Capítulo 14

Programação linear, Análise de dados

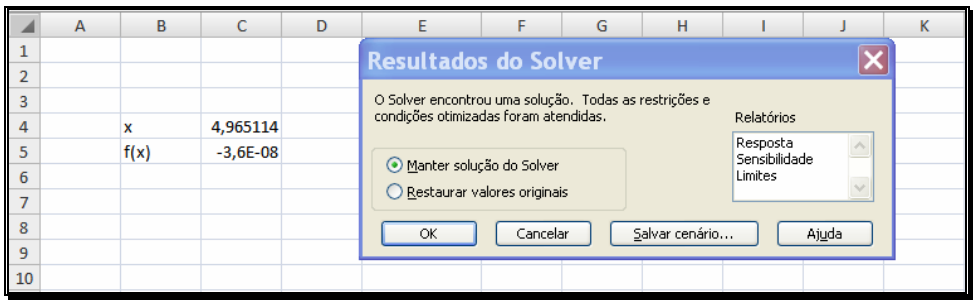
Trabalhando com o SOLVER

O Excel oferece mais ferramentas estadísticas. Via *Dados* encontra você **Analise de Dados** e o Add-in **Solver**. Se não encontrar, deve carregá-los.

- Clique no botão do Microsoft Office e, em seguida, clique em *Opções do Excel*.
- Clique em *Suplementos* e, na caixa *Gerenciar*, selecione *Suplementos do Excel*. Clique em *Ir para.*
- Na caixa *Suplementos disponíveis*, marque as caixas *Ferramentas de Análise, Ferramentas de Análise – VBA* e *Solver*. OK

Com o *Analise de Dados* vamos trabalhar mais à frente. Neste momento, dedicamo-nos ao **Solver** com o qual podemos, entre outros, resolver problemas que são complicados demais para a ferramenta *Atingir meta* (Goal Seek) que utilizamos no começo do capítulo 8. Para familiarizar-nos com o Solver, vamos resolver outra vez o problema anterior.

O Solver encontra o mesmo resultado que encontramos usando *Atingir Meta*, possivelmente com mais precisão:



Outros problemas para o Solver resolver lidam com **programação linear** (PL). Ela é usada para maximizar ou minimizar diversos tipos de problemas, por exemplo problemas da **ótima mistura** de produtos. Como exemplo podemos citar as distribuidoras de petróleo que precisam determinar a quantidade de aditivos a ser adicionada ao petróleo de forma a obter um certo tipo de gasolina ao menor custo possível ou, em certos casos, quere-se conhecer a quantidade de água que se pode adicionar a fim de atender às expectativas mínimas dos clientes -como poder ligar o motor ou poder dirigir pelo menos um quilômetro sem problemas sérias.

Assim, temos o problema de buscar um valor extremo de uma grandeza que depende de várias variáveis. Esta busca depende, muitas vezes, de *restrições* laterais que, em geral, podem ser formuladas em forma de igualdades ou desigualdades. Geralmente, trata-se de uma otimização *linear* onde se busca minimizar ou maximizar o valor de uma *função objetivo* linear $z(x_1,...,x_n) = a_1x_1 + ... + a_nx_n$. Neste caso, também as restrições são equações ou inequações lineares, ou seja, as equações ou inequações dos modelos de programação linear (PL) têm a seguinte conotação:

$$a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n <=/>= b_1 \text{ etc.}$$

O método implementado no Solver é chamado de *método simplex* que é um algoritmo que se aproxima iterativamente à solução ótima. (O Solver foi desenvolvido pela FrontLine Systems, mas, existem no mercado e no domínio publico outros Solvers, por exemplo o LP_solve.)

Para conhecer o Solver, definimos um problema no qual se produz um produto pela mistura das sustâncias S_1, ... ,S_n. A sustância S_i contém as sustâncias básicas B_1, ... ,B_m.

Suponhamos que um fabricante de comidas para animais de estimação pretenda fabricar um produto novo pela mistura de S_1 e S_2 (cereais e carne) que contém pelo menos 150g de gorduras, 200g de proteínas, 250g carboidratos e com um calor de combustão de 6800KJ, -e que deve ser, obviamente, o menos custoso possível. (Em comidas para animais de estimação, boas fontes de minerais incluem suplementos minerais, peixe, carne, fígado, lácteos e cereais.)

