UNIDADE 3 – ANÁLISE BIDIMENSIONAL E ANÁLISE DE REGRESSÃO
MÓDULO 1 – ANÁLISE BIDIMENSIONAL – GRAUS DE LIBERDADE

01

1 - CARACTERIZAÇÃO DA ANÁLISE BIDIMENSIONAL

Nos estudos anteriores nossa análise recaiu sobre o comportamento de uma variável estudada de forma "individual", ou seja, não relacionada com outra(s).

A partir de agora nossa atenção estará voltada para análise de duas variáveis observadas simultaneamente em um conjunto de indivíduos/objetos, de forma que os diferentes "níveis" das duas variáveis aparecerão cruzados/inter-relacionados, mostrando, efetivamente, um comportamento conjunto.

Existem dificuldades práticas para geração das medidas descritivas quando a variável não é tipicamente quantitativa. Essa técnica de análise permite que variáveis quantitativas e qualitativas (tanto ordinais como nominais) possam ser tratadas, em três diferentes combinações, como:

(a) duas variáveis quantitativas;
(b) uma variável quantitativa e outra qualitativa;
(c) duas variáveis qualitativas.

02

O exemplo abaixo nos mostra a relevância desta abordagem.

Um anunciante com uma grande verba para investir em publicidade suspeita que a preferência dos leitores por revistas semanais de circulação nacional independe das regiões geográficas onde moram. Considerando que uma nova revista semanal entrou em circulação, imagine que, após algumas semanas de vendas, fez-se uma pesquisa com

9000 leitores de três regiões do país (3000 em cada uma delas).

Cada um daqueles leitores, depois de ler as edições de um mês de cada revista, deveria escolher uma delas, manifestando sua preferência. Os resultados foram:

REGIÕES	REVISTAS			TOTAL
	A	B	C	
Norte - Nordeste	750	500	1750	3000
Leste	1200	850	950	3000
Centro-Sul	1050	1100	850	3000
Total	**3000**	**2450**	**3550**	**9000**

Se você for contratado para dar seu parecer para este anunciante, o que você concluiria em seu relatório? A premissa do anunciante está correta? Por quê?

03

Essa é uma situação que envolve a análise de duas variáveis simultaneamente (o comportamento conjunto de duas variáveis, sendo que ambas são qualitativas nominais: região geográfica e preferência por uma revista). Aqui podemos estabelecer duas hipóteses que devem ser avaliadas:

H_0; a preferência por uma revista independe da região de residência do leitor (as variáveis estudadas são independentes ou não relacionadas);

H_1: a preferência por uma revista está associada/relacionada à região de residência do leitor (as variáveis são dependentes).

O primeiro passo é ter uma ideia dos percentuais associados à realidade apresentada, uma vez que as frequências absolutas, tais como foram lançadas, podem induzir uma visão equivocada do que de fato está acontecendo.

É importante estabelecer com clareza que, caso haja relacionamento/associação entre as variáveis em questão, a preferência estaria sendo impactada pela região de residência do leitor e não o contrário.

Também é muito importante deixar suficientemente claro que em nenhum momento está sendo insinuado que uma das hipóteses está vinculada à não existência de uma preferência bem caracterizada por alguma das revistas.

O que isso quer dizer? Quer dizer que não estamos querendo saber qual a revista seria mais lida, ou seria a melhor escolha. O que estamos querendo saber é **se a escolha da revista está ou não relacionada com a região do leitor.**

Assim, o que está em questão é se o perfil de preferência nas três regiões geográficas pode ser considerado aproximadamente o mesmo (o que caracteriza um quadro de não relacionamento, contemplado pela hipótese H0) ou se há padrões de preferência distintos para as regiões investigadas.

Feitas essas considerações, a tabela com as frequências **percentuais** ficaria:

REGIÕES	REVISTAS			TOTAL
	A	B	C	
NORTE-NORDESTE	25,00%	16,67%	58,33%	100%
LESTE	40,00%	28,33%	31,67%	100%
CENTRO-SUL	35,00%	36,67%	28,33%	100%
TOTAL	33,33%	27,22%	39,44%	100%

Veja como cada um dos percentuais acima foi obtido.

Revista A, Região Norte-Nordeste: 750/3000 = 0,25 = 25%

Revista A, Região Leste: 1200/3000 = 0,40 = 40%

Revista A, Região Centro-Sul: 1050/3000 = 0,35 = 35%

Revista B, Região Norte-Nordeste: 500/3000 = 0,1667 = 16,67%

Revista B, Região Leste: 850/3000 = 0,2833 = 28,33%

Revista B, Região Centro-Sul: 1100/3000 = 0,3667 = 36,67%

Revista C, Região Norte-Nordeste: 1750/3000 = 0,5833 = 58,33%

Revista C, Região Leste: 950/3000 = 0,3167 = 31,67%

Revista C, Região Centro-Sul: 850/3000 = 0,2833=28,33%

Total de Leitores, Revista A: 3000/9000 = 0,3333 = 33,33%

Total de Leitores, Revista B: 2450/9000 = 0,2722 = 27,22%

Total de Leitores, Revista C: 3550/9000 = 0,3944 = 39,44%

05

Essa tabela parece revelar algum grau de dependência/relacionamento entre as variáveis *região de residência do leitor e preferência por revista semanal de circulação nacional*, uma vez que, se não houvesse, seriam esperadas proporções próximas a 33,33%; 27,22% e 39,44% em cada uma das regiões (de tal modo que não seria possível a identificação da região, caso fosse dado o perfil de preferência, ou, em outras palavras, para todas as regiões valeria o padrão de preferência revelado para o Brasil como um todo).

Dito de outra forma, se a distribuição de preferência considerando todas as regiões é 33,33%; 27,22% e 39,44%, então caso a preferência fosse independente da região, encontraríamos essa mesma proporção em todas as regiões. Por exemplo, para a revista A na região Norte-Nordeste, teríamos 33,33% do total da região, ou seja, 33,33% * 3000 = 1000. Para a revista B teríamos 27,22% do total, nesse caso 27,22%*3000 = 817 e, por fim, para a revista C teríamos 39,44% * 3000 = 1183.

Assim, as respectivas frequências **hipotéticas** para essa situação (de independência) seriam:

REGIÕES	REVISTAS			TOTAL
	A	B	C	
NORTE-NORDESTE	1000	817	1183	3000
LESTE	1000	817	1183	3000
CENTRO-SUL	1000	817	1183	3000
TOTAL	3000	2450	3550	9000

Veja como cada uma das frequências desse quadro foi obtida.

Revista A, Região Norte-Nordeste : $0,33 \times 3000 = 1000$
Revista A, Região Leste : $0,2722 \times 3000 = 817$
Revista A, Região Centro-Sul : $0,3944 \times 3000 = 1183$
Revista B, Região Norte-Nordeste : $0,33 \times 3000 = 1000$
Revista B, Região Leste : $0,2722 \times 3000 = 817$
Revista B, Região Centro-Sul : $0,3944 \times 3000 = 1183$
Revista C, Região Norte-Nordeste : $0,33 \times 3000 = 1000$
Revista C, Região Leste : $0,2722 \times 3000 = 817$
Revista C, Região Centro-Sul : $0,3944 \times 3000 = 1183$

06

Outra observação que se torna necessária, por ser oportuna e muito importante, é que o fato de haver frequências idênticas nas três regiões (nesse caso, 3000 leitores) foi meramente ilustrativo e com efeito didático. Na prática, isso pode ocorrer ou não, e o procedimento até agora adotado fica inalterado.

Na situação que está sendo discutida, fica nítido que há "desvios" entre as frequências observadas (**reais**) e as frequências esperadas (**hipotéticas**, no caso da independência ser verdadeira). Devemos ser capazes de responder se estas diferenças são suficientemente grandes ou pequenas (a resposta a essa pergunta vai efetivamente validar uma das duas hipóteses levantadas).

É necessário, então, definir um critério objetivo que nos auxilie nessa tarefa. Por analogia com a lógica utilizada para a definição de variância, podemos chegar à seguinte medida:

$$\sum_i \sum_j \left[\frac{(frequencia\ observada_{ij} - frequencia\ hipotética_{ij})^2}{frequencia\ hipotética_{ij}} \right]$$

O que está sendo indicado é que se deve tomar cada frequência de fato observada e subtrair a correspondente frequência hipotética. Em seguida, deve-se elevar esse resultado ao quadrado para que não se tenham valores negativos, como no cálculo de variância, e então dividir pela frequência hipotética para que se tenha uma ideia mais adequada da diferença entre as frequências.

O raciocínio poderia ser também assim.

Se a diferença entre uma frequência real e sua correspondente frequência hipotética resulta 100, pode ser difícil interpretar se isto significa uma grande diferença ou uma pequena diferença entre as frequências, afinal 100 como resultado de 3100 - 3000 tem um significado e como resultado de 300 - 200 tem significado bastante distinto, logo, ao dividir essa diferença por uma referência (nesse caso, a própria frequência hipotética), tem-se uma ideia muito melhor de sua ordem de grandeza.

Como a pretensão é estabelecer uma medida que reflita a diferença entre duas situações (dois conjuntos de dados) e não apenas entre "pares" específicos, faz-se necessário somar todas as parcelas, sendo que, no caso sob análise, são nove, uma vez que há três colunas (correspondentes às três revistas) e três linhas (correspondentes às três regiões geográficas).

07

Passemos agora ao cálculo desse "indicador":

$$\frac{(750-1000)^2}{1000}+\frac{(500-817)^2}{817}+\frac{(1750-1183)^2}{1183}+\frac{(1200-1000)^2}{1000}+\frac{(850-817)^2}{817}+$$
$$+\frac{(950-1183)^2}{1183}+\frac{(1050-1000)^2}{1000}+\frac{(1100-817)^2}{817}+\frac{(850-1183)^2}{1183}=$$

Cujo resultado final fica:

$$62,5+123+271,76+40+1,33+45,89+2,5+98,03+93,74=738,75$$

É indispensável, agora, concluir se o valor de 738,75 (unidades de medida) reflete uma grande diferença entre a situação real e a hipotética ou, se na verdade, a diferença não pode ser considerada significativa. Para que essa decisão seja possível, será necessário abordar dois novos aspectos teóricos:

Graus de liberdade **Distribuição qui-quadrado**

Estudaremos esses conceitos a seguir.

08

2 - Graus de Liberdade

Para termos uma noção intuitiva do que vem a ser o número de graus de liberdade, é indicado realizar a análise de uma tabela de dupla entrada.

Considere que seja apresentada uma tabela contendo apenas os totais parciais, como mostrado a seguir:

REGIÕES	REVISTAS			TOTAL
	A	B	C	
NORTE-NORDESTE				3000
LESTE				3000
CENTRO-SUL				3000
TOTAL	3000	2450	3550	9000

Os totais parciais espelham com fidelidade o resultado da pesquisa realizada quando analisada cada variável separadamente. O que está em questão é o comportamento conjunto das duas variáveis.

