplência de pessoas físicas no Brasil (dados do Banco Central, 2018) para os produtos de cheque especial e crédito pessoal. Encontre a média, a mediana, a moda, o mínimo e o máximo. Qual produto apresentou a maior inadimplência?

|  |  |  |
| --- | --- | --- |
| **Mês** | **Cheque Especial** | **Crédito Pessoal** |
| Janeiro | 15,19 | 3,74 |
| Fevereiro | 13,47 | 3,65 |
| Março | 13,96 | 3,65 |
| Abril | 13,29 | 3,67 |
| Maio | 13,18 | 3,62 |
| Junho | 13,3 | 3,57 |
| Julho | 13,26 | 3,58 |
| Média | 13,6643 | 3,64 |
| Mediana | 13,3 | 3,65 |
| Moda | - | 3,65 |
| Mínimo | 13,18 | 3,57 |
| Máximo | 15,19 | 3,74 |

**DISTRIBUIÇÃO DE PROBABILIDADE**

1. O RH levantou alguns dados dos seus funcionários e descobriu que o salário médio dos gerentes de sua empresa é de 6 mil de reais com um desvio padrão de 2 mil de reais. Calcule as probabilidades:
2. Qual a probabilidade de que um gerente tenha salário superior a 10 mil reais?

Z = (10-6)/2 = 2

Probb = 0,5-0,4772 = **0,0228**

1. Qual a probabilidade de que um gerente tenha salário inferior a 4 mil reais?

Z = (4-6)/2 = -1

Probb = 0,5-0,3413= **0,1587**

1. Qual a probabilidade de que um gerente tenha salário entre 5 e 7 mil de reais?

Z1 = (5-6)/2 = -0,5 e z2 = (7-6)/2 = 0,5

Probb = 0,1915 + 0,1915= **0,3830**

**TESTE DE HIPÓTESE**

1. Um distribuidor de milho afirma que a quantidade de milho colocada em cada pacote segue uma distribuição Normal com média 200 g e desvio padrão 20 g. O comprador do lote tirou uma amostra de 120 pacotes e obteve uma média de 189 g. O comprador deseja testar se a média é 200 g ou diferente de 200 g.

OBS.: Colocar a fórmula do z??????

µ = 200

δ = 20

n = 120

z = 1,96 (2,5% de área para a direita e 2,5% de área à esquerda)

H0: Mi = 200 e H1: Mi diferente de 200

1. Obtenha a região crítica para o teste.

Área à direita de 1,96 e à esquerda de -1,96

1. Através de um teste de hipótese verifique se a quantidade média de milho colocada em cada pacote é igual a 200 g ou diferente de 200 g (usar α=0,05).

Área à direita de 203 e à esquerda de 196

Como 189 (está na região crítica) está à esquerda de 196, rejeito H0, ou seja, a média é diferente de 200.

1. O gerente de uma EMPRESA DE SEGURO SAÚDE afirma que o tempo médio de atendimento a seus clientes é de 8 minutos com um desvio padrão de 2 minutos. Um auditor, acreditando que o tempo médio é superior ao mencionado, retirou uma amostra de 80 ligações e obteve um tempo médio de 9 minutos.

µ = 8

δ = 2

n = 80

z = 1,64 (5% de área para a direita)

H0: Mi = 8 e H1: Mi > 8

A) Obtenha a região crítica para testar a hipótese do auditor.

1. Teste a hipótese de interesse (usar α=0,05).

Como 9 minutos é maior que 8,36, rejeito H0, ou seja, o tempo médio é maior que 8.

1. Um gerente do Bradesco deseja saber se em média os clientes gastam mensalmente R$ 2500,00 no cartão de crédito ou se em média gastam um valor inferior a R$ 2500,00. Para se testar a hipótese de interesse retirou uma amostra de 250 clientes e obteve-se uma média de R$ 2780,00 com um desvio padrão amostral de R$ 340,00.

Mi = 2500

Desvio ?????

N = 250

Desvio amostral = 340

tc = -1,2849

H0: Mi = 2500 e H1: Mi < 2500

1. Obtenha a região crítica para testar a hipótese.
2. Teste a hipótese de interesse (usar α=0,10).

X = -1,2849\*340/raiz(250) + 2500 = 2.472,37

Não rejeito H0, logo, o valor é igual a 2500.