Tabela dos materiais básicas em gramas por kg de cereais/carne

	Cereais	Carne	Mínimo
Gorduras	100g	500g	150g
Proteínas	500g	100g	200g
Carboidratos	400g	400g	250g
Combustão	8400kJ	17000kJ	6800kJ
Preço/kg	3,50	5,20	

Sejam x = quantidade em kg de cereais por ração e y = quantidade em kg de carne por ração.

Para as restrições temos

Gorduras: 100x + 500y >= 150
Proteínas: 500x + 100y >= 200
Carboidratos: 400x + 400y >= 250
Combustão: 8400x + 1700y >= 6800

A função objetivo a minimizar é: z = 3,5x + 5,2y

Entradas na planilha:

Coloque os dados numa planilha, veja o exemplo a seguir.

1. O Solver precisa duas células, por exemplo F1 e F2 (células variáveis), para armazenar as duas soluções x e y.
2. F4 contém a *função objetivo* =F1*3,5+F2*5,2
3. As *condições laterais* colocamos em H1:H4
   H1: =F$1*B1+F$2*C1, copiar até H4.
4. Active o Solver.
5. *Definir célula de destino*: fazer clique na célula F4 e, depois, selecionar *Min*. As *Células variáveis* são F1:F2, faça um clique nelas. Clique no botão *Adicionar* para adicionar as restrições. A caixa de dialogo está dividida em três partes. Com o cursor no campo *Referência de célula*, faça um clique em H1; mude o símbolo <= para >= (para cada desigualdade) e, com o cursor em *Restrição*, faça um clique na D1. Em seguida, clique no botão *Adicionar* para colocar a desigualdade na lista das restrições. Agora o mesmo procedimento com H2 e D2 etc. Você deve terminar a última restrição com *OK*.

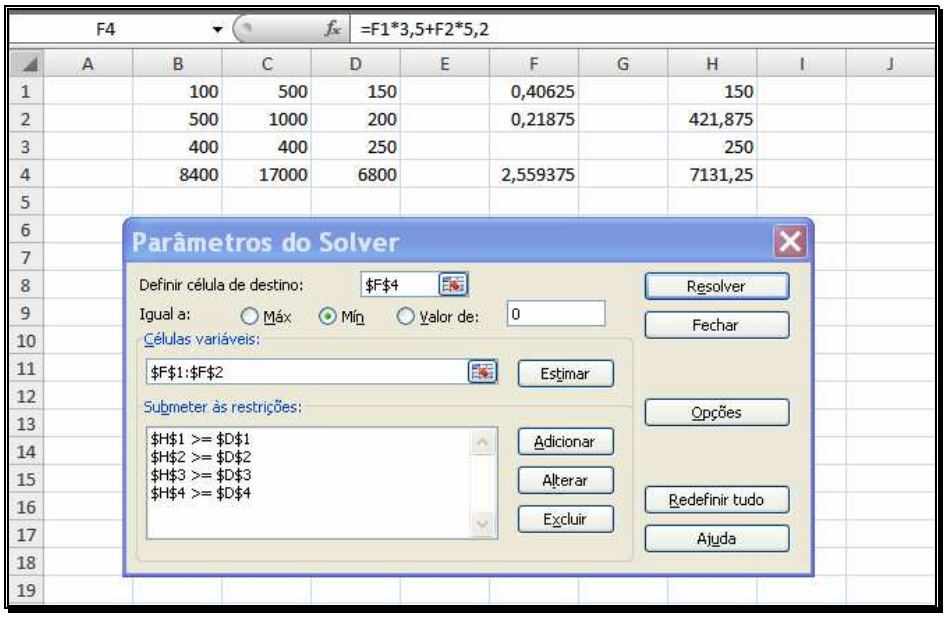
Se depois clicar em *Resolver*, aparecerá em F1 a informação de que deve usar, por ração, 406 gramas de cereais. Em F2 diz y = 219g de carne. Na célula F4 fica o preço da ração: 2,56 Reais.

Após terminar, veremos a caixa de dialogo do Solver:

Manter solução do Solver: Neste caso, vai manter na planilha atual os valores encontrados pelo Solver.