09

Considere agora que podemos arbitrar um valor para preenchimento do cruzamento: revista A e região Norte-Nordeste. Qualquer número entre 0 e 3000 pode ser escolhido, uma vez que não estará ultrapassando os limites impostos pelos totais parciais (tanto total da linha como da coluna).

Admita que seja escolhido o número 500 (como poderia ter sido 600, 700 ou outra opção):

REGIÕES	REVISTAS			TOTAL
	A	B	C	
NORTE-NORDESTE	500			3000
LESTE				3000
CENTRO-SUL				3000
TOTAL	3000	2450	3550	9000

Façamos agora o mesmo para o cruzamento da revista A com região geográfica Leste. Embora já não tenhamos tantas opções de números, pois a primeira célula já foi preenchida, continuamos com um bom leque de alternativas. Admita, mais uma vez, por hipótese, que seja escolhido o número 900:

REGIÕES	REVISTAS			TOTAL
	A	B	C	
NORTE-NORDESTE	500			3000
LESTE	900			3000
CENTRO-SUL				3000
TOTAL	3000	2450	3550	9000

10

O que você acha que acontece agora para o preenchimento do cruzamento revista A e região Centro-Sul? Existe alguma liberdade de escolha do número que irá ocupar aquela célula?

Caso queiramos preservar os totais da primeira coluna e da terceira linha não há alternativa que não seja o número 1600. Assim, podemos dizer que, ao preencher "com liberdade" as duas primeiras células, a terceira ficou "presa", "automaticamente" condicionada a elas.

Esse raciocínio pode continuar sendo desenvolvido agora para o cruzamento: revista B e região Norte-Nordeste. Mais uma vez, embora não haja tantas opções disponíveis (se compararmos com as possibilidades da primeira célula, na qual inserimos o 500), há um número considerável de alternativas, e podemos, por exemplo, optar por 1200. A tabela ficaria:

REGIÕES	REVISTAS			TOTAL
	A	B	C	
NORTE-NORDESTE	500	1200		3000
LESTE	900			3000
CENTRO-SUL				3000
TOTAL	3000	2450	3550	9000

11

Se pensarmos no preenchimento do cruzamento revista C e região Norte-Nordeste, é fácil constatar que só há uma possibilidade que atende corretamente à condição de 3000 leitores naquela região, a saber, o número 1300.

Mais uma vez, não há "liberdade" de preenchimento para aquela célula. Para o cruzamento revista B e região Leste, por analogia com as demais situações, há diversas possibilidades numéricas que podem ser consideradas.

Pode-se arbitrar, por exemplo, 1100 e assim a tabela estaria completa, pois tanto o cruzamento revista B e região Centro-Sul, como o cruzamento revista C e região Centro-Sul ficariam "atrelados":

REGIÕES	REVISTAS			TOTAL
	A	B	C	
NORTE-NORDESTE	500	1200		3000
LESTE	900	1100		3000
CENTRO-SUL				3000
TOTAL	3000	2450	3550	9000

12

A constatação que se quis passar é que em uma tabela com três linhas e três colunas, uma vez fixados/arbitrados quatro valores, os outros cinco ficam a eles condicionados, e fica visível que a "perda de liberdade" se dá em uma linha e uma coluna:

REGIÕES	REVISTAS			TOTAL
	A	B	C	
NORTE-NORDESTE	500	1200		3000
LESTE	900	1100		3000
CENTRO-SUL				3000
TOTAL	3000	2450	3550	9000

As células hachuradas significam apenas que estão "presas", condicionadas ao preenchimento das demais. Logo, pelo exposto:

> Número de graus de liberdade = (número de linhas - 1) X (número de colunas - 1)

No caso em questão, como há 3 linhas e 3 colunas: (3 - 1) x (3 - 1) = 2 x 2 = 4

Como visto no cálculo da medida de diferença entre as situações real e hipotética, havia a soma de nove parcelas, fruto da existência de três linhas e três colunas. O número de graus de liberdade está vinculado ao tamanho da tabela, ou seja, ao número de linhas e colunas, uma vez que parece razoável supor que à medida que o número de células aumenta, a medida da diferença entre as duas situações (real e hipotética) tende a aumentar. Isso significa que o resultado da diferença para uma determinada tabela pode ter interpretação distinta do mesmo resultado para outra tabela maior ou menor.

13

RESUMO

Na análise bidimensional, é estudado o comportamento conjunto de duas variáveis, que tanto podem ser quantitativas como qualitativas. O objetivo maior é concluir se há independência ou dependência entre elas (relacionamento ou não).

Dada uma situação para a qual foram observadas frequências relativas ao cruzamento das diferentes respostas de uma variável com as diferentes respostas de outra, parte-se para a determinação de frequências percentuais, com o cuidado de verificar qual variável influencia o comportamento da outra, caso exista relacionamento entre elas.

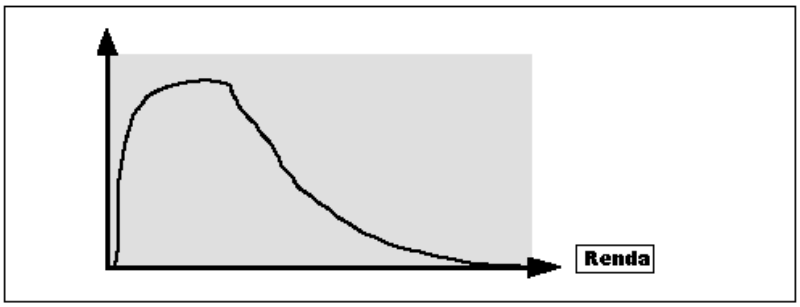
Feito isto, parte-se para a determinação de um conjunto de frequências hipotéticas, admitindo-se que a independência fosse verdadeira. Isso permitirá uma comparação com as frequências verdadeiras, a partir do cálculo de uma medida da diferença entre os conjuntos de frequências. Essa medida, denominada qui-quadrado servirá para analisarmos se existe um relacionamento entre as duas variáveis e será desenvolvida no próximo módulo.

Vimos ainda que só poderemos interpretar corretamente o resultado obtido conhecendo o número de **graus de liberdade**.

> Número de graus de liberdade = (número de linhas - 1) X (número de colunas - 1)

O número de graus de liberdade está vinculado ao tamanho da tabela, isto é, ao número de linhas e colunas, uma vez que parece razoável supor que à medida que o número de células aumenta, a medida da diferença entre as duas situações (real e hipotética) tende a aumentar. Isso significa que o resultado da diferença para uma determinada tabela pode ter interpretação distinta do mesmo resultado para outra tabela maior ou menor.

14

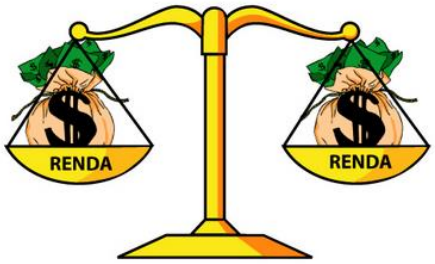


O que está sinalizado é que à medida que a renda aumenta, a quantidade de pessoas que ganha aquela quantia vai diminuindo sensivelmente. Esse aspecto é típico de uma distribuição denominada qui-quadrado (muito embora o gráfico desse tipo de distribuição possa ser bastante distinto, dependendo do número de graus de liberdade). O símbolo adotado para essa distribuição é:

X^2

15

A decisão, se uma determinada renda é considerada grande ou pequena, passa por uma avaliação comparativa, afinal um mesmo salário pode ser considerado baixo em um contexto e alto em outro. Assim, ao indagar se um salário de 50 unidades monetárias, por exemplo, é alto ou baixo, deve-se ter uma ideia da frequência percentual que fica abaixo e acima desse valor.



Caso tenha-se uma proporção de 95 ou 99% abaixo do valor de referência (e, consequentemente, 5 ou 1% acima, respectivamente), teremos de admitir que aquele valor pode ser considerado alto. O procedimento para a avaliação no nosso contexto é análogo: se formos capazes de determinar qual o percentual que fica acima e qual fica abaixo do valor calculado, teremos um indicador nítido de sua ordem de grandeza.

É com base nesse conceito que iremos aplicar a **distribuição qui-quadrado** na análise bidimensional, pois vimos no módulo anterior que chegamos a calcular um valor que representava a diferença entre a frequência real e a hipotética, lembra-se? Pois bem, o problema é que não sabíamos se o valor obtido

era grande ou pequeno. Se a diferença fosse grande, então a hipótese que dizia que a preferência variava com a região seria confirmada. Caso o valor fosse pequeno, ou seja, a diferença entre as frequências reais e as hipotéticas fosse pequena, então poderíamos concluir que a preferência não variava com a região do leitor.

16

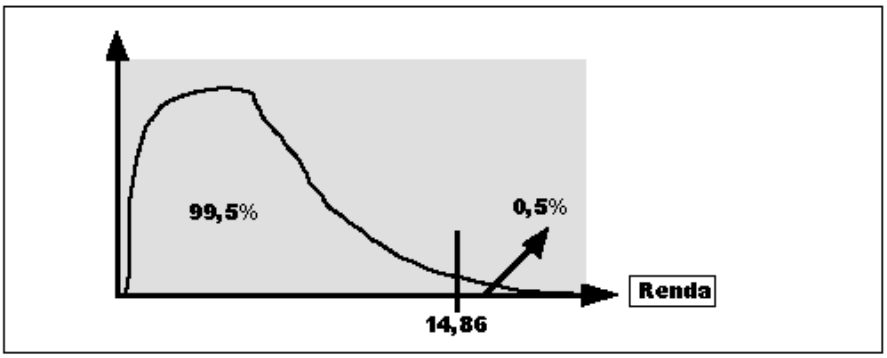
Pode-se, então, optar entre duas alternativas para determinação se o valor calculado para a diferença entre a tabela real e a hipotética é suficientemente grande ou não:

(a) recorrer à tabela qui-quadrado, disponível na grande maioria dos livros de Estatística Básica (Mario Triola; Wilton Bussab; Pedro Morettin). O primeiro passo é determinar o número de graus de liberdade para, em seguida, consultar na linha correspondente a esse número um conjunto de probabilidades que irão sinalizar como posicionar/avaliar o valor calculado;

Exibimos, a seguir, um trecho da tabela A-4 que se encontra no livro de Mario Triola, (1999) pág. 356:

Graus de liberdade	Área à direita do valor crítico									
	0,995	0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01	0,005
3	0,072	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	9,236	11,071	12,833	15,086	16,750

Constata-se que, para 4 graus de liberdade, o valor 14,860 "quebra" a distribuição qui-quadrado deixando 99,5% dos valores à sua esquerda e 0,5% à sua direita.