1. Considere um diretor que deseja testar se a proporção de clientes satisfeitos com os serviços oferecidos é 75 % ou superior a 75 %. Para se testar a hipótese de interesse retirou-se uma amostra de 600 clientes e obteve-se uma proporção amostral de 85 %. Considerar 90 % de confiança.
2. Obtenha a região crítica para testar a hipótese.

p0= 75%

n = 600

p = 85%

H0: p = 75% e H1: p > 75%

Z\* raiz(p0(1-p0)/n) +p0 = 0,7726

1. Teste a hipótese de interesse (usar α=0,10).

Como 85% (proporção da amostra) > 0,7726, rejeito H0, ou seja, meus clientes estão satisfeitos.

1. Obtenha a Margem de erro

Margem de erro = z \*raiz (0,25/n) = 1,645 \* raiz(0,25/600) = 3,35%

1. Obtenha um intervalo de confiança para a proporção de clientes satisfeitos com os serviços oferecidos. Considerar 90 % de confiança.

(85% - 3,35% ; 85% + 3,35%) = (81,65%; 88,35%)

A cada 100 amostras de clientes, 90 delas terão a proporção de clientes satisfeitos entre 81,65% e 88,35%.

**REGRESSÃO LINEAR**

1. Uma empresa deseja saber se o **valor de sinistro de um seguro** está relacionado com **a renda do cliente**. O Modelo de regressão linear simples ajustado para explicar a relação entre estas variáveis é dado por:

|  |  |
| --- | --- |
| *Estatística de regressão* | |
| R múltiplo | 0,7797 |
| R-Quadrado | 0,607932 |
| R-quadrado ajustado | 0,591596 |
| Erro padrão | 5048,004 |
| Observações | 26 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coeficientes* | *Erro padrão* | *Stat t* | *valor-P* |
| Interseção | 32624,03 | 2072,292 | 15,74297 | 3,77E-14 |
| renda | -3,59298 | 0,588983 | -6,10032 | 2,66E-06 |

1. Qual o coeficiente de correlação linear entre as variáveis? Interprete o coeficiente.

0,7797. É uma correlação forte, acima de 0,7.

1. Qual o coeficiente de determinação? Interprete o coeficiente.

0,6079. O modelo capturou 60,79% da variabilidade dos dados.

1. Teste se a renda explica o valor do sinistro. Teste com 90 % de confiança?

Como o p-valor é < 0,1, a renda explica o sinistro.

1. Teste se o beta 0 deve fazer parte do modelo. Teste com 90 % de confiança?

Como o p-valor é <0,1, o beta 0 deve fazer parte do modelo.

1. Qual o valor de sinistro esperado quando a renda é R$2.500,00?

Y (sinistro) = 32624,03 – 3,59298\*renda

Y (sinistro) = 32624,03 – 3,59298\*2500

Y = 23.641,6

1. Um banco deseja obter um modelo para projetar o PIB de um município em função de várias variáveis. Com base no ajuste do modelo responda as questões.
2. Qual o coeficiente de determinação? Interprete o coeficiente.

O coeficiente é igual a 0,907214. Isso indica que o modelo capturou 90,7% da variabilidade dos dados.

1. Qual variável deve ser excluída em primeiro lugar? Justifique.

População > 60 anos, pois é a variável que possui maior p-valor, ou seja, possui menor relação com o PIB.

|  |  |
| --- | --- |
| *Estatística de regressão* | |
| R múltiplo | 0,952932 |
| R-Quadrado | 0,90808 |
| R-quadrado ajustado | 0,907214 |
| Erro padrão | 1356,933 |
| Observações | 644 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coeficientes* | *Erro padrão* | *Stat t* | *valor-P* |
| Interseção | -183,121 | 694,0552 | -0,26384 | 0,691987 |
| População > 60 anos | 7,224183 | 20,7104 | 0,348819 | 0,72734 |
| Habitantes | -0,00106 | 0,001552 | -0,68268 | 0,495055 |
| Área | -0,64583 | 0,170036 | -3,79821 | 0,00016 |
| Taxa de mortalidade | 31,42039 | 28,18665 | 1,114726 | 0,265388 |
| Esgoto | -2,17772 | 3,791788 | -0,57433 | 0,56595 |
| Emprego | 0,116091 | 0,004447 | 26,10379 | 1,1E-02 |

1. Qual o objetivo da análise de regressão linear?

Fazer uma previsão de uma variável em função de outras variáveis.