Restaurar valores originais: Neste caso, vai manter os valores originais.

Existem mais Opções: Tempo máximo, Precisão, Mostrar resultado de iteração ....



Análise de dados

Para descrever uma amostra, utiliza-se as seguintes estimativas:

Freqüência, média amostral, desvio padrão amostral, mediana amostral. Estas estimativas estimam a verdadeira média, o desvio padrão e a mediana da população, que são desconhecidos. Chama-se os verdadeiros, mas dês-conhecidos, valores populacionais de *parâmetros*, definidos com letras Gregas. As letras Romanas referem-se aos valores amostrais que são chamadas de *estadísticas*. A pergunta básica a responder é: Como podemos obter estimativas dos parâmetros populacionais, a partir das estadísticas amostrais, e quão precisas serão tais estimativas?

Para o seguinte exemplo precisamos das seguintes expressões:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{m} f_i x_i$$ média amostral; =MÉDIA

$$s^2 = \frac{1}{n-1}(\sum_{i=1}^{m} f_i x_i^2 - n\overline{x}^2)$$ variância amostral; =VAR

$s = \sqrt{s^2}$ = desvio padrão amostral: =DESVAP

f_i = freqüência absoluta =FREQÜÊNCIA, $F_i := f_i/n$ = freqüência relativa

Existe outra definição da variância com 1/n em vez de 1/(n-1). A diferença entre as duas fórmulas será insignificante, se n fosse muito grande. A formula com 1/n pode ser escrita como $s^2 = \sum_{i=1}^{m} x_i^2 F_i - \overline{x}^2$. Esta é, geralmente, uma expressão mais conveniente para usar no cálculo da variância de uma distribuição de freqüência do que a anterior.

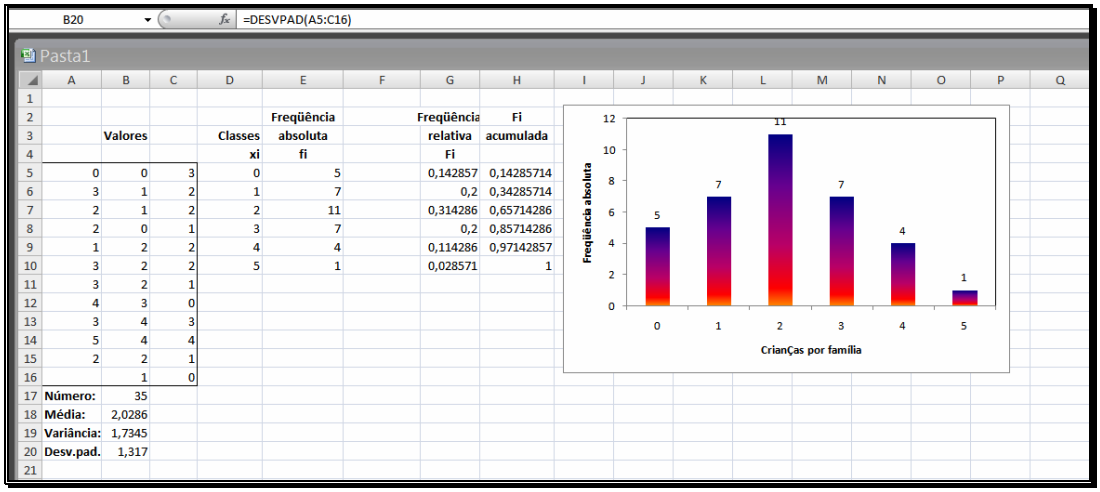
Exemplo: Temos uma amostra de 35 valores (crianças por família) que foram anotados no momento de recebê-los, sem ser ordenados. Queremos determinar os valores das estadísticas.

Na planilha a seguir, temos em A5:C16 os valores da amostra. Ao lado, D5:D10, temos uma pequena lista das classes (0, 1, 2, ...,5).

E5: Selecionar E5:E10 e inserir a fórmula =FREQÜÊNCIA(A5:C16;D5:D10), Ctrl +Shift+Enter
B17: =SOMA(E5:E10)
B18: =MÉDIA(A5:C16)
B19: =VAR(A5:C16)
B20: =DESVAP(A5:C16)

G5: =E5/B$17, copiar até G10
Na coluna H calculamos a função F da distribuição empírica,Fi, acumulada.
H5: =G5; H6: =G6+H5, copiar até H10. O último valor da 1.