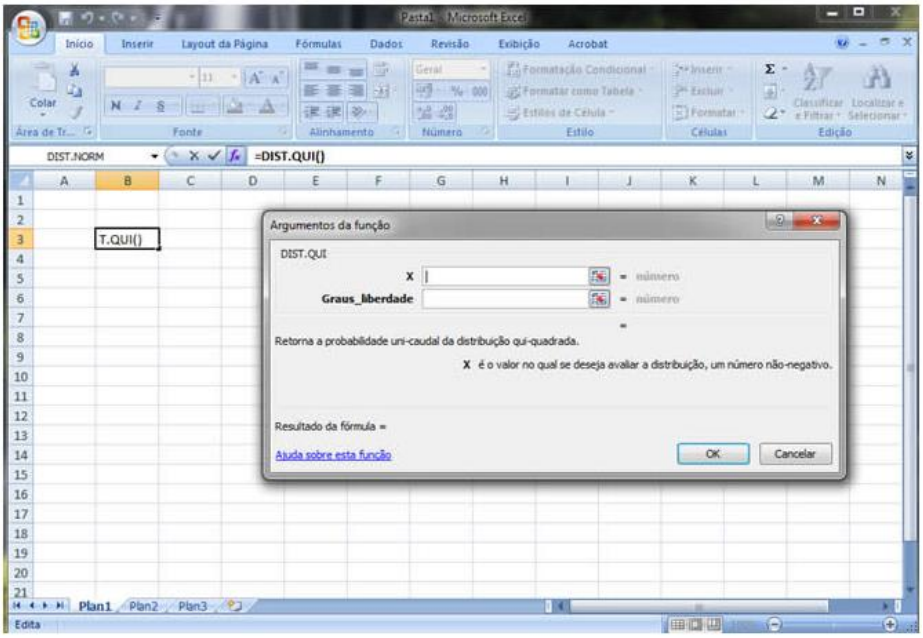


Consequentemente, o próprio 14,86 já poderia ser considerado um valor suficientemente grande para esse caso. O que dizer então do valor resultante de nossa medida, ou seja, 738,75? Parece inquestionável que esse valor seja de fato muito grande. Logo, como a diferença "medida" entre as duas tabelas (real e hipotética) é muito grande, e a situação hipotética parte do pressuposto de independência das variáveis, nossa conclusão é:

Para o caso sob análise, a hipótese de independência deve ser rejeitada, admitindo-se então que existe relacionamento/dependência entre a região de residência do leitor e a preferência pela revista semanal de circulação nacional.

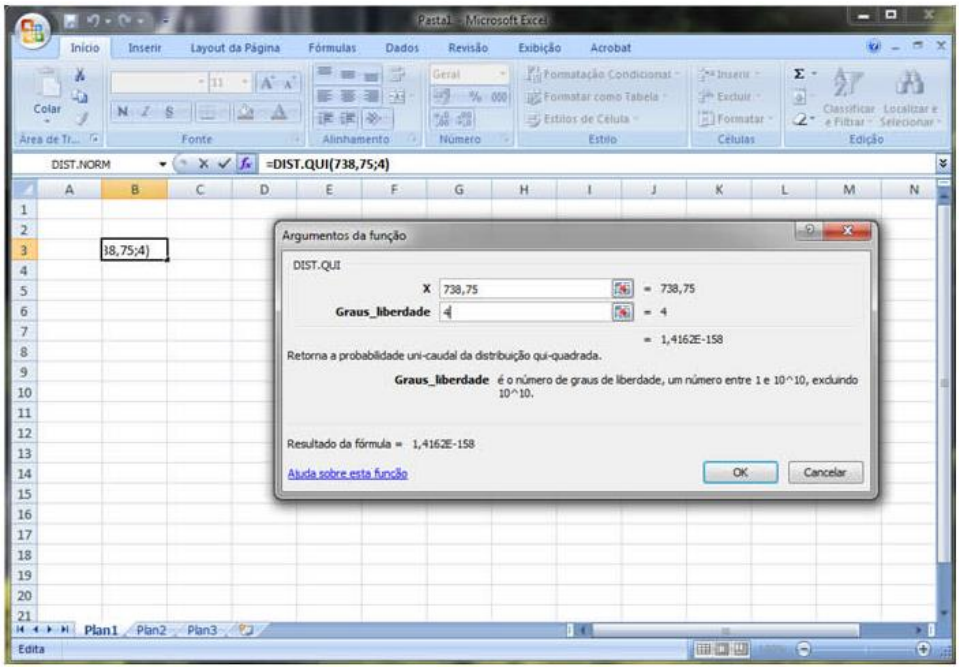
17

(b) recorrer à planilha Excel, com essa sequência de comandos, após posicionar o cursor na célula na qual se deseja a inserção do resultado: fx, Estatística, DIST.QUI, OK, e então teremos a seguinte janela.



18

Deve-se agora digitar o valor calculado de 738,75 ao lado do X e digitar 4 no campo destinado aos graus de liberdade, em seguida OK (observe que antes de clicar em OK já é possível visualizar o resultado).



O resultado mostrado pela planilha revela que a frequência relativa à esquerda do valor 738,75 é muitíssimo baixa, tanto é que foi usada a notação exponencial.

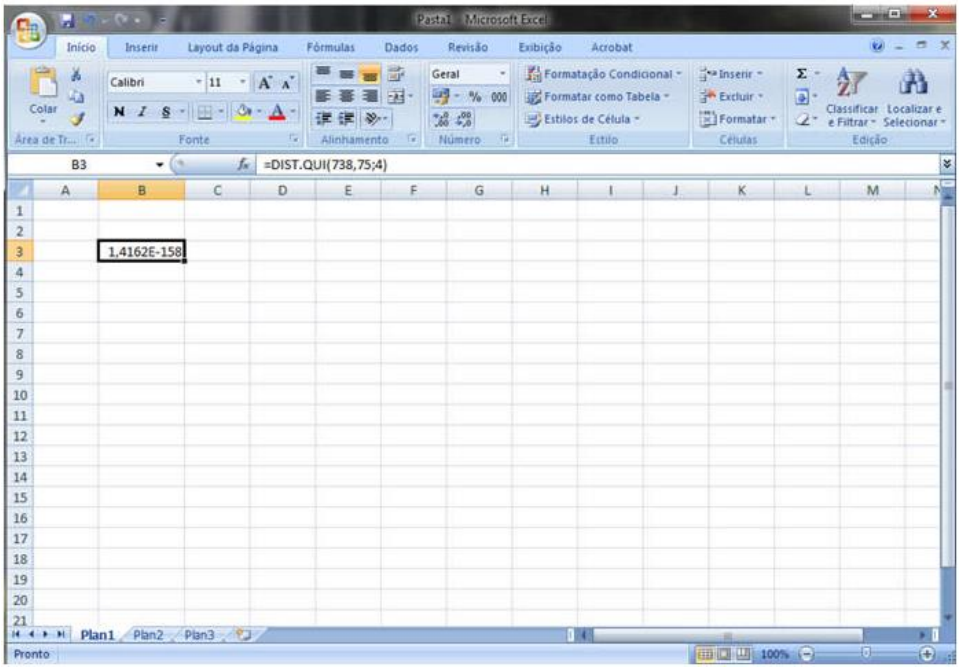
àrea à direita de 738,75 = 1,4162 x 10-158

19

Você consegue imaginar quão pequeno é esse número? Logo só há uma conclusão possível:

Para o caso sob análise, a diferença encontrada entre a situação real e a situação hipotética é muito grande, então a independência deve ser rejeitada, admitindo-se como verdadeira a hipótese alternativa de relacionamento/dependência entre as variáveis região de residência do leitor e preferência pela revista semanal de circulação nacional.

O valor calculado para a diferença entre a situação real e a hipotética recebe o nome de **qui-quadrado calculado**.



20

Dizer que existe relacionamento entre as variáveis estudadas, embora já agregue significativo valor à nossa análise, ainda deixa margem a uma pergunta muito interessante: que associações foram mais relevantes (mais contribuíram) para a rejeição da independência?

Para responder devem ser consideradas as maiores parcelas da soma que gerou o qui-quadrado calculado (expressam as maiores diferenças entre a situação hipotética de independência e a situação real), até o limite dado pelo número de graus de liberdade. No exemplo dado, são quatro graus de

liberdade, então devem ser analisadas as quatro maiores parcelas da soma. Isso não deve ser visto de forma cartesiana, mas sim como uma regra que deve orientar a análise/interpretação da situação, pois a 5ª maior parcela pode ser muito próxima da 4ª maior, merecendo, então, ser incluída no rol das análises mais pontuais. Por outro lado, apenas as três maiores poderiam "esgotar" praticamente a totalização do qui-quadrado calculado, não sendo necessário fazer uma análise específica de mais uma parcela.
Voltando à expressão que gerou o 738,75:

$$62{,}5 + 123 + 271{,}76 + 40 + 1{,}33 + 45{,}89 + 2{,}5 + 98{,}03 + 93{,}74 = 738{,}75$$

271,76 ⇒ referente ao cruzamento revista C e região Norte - Nordeste;
123 ⇒ referente ao cruzamento revista B e região Norte - Nordeste;
98,03 ⇒ referente ao cruzamento revista B e região Centro - Sul;
93,74 ⇒ referente ao cruzamento revista C e região Centro - Sul.

21

Partindo para as interpretações específicas, as quatro maiores parcelas são, pela ordem:

1. Número de leitores da região Norte-Nordeste que prefere a revista C é significativamente **maior** do que aquele que seria esperado em uma situação de independência, ou seja, caso o padrão nacional de preferência valesse para todas as regiões indistintamente. Saiba +
2. Número de leitores da região Norte-Nordeste que prefere a revista B é significativamente **menor** do que aquele que seria esperado em uma situação de independência.
3. Número de leitores da região Centro-Sul que prefere a revista B é significativamente **maior** do que aquele que seria esperado em uma situação de independência.
4. Número de leitores da região Centro-Sul que prefere a revista C é significativamente **menor** do que aquele que seria esperado em uma situação de independência.

Também se diz que a hipótese de independência foi rejeitada com confiança superior a 99,5%. Ou seja, o risco de nossa conclusão estar equivocada é muito pequeno. Em outras palavras, para um nível de confiança de 99,5% o valor do qui-quadrado tabelado, para 4 graus de liberdade, é 14,86. Como o qui-quadrado calculado resultou um valor superior a este, deve-se rejeitar a hipótese H_0 (de independência) para esse nível de confiança.

Saiba +

É óbvio que nunca houve dúvida de 1750 leitores ser um número superior a 1183 leitores. O que se está dizendo agora é que o primeiro é significativamente maior, o que talvez não fosse possível garantir, caso tivéssemos obtido, por exemplo, 1650 como frequência hipotética.