**ANÁLISE DISCRIMINANTE**

1. Qual o objetivo da análise discriminante?

Separar suas observações em grupos, discriminá-las.

1. Quais as 4 suposições da análise discriminante?

Normalidade multivariada

Ausência de outlier

Ausência de multicolinearidade

Homogeneidade das matrizes de variância e covariância

1. Teste a hipótese, considerar 90 % de confiança.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| H0: As populações possuem matrizes de variâncias iguais  H1: As populações possuem matrizes de variâncias diferentes | | **Test Results** | | | | --- | --- | --- | | Box's M | | 1,989 | | F | Approx. | ,587 | | df1 | 3 | | df2 | 99783,919 | | Sig. | ,623 | | Tests null hypothesis of equal population covariance matrices. | | | |

Como o p-valor é > 0,1 (0,63), não rejeito H0, ou seja, as populações possuem matrizes de variâncias iguais.

1. Teste a hipótese, considerar 90 % de confiança.

H0: Não há função de discriminação

H1: Há função de discriminação

| **Wilks' Lambda** | | | | |
| --- | --- | --- | --- | --- |
| Test of Function(s) | Wilks' Lambda | Chi-square | df | Sig. |
| 1 | ,415 | 15,825 | 2 | ,000 |

Como p-valor é < 0,1, rejeito H0, há função de discriminação.

1. Um diretor de relacionamento com o cliente deseja classificar os clientes de seguro de automóvel em clientes que reclamam – grupo 1 e em clientes que não reclamam – grupo 0 após um sinistro de automóvel. Com base no teste de igualdade de médias, para qual variável há evidência de que as médias são iguais nos dois grupos? Considerar 90 % de confiança.

| **Tests of Equality of Group Means** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | Wilks' Lambda | F | df1 | df2 | Sig. |
| tempo | ,454 | 22,862 | 1 | 19 | ,000 |
| Valor\_carro | ,990 | ,185 | 1 | 19 | ,672 |

Como o p-valor da variável Valor do carro é maior que 0,1, não rejeito H0, ou seja, as médias são iguais. Preciso retirá-la do modelo.

1. Suponha que, no exercício anterior, as médias do valor do carro para os dois grupos sejam diferentes. Escreva a função discriminante para classificar os grupos.

| **Standardized Canonical Discriminant Function Coefficients** | |
| --- | --- |
|  | Function |
| 1 |
| tempo | 1,046 |
| valorcarro | -,401 |

F = 1,046\*tempo – 0,401\*valor do carro

1. Obtenha o dc para a classificação das observações. Sabe-se que no grupo 0 tem-se 20 clientes e no grupo 1, tem-se 22 clientes.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **Functions at Group Centroids** | | | --- | --- | | reclama | Function | | 1 | | 0 | -1,077 | | 1 | 1,184 | | Unstandardized canonical discriminant functions evaluated at group means | | |

**dc = (20\*1,184 + 22\*-1,077)/(20+22)**

**dc = (23,68-20,923)/42 = -0,00033**

1. Suponha que, no exercício anterior, as médias do valor do carro para os dois grupos sejam diferentes. Classifique os 2 clientes como grupo 0 (não reclamam) ou 1 (reclamam).

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Tempo** | **Valor do carro** | **Classe** |
| **Cliente 1** | 10 | 30000 | 0 |
| **Cliente 2** | 45 | 50000 | 0 |

F Cliente 1 = 1,046\*10 – 0,401\*30000= -12.019,5 (como esse valor é < **-0,00033** então é um cliente que **NÃO** reclama)

F Cliente 2 = 1,046\*45 – 0,401\*50000= -20.002,9 (como esse valor é < **-0,00033** então é um cliente que **NÃO** reclama)

**REGRESSÃO LOGÍSTICA**

1. Qual o objetivo da regressão logística?

Estimar a probabilidade de uma observação pertencer à um grupo.