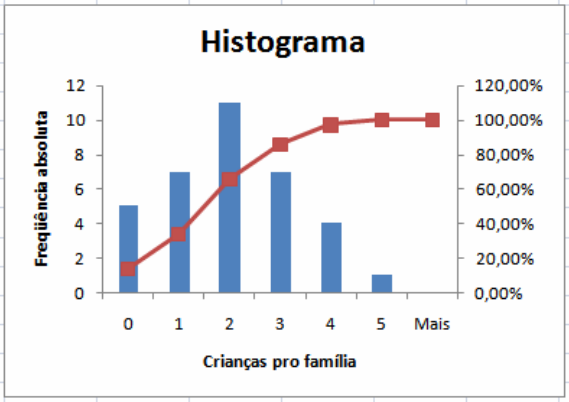
Para criar o gráfico, selecionamos *Barras>Barras agrupadas* e *Adicionar Rótulos*



A seguir utilizamos para o mesmo problema a ferramenta **Análise de Dados** que mostra os mesmos valores para a função F, mas em %.

Selecione *Histograma* com Intervalo de entrada: A5:C16. Intervalo do bloco (=bin range, bin = intervalo): D5:D10. Intervalo de saída: $E$14 (ou outro)

Bloco	Freqüência	cumulativo
0	5	14,29%
1	7	34,29%
2	11	65,71%
3	7	85,71%
4	4	97,14%
5	1	100,00%
Mais	0	100,00%



Na caixa de dialogo, foi selecionado *Percentagem Cumulativo* e *Resultado do gráfico*. (Um diagrama *Pareto* é um diagrama ordenado de barras.) O diagrama mostra também a curva da função F que termina em 100% = 1. O que é chamado de "bloco" é o intervalo de classe que, em inglês, é chamado "bin".

Distribuições

Neste parágrafo, estudamos distribuições contínuas. Em muitos casos práticos, podemos supor que os dados têm uma *distribuição Normal*.

A distribuição Normal ocupa um lugar de preeminência dentre as distribuições da teoria estadística. Ela é especificada por 2 parâmetros: a média populacional, **μ**, e o desvio padrão populacional, **σ**, ou também a variância $\boldsymbol{\sigma^2}$.

A função *gaussiana* de *densidade* de probabilidades, FGDP, é definida por

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \quad (1)$$

Esta função também é chamada *função normal de erros*. (No caso de a variável X sendo discreta, f(x) também é chamada *função* de probabilidades. A variável aleatória X é dita *discreta*, se assume valores num conjunto finito ou infinito enumerável.) A distribuição normal é simétrica em torno da média o que implica que e média, a mediana e a moda são todas coincidentes.
A Probabilidade do evento "X<=x", ou seja P(X<=x) = F(x), será calculada pela função

$$F(x) = \int_{-\infty}^{x} f(t)dt \quad = P(X<=x) \quad (2)$$

F(x) = função distribuição de probabilidade, ou função de distribuição cumulativa (FDA).

É convenção usar um F maiúsculo para a FDA, em contraste com o f minúsculo usado para a função densidade de probabilidade (ou função massa de probabilidade).

Usando μ=0 e σ=1, proporciona a *distribuição normal padrão*. Neste caso, escreve-se, normalmente, φ e Φ em vez de f e F.
Na prática desejamos calcular probabilidades para diferentes valores de μ e σ, (usando =DIST.NORM). Mas, não é necessário trabalhar com diferentes distribuições, para resolver um dado problema, basta transformar a variável X numa forma padronizada $Z = \frac{X - \mu}{\sigma}$, pois Z tem distribuição N(0,1). Podemos, então, escrever $F(x) = \Phi(\frac{x - \mu}{\sigma})$. Em Excel temos a função =DIST.NORMP que retorna a função da distribuição cumulativa normal padrão.

No seguinte exemplo vamos usar a função DIST.NORM

Exemplo: Suponha que as espessuras de um particular tipo de pranchas possam ser descritas por uma distribuição Normal, com média μ = 1,4cm e desvio padrão σ = 0,05cm. (Diremos, então, que a variável aleatória X = *espessura* varia continuamente, e teremos uma distribuição contínua. Tomamos a média aritmética $\overline{x}$ e o desvio *s* como "boas" estimativas de μ e σ.)