22

2 - "Medindo" a Dependência

Uma vez rejeitada a independência, isto é, existe associação entre as variáveis, já se tem um "sentimento" a respeito do grau de associação a partir do valor do qui-quadrado calculado (ou seja, quanto maior for o valor calculado para o qui-quadrado, maior o grau de associação existente). Isso, no entanto, pode soar um tanto subjetivo e de difícil dimensionamento do que venha a ser uma alta ou baixa associação. Assim, Pearson desenvolveu o coeficiente de contingência C, definido por

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Onde **n** é o número de observações.

Do ponto de vista teórico, esse coeficiente é um número entre zero e um. Fica claro que o valor zero ocorre quando se tem um caso de independência "total" (ou "perfeita"). Uma variação possível para C é

$$C^* = \frac{C}{\sqrt{(t-1)/t}}$$

Onde **t** = mínimo entre o número de colunas e o número de linhas da tabela.

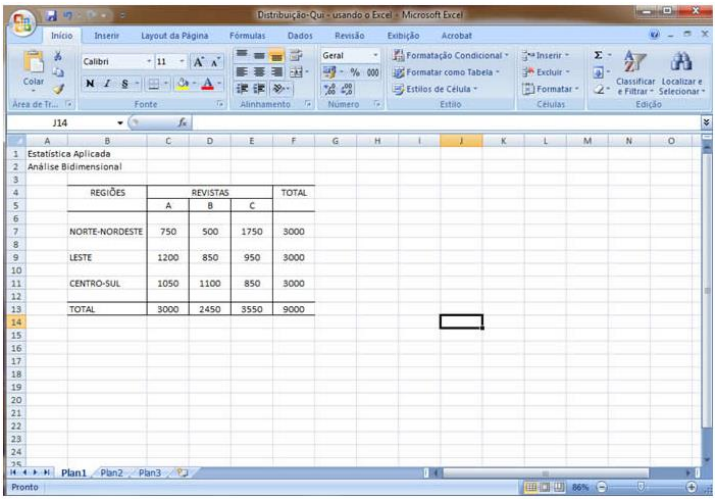
Do ponto de vista prático, o percentual à direita do valor do qui-quadrado calculado acaba sendo um "termômetro" muito mais calibrado a respeito da intensidade da dependência presente entre as variáveis.

23

3 - UTILIZANDO O EXCEL

Mais uma vez a planilha Excel pode ser um recurso valioso para a prática da análise bidimensional, permitindo, particularmente, um ganho expressivo de tempo, principalmente quando as tabelas são de porte maior do que a apresentada anteriormente.

Uma vez aberta uma planilha, deve-se entrar com os dados em um formato similar ao apresentado a seguir:

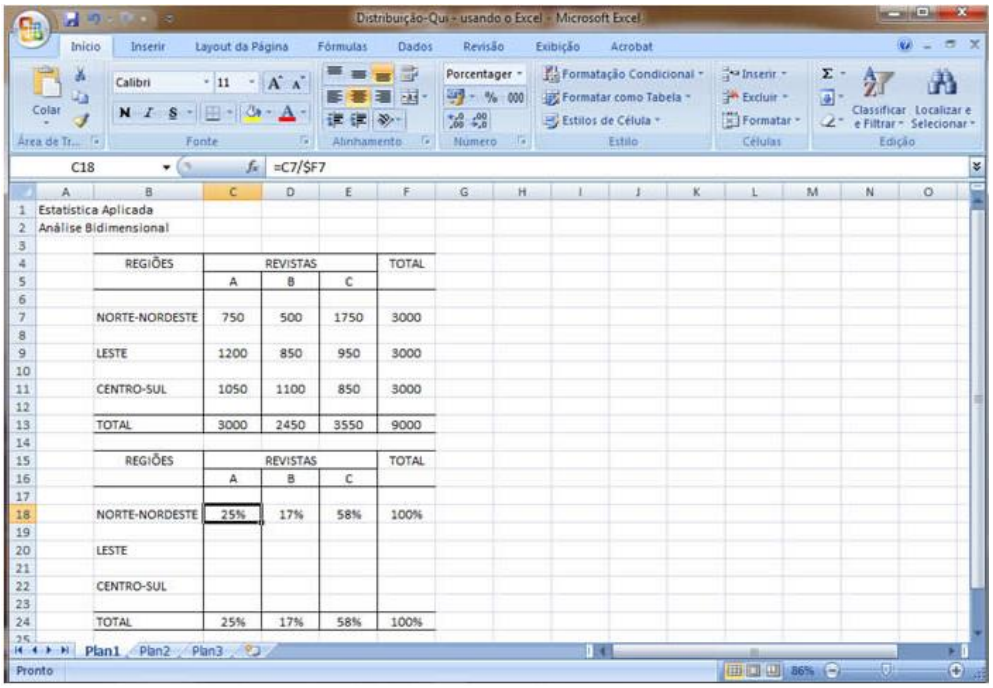


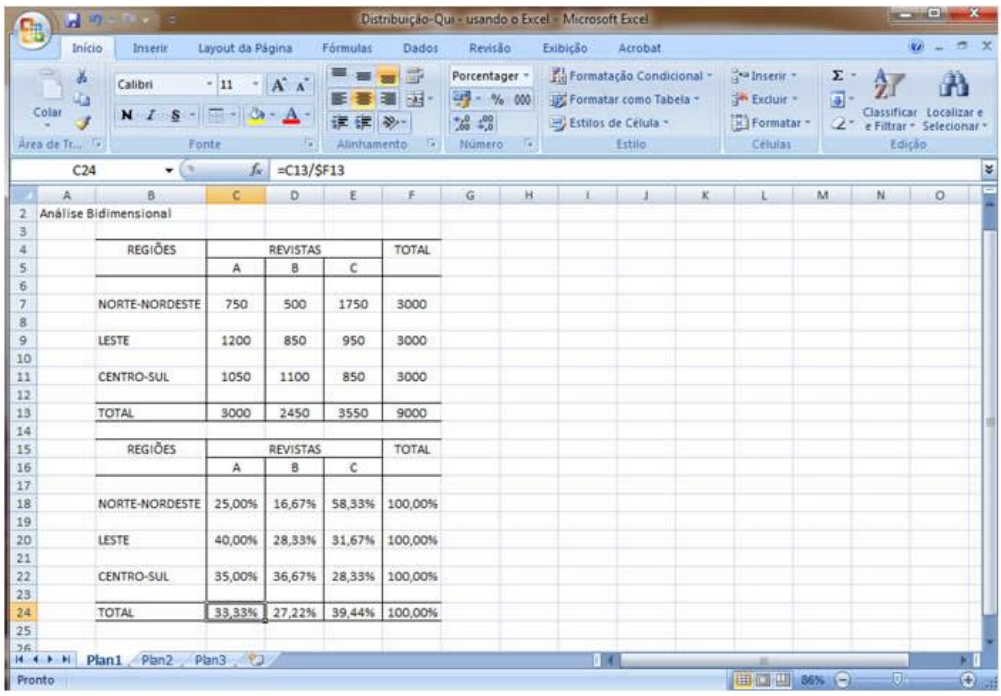
Note que, para a linha e a coluna destinadas à inserção dos totais, deve-se introduzir as fórmulas adequadas para que o cálculo das somas seja executado.

24

Em seguida, prepara-se uma tabela com as frequências percentuais (lembrando a necessidade de considerar qual variável seria condicionada pela outra, caso houvesse relacionamento das variáveis estudadas).

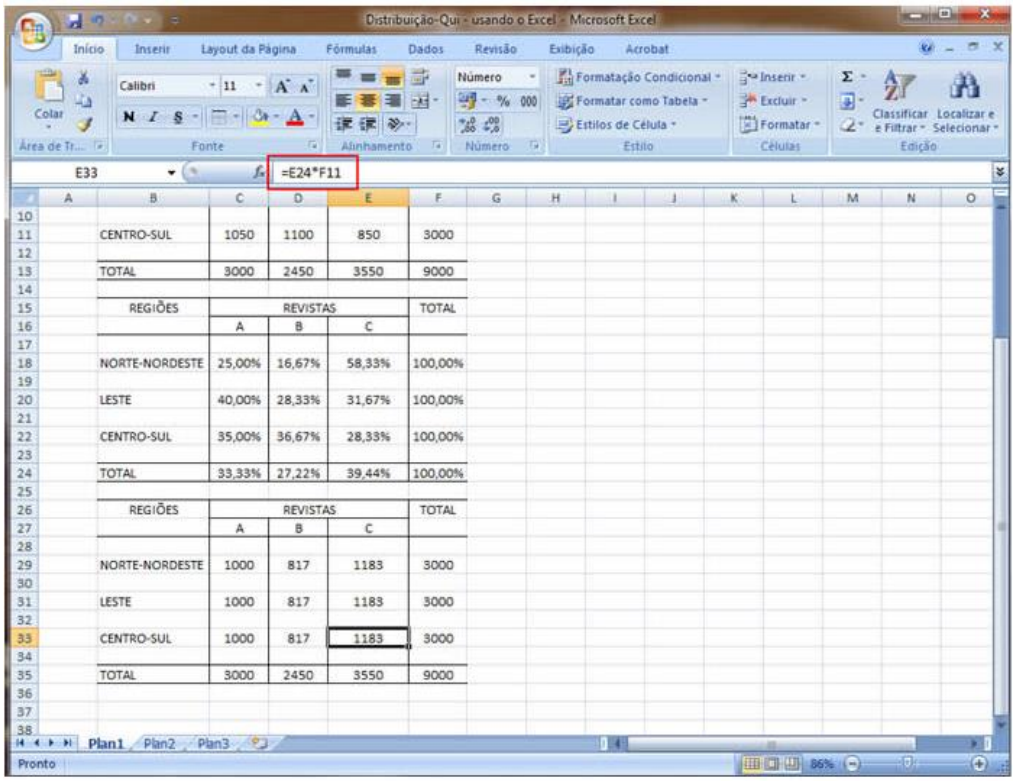
Observe atentamente como foi obtido cada percentual das planilhas apresentadas.