1. Descreva quatro exemplos de aplicação da regressão logística.
2. Estimar a probabilidade de um cliente comprar um produto no seu site.
3. Estimar a probabilidade de um cliente/transação ser ou não fraude.
4. Estimar a probabilidade de um cliente continuar comprando no seu supermercado por um ano.
5. Estimar a probabilidade de um cliente ter ou não um sinistro de automóvel.
6. Um diretor de RH tem o objetivo de classificar seus funcionários como funcionários que possuem **alto nível de estresse (p) e baixo nível de estresse (1-p)**. As variáveis consideradas no modelo foram: cargo, diretoria que o funcionário trabalha e distância entre a residência e o trabalho.
7. Quais variáveis não são importantes para explicar o nível de estresse do funcionário? Considerar na decisão 90 % de confiança.

| **Parameter** |  | **B** | **SE** | **Wald** | **Sig.** |
| --- | --- | --- | --- | --- | --- |
| Intercept |  | -10.1540 | 1.3107 | 60.0194 | <.0001 |
| Distancia |  | 0.6578 | 0.0876 | 56.3599 | <.0001 |
| Diretoria | A | 0.6077 | 0.5169 | 1.3821 | 0.2397 |
| Cargo | Analista | -3.0015 | 0.8890 | 11.3989 | 0.0007 |
| Cargo | Diretor | 4.3615 | 0.9972 | 19.1312 | <.0001 |

A variável diretoria não é importante para prever o stress do funcionário.

Após a retirada da variável Diretoria, foi obtida a equação abaixo:

| **Parameter** |  | **B** | **SE** | **Wald** | **Sig.** |
| --- | --- | --- | --- | --- | --- |
| Intercept |  | -9.8337 | 1.2562 | 61.2785 | <.0001 |
| Distancia |  | 0.6610 | 0.0881 | 56.3152 | <.0001 |
| Cargo2 | Analist | -3.0098 | 0.8690 | 11.9961 | 0.0005 |
| Cargo2 | Diretor | 4.1483 | 0.9795 | 17.9373 | <.0001 |

1. Obtenha a equação de p (clientes que pagam em dia o empréstimo) com base na tabela a seguir. **Escrever os coeficientes com 3 casas decimais**

P = 1 / (1+ e-(-9.8337+0.6610\*Distancia-3.0098\*Cargo Analista+4.1483\*Cargo Diretor))

1. Obtenha a probabilidade de estresse (p) para cada funcionário. Obter as probabilidades com 4 casas decimais.
2. Classifique como grupo 1 (funcionário que pode ter alto nível de stress) o funcionário com p > 0,5 e como grupo 0 (funcionário que pode ter baixo nível de stress) o funcionário com p < 0,5.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Funcionário | Diretoria | Distancia | Cargo | p | Grupo |
| 112 | 2 | 19,41 | Diretor | 0,999212208 | 1 |
| 137 | 1 | 12,83 | Gerente | 0,205465594 | 0 |

1. Considerando a seguinte tabela de classificação, qual o percentual de classificação correta?

|  | Observed | | Predicted | | |
| --- | --- | --- | --- | --- | --- |
|  | stress | |  |
|  | 0 | 1 |
| Step 1 | stress | 0 | 346 | 8 |  |
| 1 | 14 | 132 |  |
|  | |  |  |  |

Percentual = (346+132)/(346+132+8+14) = 95,6%

**ANÁLISE DE CLUSTER**

* 1. Qual o objetivo da análise de cluster?

Agrupar minhas observações (clientes, por ex.) em grupos de características iguais entre si.

* 1. Descreva quatro exemplos de aplicação da análise de cluster.
     1. Agrupar clientes que possuem mesmo tempo de casa e valor em grupos diferentes
     2. Agrupar agências semelhantes entre si
     3. Agrupar fundos de investimento semelhantes (mesma taxa, mesmo retorno, índices iguais)
     4. Agrupar clientes PJ de faturamentos, área, quantidade funcionários para analisa-los
  2. Quais os 2 métodos de agrupamento utilizados para gerar os clusters?
     1. Hierárquico
     2. K-means
  3. Qual o objetivo do dendograma?