Aleatoriamente tiramos da produção uma prancha e perguntamos:

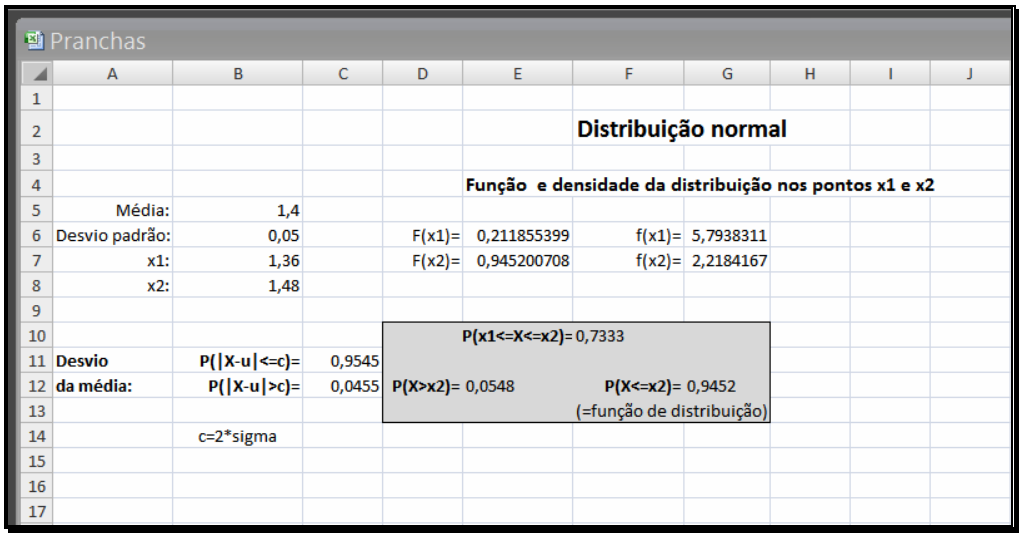
a. Qual a probabilidade de que a espessura esteja entre 1,36cm e 1,48cm?
b. Qual a probabilidade de que ela seja major do que 1,45cm?

Ajuda:

a. Dada f(x), eq.(1), a probabilidade de X se encontrar no intervalo (x_1,x_2) pode ser calculada através de integração segundo eq. (2).

$$P(x_1<=X<=x_2) = F(x_2) - F(x_1).$$

b. F(x) calculamos com =DIST.NORM(x;μ;σ;1). Com o parâmetro 0 obtém-se f(x). O lado direito da eq. (2) representa a probabilidade de que a variável X tome um valor inferior ou igual a x.



Pranchas

	A	B	C	D	E	F	G	H	I	J
1										
2						**Distribuição normal**				
3										
4					**Função e densidade da distribuição nos pontos x1 e x2**					
5	Média:	1,4								
6	Desvio padrão:	0,05		F(x1)=	0,211855399	f(x1)=	5,7938311			
7	x1:	1,36		F(x2)=	0,945200708	f(x2)=	2,2184167			
8	x2:	1,48								
9										
10				**P(x1<=X<=x2)=** 0,7333						
11	**Desvio**	**P(|X-u|<=c)=**	0,9545							
12	**da média:**	**P(|X-u|>c)=**	0,0455	**P(X>x2)=** 0,0548		**P(X<=x2)=** 0,9452				
13						(=função de distribuição)				
14		c=2*sigma								
15										
16										
17										

Entradas:

1. Dados em B5:B8
2. E6: =DIST.NORM($B$7;$B$5;$B$6;1) (= $F(x_1)$)
   E7: =DIST.NORM($B$8;$B$5;$B$6;1) (= $F(x_2)$)
3. G6: =DIST.NORM($B$7;$B$5;$B$6;0) (= $f(x_1)$ segundo eq.(1))
   G7: =DIST.NORM($B$8;$B$5;$B$6;0) (= $f(x_1)$)
4. F10: =SE(B7="";"";E7-E6)
5. C11: =2*DIST.NORM(2;0;1;1)-1 ou =2*DIST.NORMP(2)
6. C12: =2*(1-DIST.NORM(2;0;1;1)) ou =2*(1-DIST.NORMP(2))
7. E12: =1-E7; G12: =E7