25

O próximo passo é calcular as frequências hipotéticas a partir dos percentuais resultantes nas posições C24, C25 e C26, multiplicando-os pelas frequências totais encontradas nas posições F7, F9 e F11. Assim:

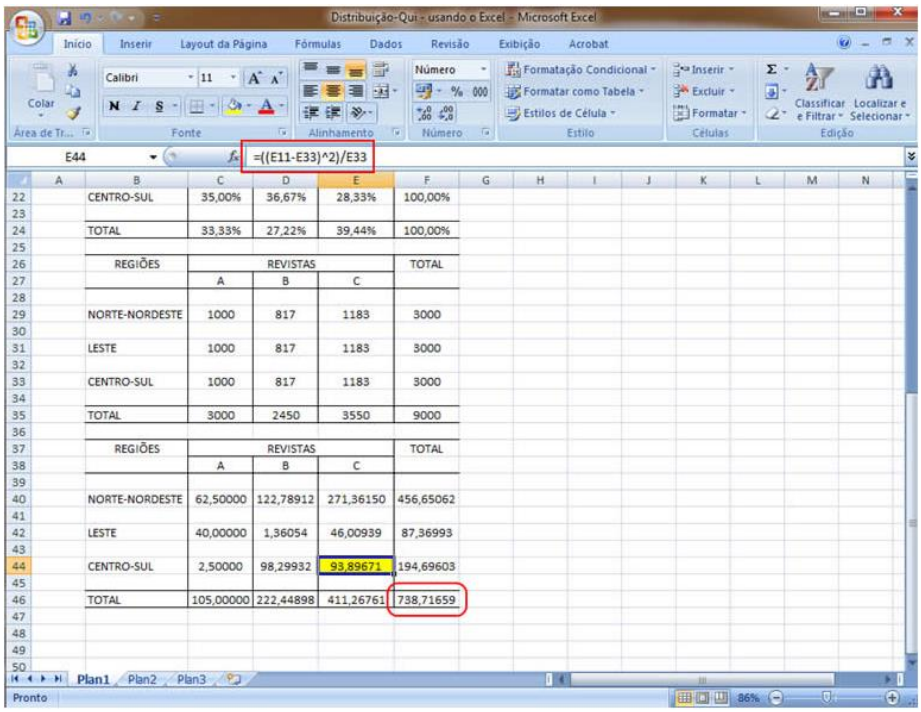


26

Deve-se agora construir uma tabela na qual apareçam as parcelas que totalizarão o qui-quadrado calculado. Para o cruzamento revista A e região Norte-Nordeste, lembrando que a medida geral é dada por

$$\sum_i \sum_j \left[\frac{(frequencia\ observada_{ij} - frequencia\ hipotética_{ij})^2}{frequencia\ hipotética_{ij}} \right]$$

faz-se: =((C7-C29)^2)/C29. Para os demais cruzamentos o procedimento é o mesmo.



22	CENTRO-SUL	35,00%	36,67%	28,33%	100,00%
23					
24	TOTAL	33,33%	27,22%	39,44%	100,00%
25					
26	REGIÕES	REVISTAS			TOTAL
27		A	B	C	
28					
29	NORTE-NORDESTE	1000	817	1183	3000
30					
31	LESTE	1000	817	1183	3000
32					
33	CENTRO-SUL	1000	817	1183	3000
34					
35	TOTAL	3000	2450	3550	9000
36					
37	REGIÕES	REVISTAS			TOTAL
38		A	B	C	
39					
40	NORTE-NORDESTE	62,50000	122,78912	271,36150	456,65062
41					
42	LESTE	40,00000	1,36054	46,00939	87,36993
43					
44	CENTRO-SUL	2,50000	98,29932	93,89671	194,69603
45					
46	TOTAL	105,00000	222,44898	411,26761	738,71659
47					
48					
49					
50					

27

A pequena diferença entre o valor do qui-quadrado calculado com a utilização do Excel e aquele apresentado anteriormente (738,75), deve-se, unicamente, ao fato de que agora não foram feitas aproximações numéricas para realização dos cálculos.

Uma vez determinado o qui-quadrado calculado, a sequência deve ser exatamente aquela já discutida:

1. Utilização do recurso da planilha para obtenção do percentual à direita desse valor.

2. Decisão pela dependência ou independência.

3. Concluindo-se pela dependência, análise dos cruzamentos que mais contribuíram para rejeição da independência.

28

RESUMO

Vimos que a análise bidimensional estuda o comportamento conjunto de duas variáveis, tendo como objetivo concluir se há independência ou dependência entre elas (relacionamento ou não). No módulo 1 apresentamos um problema que consistia em verificar se a hipótese variável A (Quantidade de revistas) varia em função da variável B (região do país). Para responder de forma conclusiva a essa questão, calculamos quais seriam as quantidades vendidas, caso essa hipótese não fosse verdadeira (as chamadas frequências esperadas). Por fim, calculamos um valor que representa a diferença entre as frequências reais e as frequências esperadas. Mas como interpretar esse valor? Como ele pode nos dizer se as duas variáveis são ou não independentes?

Para responder a essas perguntas, estudamos neste módulo a distribuição qui-quadrado. A distribuição qui-quadrado representa o valor da dispersão para duas variáveis. Tomando por base o valor obtido e o número de graus de liberdade, usa-se a tabela da distribuição qui-quadrado ou a planilha Excel. Quanto maior o valor de qui-quadrado, maior será a dependência entre as duas variáveis. De maneira geral, a literatura rejeita a hipótese de independência quando a área à direita do valor seja inferior ou igual a 0,5%. Chamamos esse valor limite também de grau de significância.

Vimos ainda que, caso haja um relacionamento entre as variáveis, podemos investigar quais os cruzamentos que mais contribuíram para que isto acontecesse. Para tanto, tomam-se as maiores parcelas do qui-quadrado calculado (em número igual ao de graus de liberdade), o que constituirá um bom indicativo do comportamento das variáveis em questão. Por fim, vimos que também é possível usar o coeficiente de contingência de Pearson para medir a dependência entre as duas variáveis.

UNIDADE 3 – ANÁLISE BIDIMENSIONAL E ANÁLISE DE REGRESSÃO
MÓDULO 3 – ANÁLISE DE REGRESSÃO

29

1 - CARACTERIZANDO A ANÁLISE DE REGRESSÃO

Muitas vezes procuramos identificar a existência de uma relação entre duas ou mais variáveis. Às vezes desejamos verificar, por exemplo, se o nível salarial das pessoas está relacionado com o tempo de experiência profissional da mesma. Pode-se ainda querer saber qual seria o valor de um apartamento de seis quartos em determinado local, onde só se tem a venda apartamentos de 3, 4 e 5 quartos. Por outro lado, pode-se ainda querer saber qual será a estimativa de consumo de energia elétrica de um determinado local ao longo do tempo.

Com o auxílio de uma **análise de regressão** poderemos conhecer esses valores.

A análise de regressão nada mais é do que a estimação de uma equação matemática, através de uma função pré-definida, que servirá para explicar a relação entre os dados pesquisados.

Contudo a correlação (verificação da existência e do grau de relação entre as variáveis) entre os dados deve ser verificada. Em função desse índice, que é conhecido como "coeficiente de correlação", poderemos chegar ao coeficiente de determinação, que servirá então para validar equação de regressão encontrada para os dados.

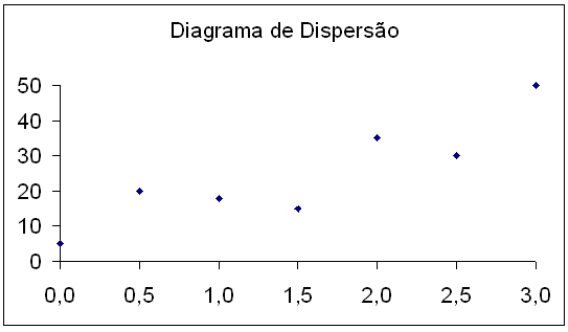
A equação alcançada pela análise de regressão será do tipo Y = f(Xi), ou seja, a variável Y será a variável dependente ou explicada pela (s) variável (is) Xi, que será (ão) a (s) variável (is) explicativa (s) ou independente (s). Se a variável dependente estiver em função de somente uma variável independente, diz-se que o modelo (equação) encontrado é **simples**. Caso a variável dependente esteja em função de mais de uma variável independente diz-se que o modelo é **composto**.

30

A função para determinação da variável dependente da relação entre as variáveis. As formas mais usuais são as funções: linear, potência, exponencial, hiperbólica, polinomial e logarítmica.

De uma maneira geral, até pela facilidade das operações matemáticas, a função linear é a mais utilizada. A análise simples, com apenas uma variável independente também é a mais utilizada. Dessa maneira, passaremos então, a detalhar as funções de regressão lineares simples.

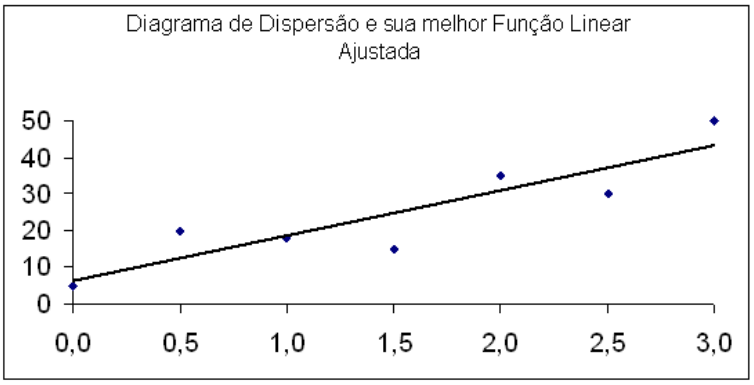
Uma equação de regressão linear simples pode ser escrita da forma genérica Y = aX + b, onde X será a variável independente; Y será a variável dependente, que será calculada em função do valor de X; "a" e "b" serão os parâmetros da função.



Analisando a "nuvem" de pontos assinalados, teremos melhores condições de especificar a função que relaciona as variáveis. No caso em análise, o ajustamento pelo modelo linear se dará em termos de uma reta. Contudo nos faltará ainda saber por onde passará a nossa reta. Assim teremos então que calcular os valores dos parâmetros "a" e "b".

31

A fim de se otimizar a equação de regressão, esses parâmetros deverão ter valores que aproximem, ao máximo, a reta dos pontos assinalados no diagrama de dispersão, conforme demonstrado na figura a seguir:



Dessa forma, o melhor método para a determinação dos parâmetros “a” e “b” que minimize as discrepâncias entre a reta e os pontos dos pares ordenados dos dados é o Método dos Mínimos Quadrados Ordinários. Segundo esse método, poderemos avaliar os parâmetros “a” e “b” pela aplicação das seguintes fórmulas:

$a = {Sxy}/{Sxx}$, onde $Sxx = \sum X_i^2 - n\overline{X}^2$ e $Sxy = \sum X_i Y_i - n\overline{X}\,\overline{Y}$

e ainda: $b = \overline{Y} - a\,\overline{X}$

32

Para avaliar o grau de relação entre as variáveis, deveremos determinar o coeficiente de correlação entre das variáveis em estudo através da formulação matemática:

onde: $$r = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{n\sum X_i^2 - (\sum X_i)^2}\sqrt{n\sum Y_i^2 - (\sum Y_i)^2}}$$

“n” será o tamanho da amostra.

$\sum X_i Y_i$ será o somatório dos valores de “X” multiplicados pelos valores de “Y” um a um.

$\sum X_i$ será o somatório de todos os valores da variável “X”.

$\sum Y_i$ será o somatório de todos os valores da variável “Y”.

$\sum X_i^2$ será o somatório de cada um dos valores da variável “X” elevado ao quadrado.