Definir a quantidade de grupos.

* 1. O dendograma pode ser sempre gerado e facilmente interpretado?

Sim. Não, ele não é facilmente interpretado.

* 1. Por qual motivo as variáveis devem ser padronizadas antes da análise de cluster?

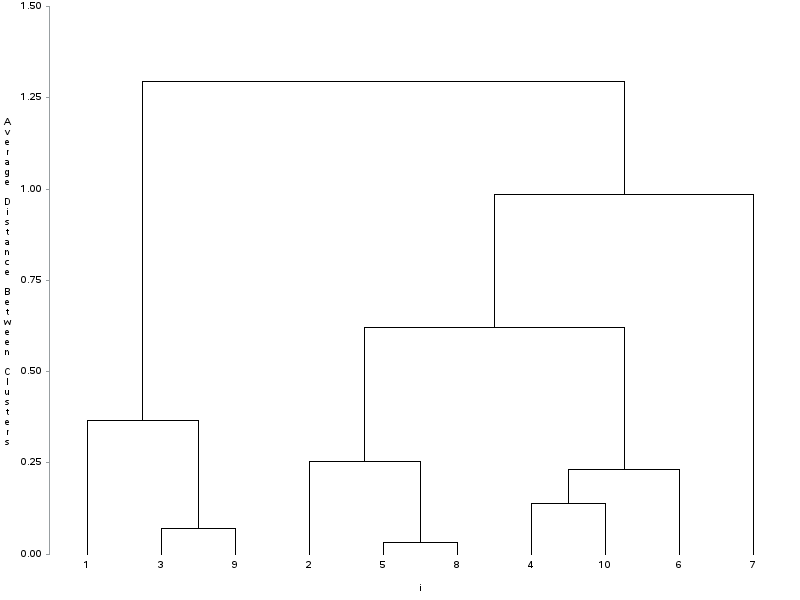
Para não termos variáveis com ordens de grandeza diferentes, assim, todas terão o mesmo peso.

* 1. Divida os clientes apresentados no dendograma em três grupos. Quais clientes estão em cada grupo?

Grupo 1: 1, 3 e 9

Grupo 2: 2, 5, 8, 4, 10, 6

Grupo 3: 7



1. Uma base de dados com 20.000 clientes de uma Telecom foi agrupada em 4 grupos por meio do método das k médias. Considere a análise de variância apresentada a seguir, responda as questões e justifique sua resposta. Considerar α = 0,05.
2. A variável **quantidade de produtos** separa os grupos? Não, pois o p-valor é > 0,05.
3. A variável **score** separa os grupos? Sim, pois o p-valor é < 0,05.

H0: As médias são iguais para os clusters

| **ANOVA** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | Cluster | | Error | | F | Sig. |
| Mean Square | Df | Mean Square | df |
| quantidade de produtos | 1,26361 | 3 | 0,90125 | 3 | 13,958 | ,543 |
| score | 10052,8 | 3 | 199069880,5 | 3 | 19802,5 | ,001 |

**TESTE QUI-QUADRADO**

1) Um executivo deseja testar se existe associação entre compra de automóvel e compra de residência.

|  |  |  |
| --- | --- | --- |
|  | Não possui residência | Possui residência |
| Não possui auto | 42 | 49 |
| Possui auto | 56 | 53 |

1. A estatística teste gerada foi de 0,5412. Obtenha a região de rejeição para testar a hipótese apresentada considerando uma confiança de 95%.