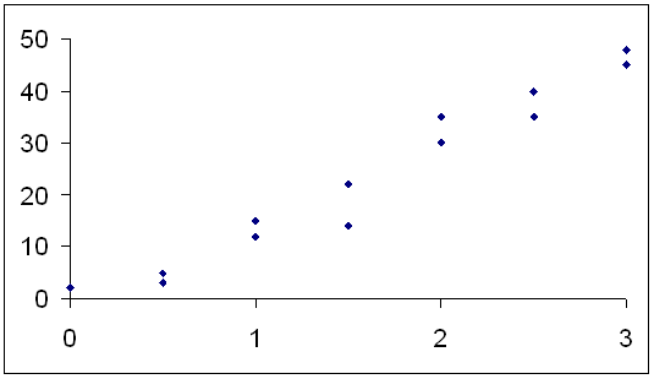
$\sum Y_i^2$ será o somatório de cada um dos valores da variável “Y” elevado ao quadrado.

O valor de “r” poderá variar de -1 a +1 passando, obviamente pelo zero. Assim, sua interpretação dependerá de seu valor numérico e de seu sinal.

33

Quando o valor de r estiver compreendido entre zero e 1, significa que temos uma **correlação positiva**, ou seja, para um incremento positivo da variável independente “X” teremos um incremento positivo da variável dependente “Y”.

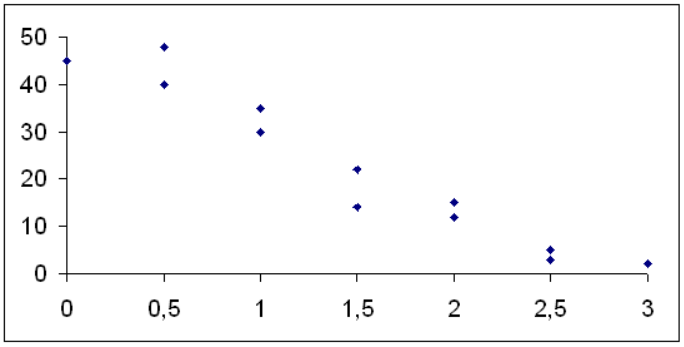
Assim, a representação no eixo cartesiano seria da seguinte forma:



34

Quando o valor de r estiver compreendido entre -1 e zero, significa que temos uma **correlação negativa**, ou seja, para um incremento positivo da variável independente “X” teremos um incremento negativo, ou ainda um decréscimo positivo da variável dependente “Y”.

Assim, a representação no eixo cartesiano seria da seguinte forma:



35

Por último devemos validar a equação da reta encontrada. Esse procedimento deve ser efetuado através do coeficiente de determinação, que nada mais é, em termos matemáticos, que o quadrado do valor do coeficiente de correlação:

$$r^2 = \left(\frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{n\sum X_i^2 - (\sum X_i)^2}\sqrt{n\sum Y_i^2 - (\sum Y_i)^2}} \right)^2$$

Teoricamente o coeficiente de determinação é o valor da variação da variável dependente Y, que é explicado pela reta de regressão. Para chegarmos a um valor de r^2 igual a 0,89, poderemos dizer que 89% da variação total da variável dependente Y está sendo explicada pela reta de regressão em função da variável independente X. Por outro lado 11% da variação de Y permanecem não explicados.

Segundo o Assis (2000) em função do valor do coeficiente de determinação, a mesma pode ser:

$r^2 > 0,90$ (90%)	►Muito Forte;
0,90 (90%) > $r^2 > 0,75$ (75%)	►Forte;
0,75 (75%) > $r^2 > 0,50$ (50%)	►Moderada;
$r^2 < 0, 50$ (50%)	►Fraca.

36

Passemos, agora, a um exemplo do emprego da análise de regressão, para resolver um questionamento do dia a dia:

Você pretende abrir um negócio, cujo foco recai sobre produtos alimentícios. Dentro de seu planejamento, você começa a prospectar possíveis pontos para a instalação de sua mercearia / minimercado / "sacolão". Surgindo um, que lhe chame particular atenção, parte-se para o levantamento de dados junto a um conjunto de 10 famílias (que servirão de base amostral) das circunvizinhanças para tentar ter uma visão preliminar dos gastos mensais que realizam com produtos alimentícios naquela região de seu interesse.

37

Os dados coletados revelam que:

Famílias	Despesas (em Reais)
A	800
B	850
C	900
D	980
E	990
F	1000
G	1050
H	1150
I	1170
J	1250

Você poderia fazer uma análise descritiva, iniciando pelo cálculo da despesa média, passando pelo cálculo da variabilidade, verificação da existência de pontos discrepantes, mas você está muito intrigado com possíveis explicações para a variação das despesas entre as diferentes famílias, daquela área da cidade. Surge logo uma potencial explicação: a renda das famílias. Como você montou um cadastro, no qual também constam informações sobre a renda familiar (admitindo que as famílias entrevistadas não fizeram nenhuma restrição para informar esse dado), vem a tabela:

Famílias	Despesas (em Reais)	Renda Familiar Líquida (em Reais)
A	800	2000
B	850	2000
C	900	2100
D	980	2250
E	990	2400
F	1000	2500
G	1050	2500
H	1150	2750
I	1170	2800
J	1250	3100

38

Busca-se então construir uma equação de regressão linear simples relacionando a variável de interesse dependente (despesas mensais com alimentação), designada por Y, e a variável que supostamente irá explicá-la (renda mensal líquida), designada por X. Assim, o que se busca é uma expressão do tipo:

$$Y = aX + b$$

Os procedimentos de cálculo, para determinar essa equação de regressão, são os seguintes:

a) Inicialmente identificamos o valor de n (número de pares de elementos amostrais).
n = 10.

b) Calculamos os valores de: $\sum X_i$, $\sum Y_i$, $\sum X_i Y_i$, $\sum X_i^2$, $\sum Y_i^2$, **X**, **Y** assim:

$$\sum X_i = 2000+2000+2100+2250+2400+2500+2500+2750+2800+3100 = 24.400$$

c) Na sequência, devemos calcular os valores dos parâmetros Sxx e Sxy:

$$Sxx = \sum X_i^2 - n\overline{X} = 60.745.000 - [10*(2440)2] = 60.745.000 - 59.536.000 = 1.209.000$$

$$Sxy = \sum X_i Y_i - n\overline{X}\,\overline{Y} = 25.209.500 - (10*2.440*1.014) = 467.900$$

d) Finalmente, poderemos calcular os valores dos parâmetros "a" e "b" de nossa equação de regressão:

$$a = {Sxy}/{Sxx} = 467.900 / 1.209.000 = 0,39$$

$$b = Y - a\,X = 1.014 - (0,39*2.440) = 62,4$$

Logo, nossa equação de regressão será: $\hat{Y} = 0,39\,X + 62,4$, o acento circunflexo sobre a variável dependente "Y" significa a estimação desse valor, ou seja, o par ordenado estabelecido, não necessariamente pertencerá ao conjunto de dados originais.

39

Com a equação de regressão estabelecida, passaremos, então, à sua validação estatística. Essa validação deve ser efetuada através do cálculo do coeficiente de determinação r^2. Para tanto calcularemos o coeficiente de correlação entre as variáveis "r" e elevaremos o valor de seu resultado ao quadrado. Assim:

$$r = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{n\sum X_i^2 - (\sum X_i)^2}\,\sqrt{n\sum Y_i^2 - (\sum Y)^2}}$$

$$r = \frac{(10*25.209.500)\ (24.400*10.140)}{\sqrt{(10*607.450.000) - (24.400)^2} * \sqrt{(10*10.469.400) - (10.140)^2}}$$

$$r = \frac{252.095.000 - 247.416.000}{\sqrt{607.450.000 - 595.360.000} * \sqrt{104.694.000 - 102.819.600}}$$

$$r = \frac{4.679.000}{\sqrt{12.090.000} * \sqrt{1.874.400}} \quad = \quad r = \frac{4.679.000}{3.477,07*1.369,09}$$

$$r = \frac{4.679.000}{4.760.409,23} \quad = \quad r = 0,983 \rightarrow \quad r^2 = 0,966$$

O resultado do coeficiente de determinação de 0,966, significa que a equação estabelecida entre a variável independente X (renda familiar líquida) e a variável dependente Y (despesas) explica 96,6% a relação que existe entre as mesmas.

Assim, podemos dizer que a determinação foi **muito forte** e, consequentemente, a equação estabelecida foi validada e poderá ser utilizada para fins de estimação.

40

Dessa forma, se quisermos estimar qual seria, por exemplo, a despesa de uma família com renda líquida mensal de R$ 5.000,00, bastaria entrar com o valor de 5.000 na equação:

$$\hat{Y} = 0{,}39X + 62{,}4 \rightarrow \hat{Y} = (0{,}39*5000) + 62{,}4 \rightarrow \hat{Y} = 2.012{,}40$$

Portanto, a despesa para uma família que possui uma renda familiar de R$ 5.000,00 (cinco mil reais) será de R$ 2.012,10 (dois mil e doze reais e dez centavos).

Por outro lado, qual seria, de acordo com a equação, a expectativa de despesa de uma família que não tivesse nenhum rendimento?

Bastaria substituir o rendimento nulo, ou seja, zero na equação, assim:

$$\hat{Y} = 0{,}39X + 62{,}4 \rightarrow \hat{Y} = (0{,}39*0) + 62{,}4 \rightarrow \hat{Y} = 62{,}40$$

A despesa seria de R$ 62,40 (sessenta e dois reais e quarenta centavos).

41

RESUMO

Na análise de regressão busca-se responder por que determinada variável está variando e como ela está variando.

Como uma primeira abordagem recorre-se a um modelo bastante simples, que é a função do primeiro grau, ou seja, Y = aX + b. A partir daí, é necessário estabelecer um critério para determinação dos parâmetros a e b. Esse método será o de mínimos quadrados, o que significa dizer que os valores obtidos serão aqueles que minimizarão a soma dos quadrados dos erros, entendendo como erros (ou resíduos) as diferenças entre os valores reais / observados de Y e aqueles valores estimados a partir da construção do modelo e substituição dos valores de X.

A reta assim obtida será a melhor reta possível, o que não é sinônimo de ser uma boa reta (ou uma reta suficientemente aderente à realidade). Assim, é necessário avaliar o quanto o modelo explica (ou justifica) a variabilidade original de Y. A medida construída para isto é o coeficiente de determinação, que revelará qual o percentual de explicação da variabilidade de Y deve-se à variabilidade de X (dentro do modelo determinado). O próximo passo é interpretar os valores dos parâmetros a e b, sendo que a será o valor de Y, quando X = 0 (o que não necessariamente tem um significado prático) e b será o

impacto sobre a variável Y quando X variar de uma unidade (no mesmo sentido se for um valor positivo e em sentidos contrários, caso seja negativo).

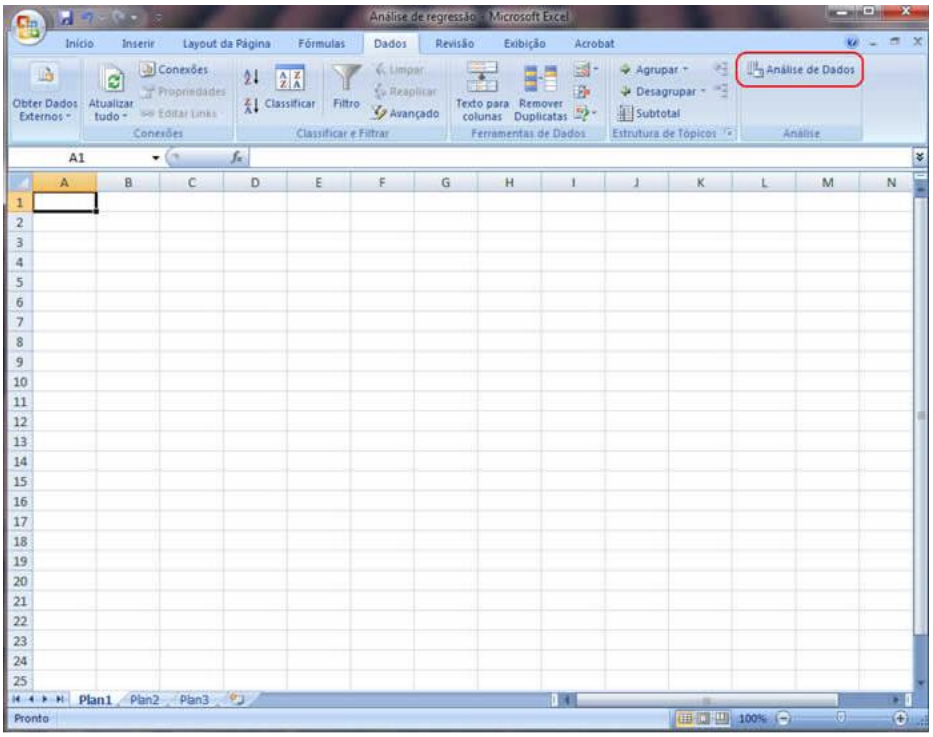
Caso o valor do coeficiente de determinação não seja julgado satisfatório, é conveniente "especular" as possíveis razões, que basicamente podem ser: insuficiência de dados, erro de especificação do modelo (pode não ser uma reta), erro de especificação da variável explicativa (pode ter sido escolhida uma variável explicativa inadequada ou insuficiente para, sozinha, explicar a variabilidade de Y) ou fatores subjetivos / de difícil mensuração.

UNIDADE 3 – ANÁLISE BIDIMENSIONAL E ANÁLISE DE REGRESSÃO
MÓDULO 1 – ANÁLISE DE REGRESSÃO COM AUXÍLIO DO EXCEL

42

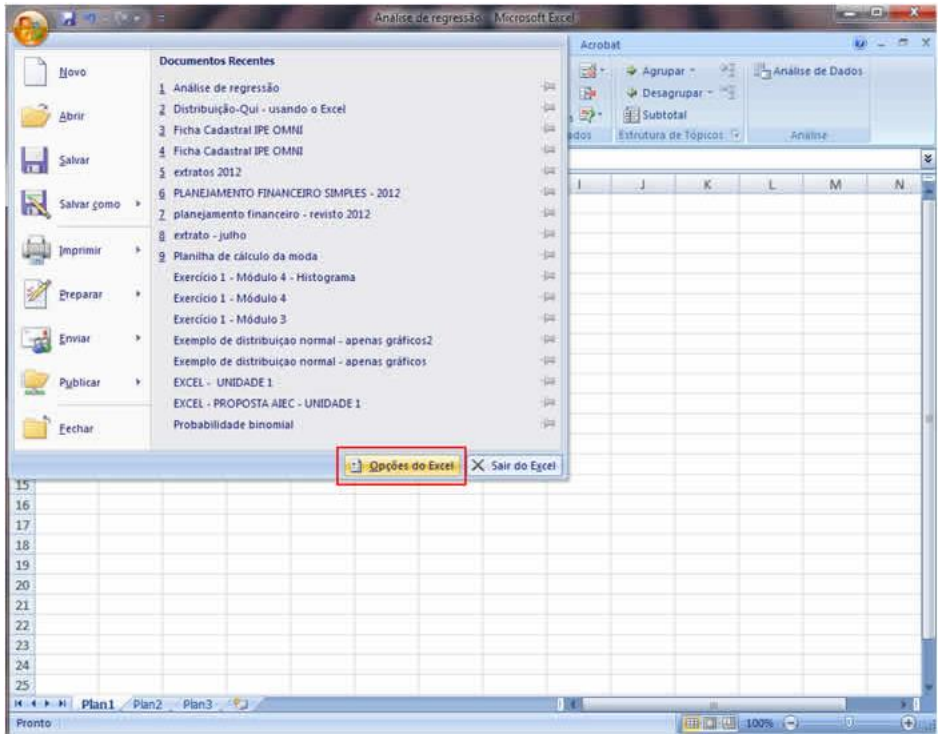
1 - ANÁLISE DE REGRESSÃO COM O AUXÍLIO DO EXCEL

Inicialmente devemos verificar se nosso programa Excel está habilitado a executar uma análise de regressão. Para isso devemos clicar, dentro do programa, na guia Dados e verificar se o botão Análise de Dados está disponível:



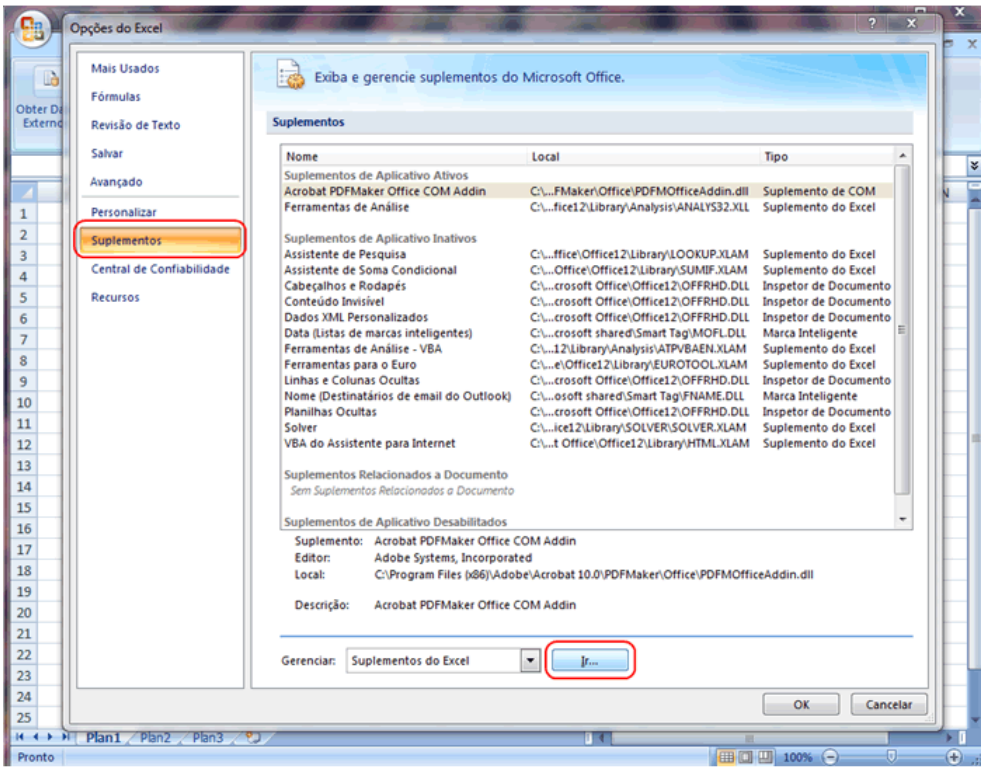
43

Caso não esteja disponível o botão “Análise de Dados” teremos que configurar o Excel incluindo o suplemento de Análise de Dados. Para habilitar tal função, que é essencial para se desenvolver uma análise de regressão, deveremos clicar no botão Office , em seguida no botão Opções do Excel:



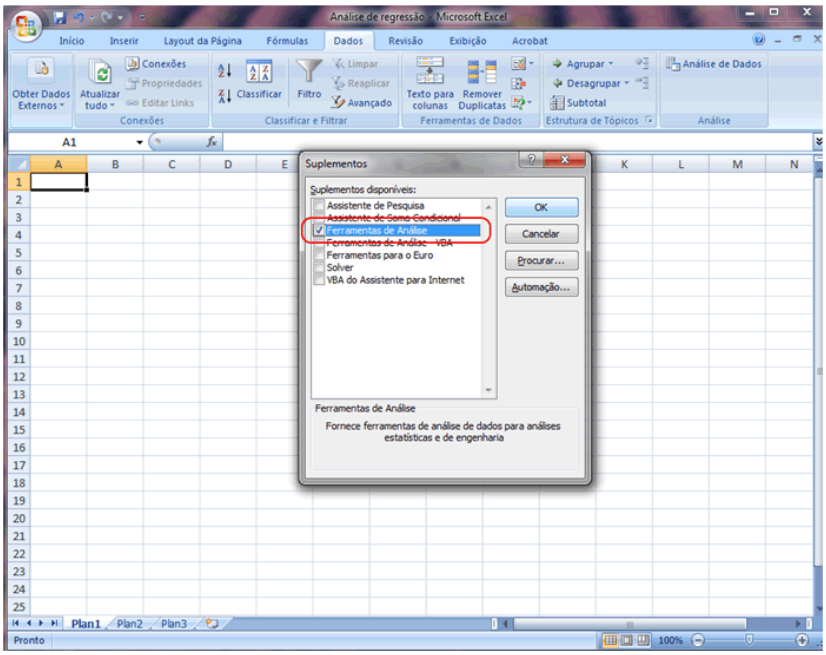
44

Na opção "Suplementos", clicar no Ir.