**Quando a estatística teste é maior que 3,84.**

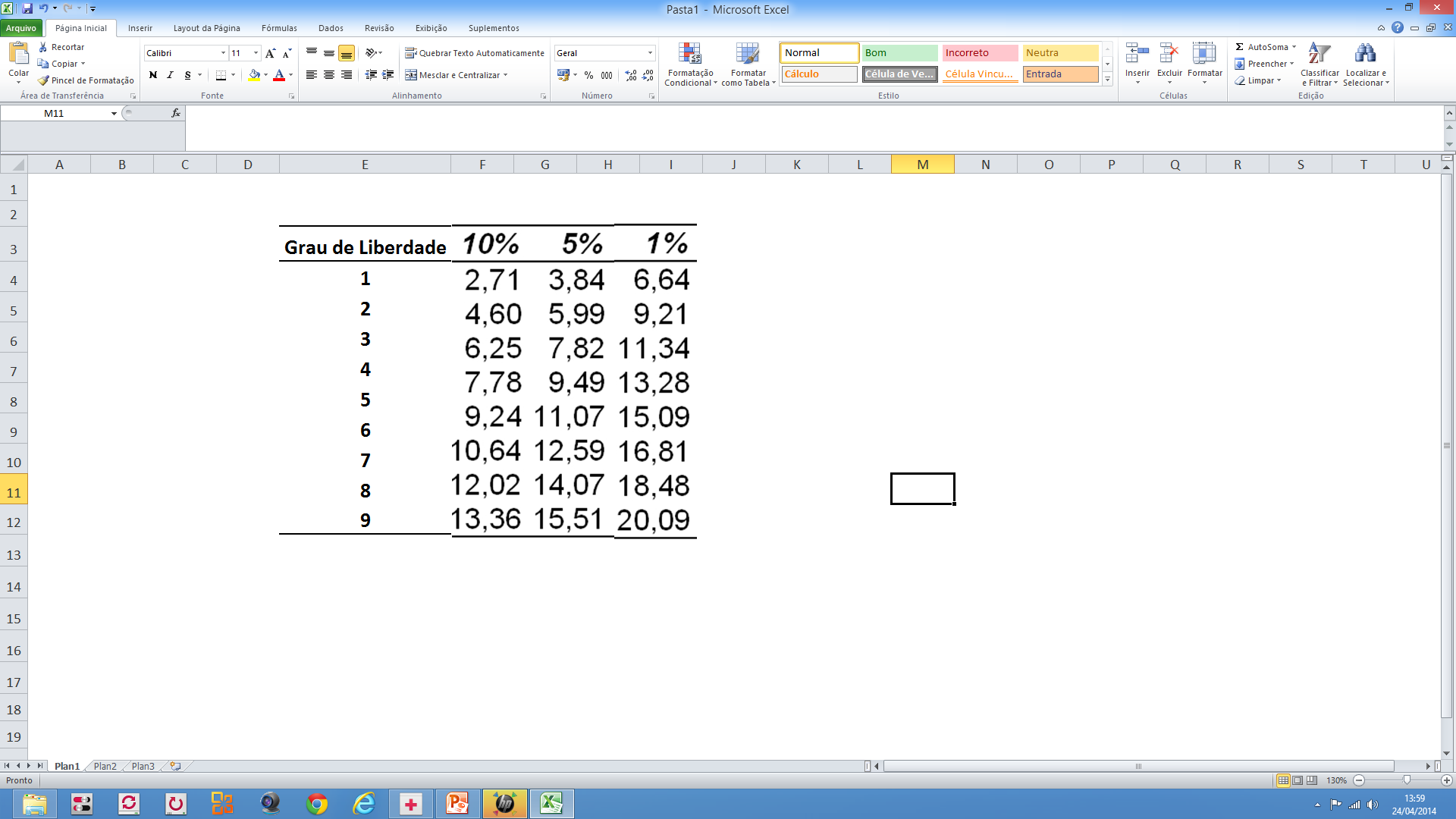
**Como 0,5412 é menor que 3,84, não rejeito (aceito) H0, logo, não existe associação. Só existiria associação se a estatística teste fosse maior que 3,84.**



H0: não existe associação

H1: existe associação

Gl = (numero de linhas -1)\*(numero de colunas – 1) = (2-1)\*(2-1) = 1



**ÁRVORE DE DECISÃO**

Um diretor do produto de planos de celulares deseja abordar clientes para oferecer um *up-sell* em um plano pós-pago. Ele fez uma árvore de decisão utilizando as variáveis de renda, reclamações e se o cliente “estourou” seu plano de internet no último ano.

1. Para quais grupos você ofereceria o aumento de plano?