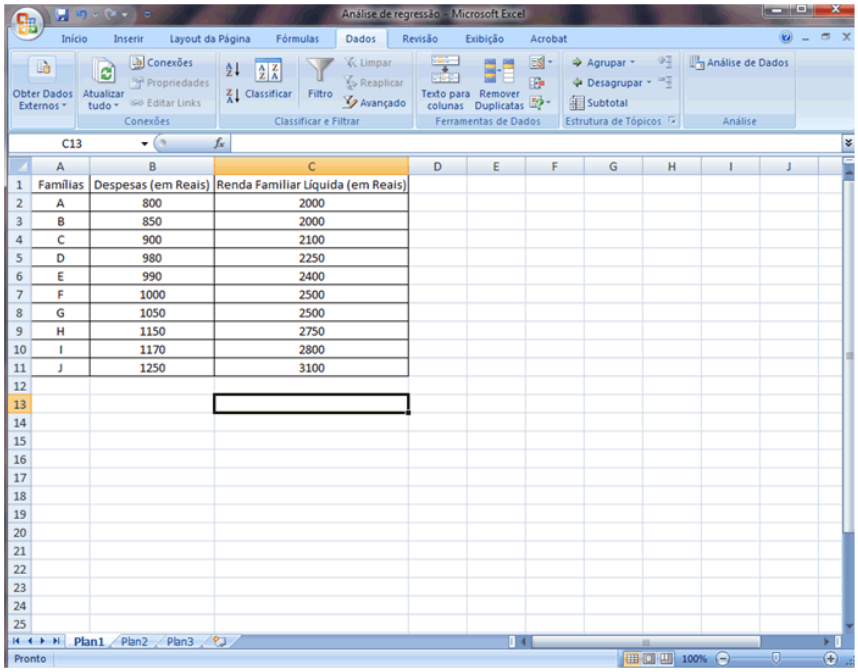
45

Em seguida, selecione a opção "Ferramentas de Análise" e confirme com o botão OK.



46

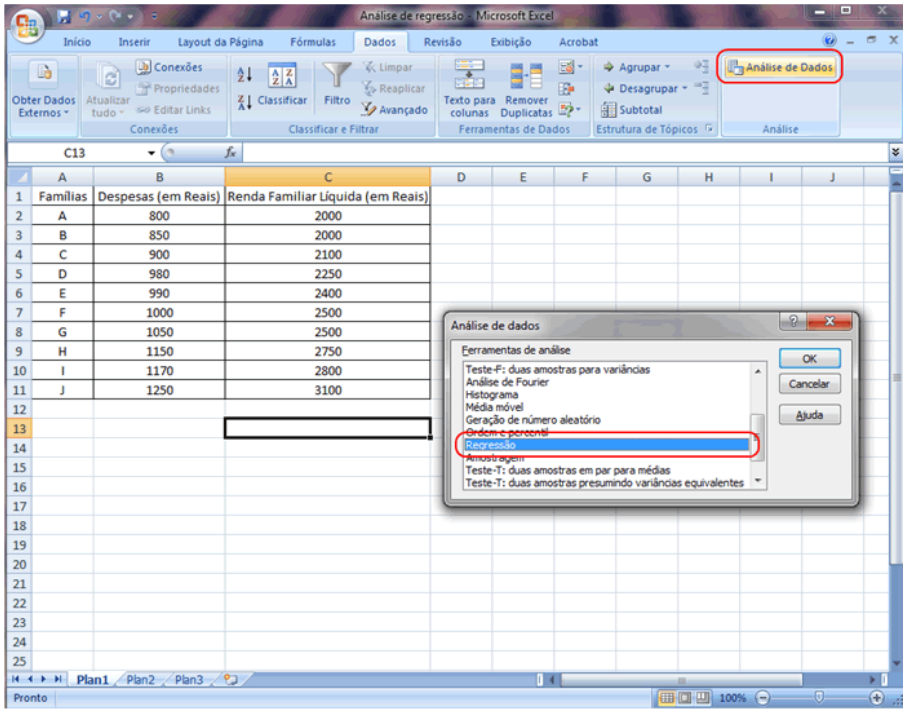
A partir daí poderemos desenvolver nossa análise de regressão. Para tanto deveremos digitar em nossa planilha os dados que iremos trabalhar. Para essa demonstração iremos adotar o mesmo exemplo já desenvolvido manualmente:



Famílias	Despesas (em Reais)	Renda Familiar Líquida (em Reais)
A	800	2000
B	850	2000
C	900	2100
D	980	2250
E	990	2400
F	1000	2500
G	1050	2500
H	1150	2750
I	1170	2800
J	1250	3100

47

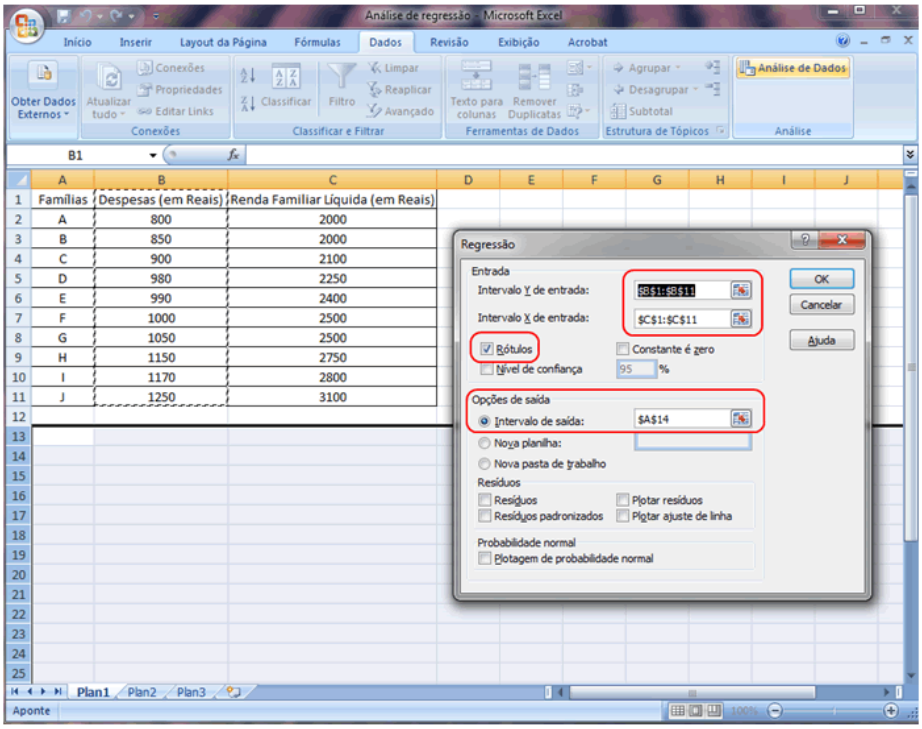
Para a determinação da equação e do coeficiente de determinação deveremos, após digitar os dados na planilha Excel, selecionar uma célula onde queremos que o resultado da análise apareça e depois, na aba Dados, clicar em “Análise de Dados”, selecionar a opção "Regressão" e clicar OK (conforme figura a seguir):



48

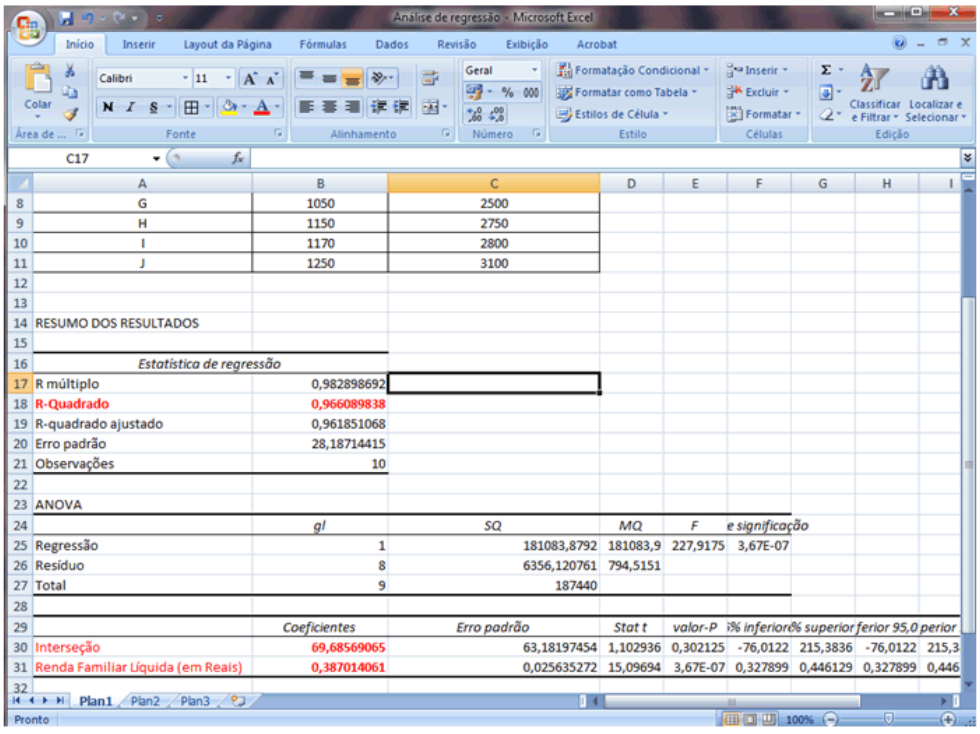
Deveremos clicar então nas pequenas setas vermelhas, que significam “entrada de dados” e selecionar os dados das variáveis independente “X” e dependente “Y”.

Deveremos, ainda, selecionar também através da pequena seta vermelha a célula que havíamos reservado anteriormente para a saída dos dados. Para que os dados já saiam na planilha final com os nomes das variáveis deveremos selecionar junto com os dados (números) as células com os próprios nomes das variáveis. Em contrapartida, para que o Excel entenda esse nosso procedimento deverá ativar o item “Rótulos”. Finalmente deveremos clicar em “ok” para que o Excel resolva as operações matemáticas de cálculo de nossa análise de regressão.



49

Essa função, ou seja, a análise de regressão com o auxílio do Excel irá fornecer vários resultados. Alguns destes não são objetos de nosso curso. Dessa maneira poderemos visualizar, em vermelho, os dados de nosso interesse nesse momento:



O coeficiente de determinação encontrado na análise de regressão com o auxílio do Excel “bateu” com o valor calculado manualmente de r2 = 0,966. Por outro lado os valos dos parâmetros “a” e “b” também “bateram” sendo iguais a “a” = 0,39 (já arredondado) e “b” = 69,69. É importante salientar que o valor desse parâmetro deu resultado um pouco diferenciado do resultado pelo cálculo manual (62,4) devido ao fato de que o Excel não arredonda os dados, o que culminou com essa pequena diferença. Se considerarmos no cálculo manual o valor do parâmetro “a” encontrado com três casas decimais, ou seja, 0,387, chegaríamos a um valor do parâmetro “b” igual a 69,72.

50

RESUMO

Vimos anteriormente que a regressão linear é uma ferramenta importante para a análise de dados nos permitindo entender o relacionamento entre duas variáveis e fazer estimações.

Neste módulo apresentamos a realização da Análise de regressão utilizando a ferramenta "Análise de Dados" do Microsoft Excel. Para usar essa ferramenta precisamos habilitar o suplemento "Ferramentas de Análise". Uma vez habilitado temos acesso a diversas ferramentas de análise, dentre as quais a ferramenta "Regressão" que permite calcular com precisão os coeficientes a e b da regressão linear e dentre diversos outros parâmetros o coeficiente de determinação (R-Quadrado) que nos permite avaliar a aderência dos dados a esse modelo.