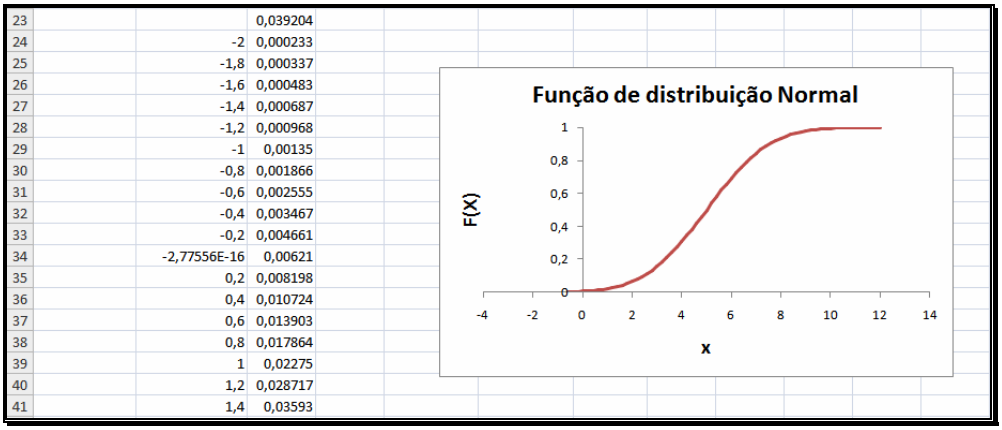
A probabilidade do desvio padrão da média foi calculada com c = 2σ. Para a distribuição Normal, a proporção de valores caindo dentro de dois desvios padrão da média, $\mu \pm 2\sigma$, é P(|X-u|<=2σ) = 2·Φ(2) -1 = 0,9545, ou ≈ 95,5%. Ou seja, veja C11, 95,5% de todas as pranchas têm uma espessura que desvia-se do valor esperado menos de c = 2σ = 0,1cm.

C12: só 4,55% desviam-se mais de 0,1cm da média.

(A desigualdade de Tschebyschew, P(|X-u|<=kσ) > 1-1/k^2, dá, com k=2, a probabilidade P > 0,75. Esta redução do limite a só 75% é o preço que se paga para a universalidade da estimação.)

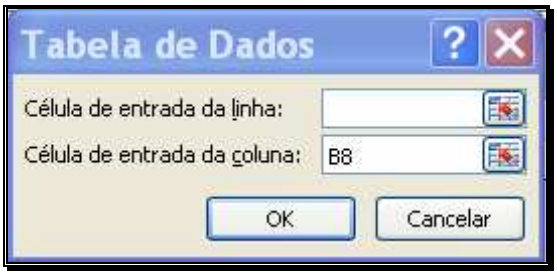
Se queremos trabalhar com $Z = \frac{X - \mu}{\sigma}$, devemos pôr μ = 0 e σ=1. $Z(x_1)$ = (1,36-1,4)/0,05 = -0,8 é $Z(x_2)$ = 1,6.

No Excel encontramos em *Dados>Teste de Hipóteses* a opção "Tabela de Dados". Por meio dela podemos substituir o valor na célula B8 sucessivamente por outros valores, por exemplo pelos valores -2, -1, 8, ... , 12 em B24:B94, veja a figura a seguir.



23		0,039204
24	-2	0,000233
25	-1,8	0,000337
26	-1,6	0,000483
27	-1,4	0,000687
28	-1,2	0,000968
29	-1	0,00135
30	-0,8	0,001866
31	-0,6	0,002555
32	-0,4	0,003467
33	-0,2	0,004661
34	-2,77556E-16	0,00621
35	0,2	0,008198
36	0,4	0,010724
37	0,6	0,013903
38	0,8	0,017864
39	1	0,02275
40	1,2	0,028717
41	1,4	0,03593

B5: 5; B6: 2; C23: =E7
B24: -2; B25: =B24+0,2 copiar até B94 (F5 e Ctrl d)
Selecione B23:C94 e em seguida selecione "Tabela de Dados" onde deixamos a primeira opção no primeiro campo em branco:

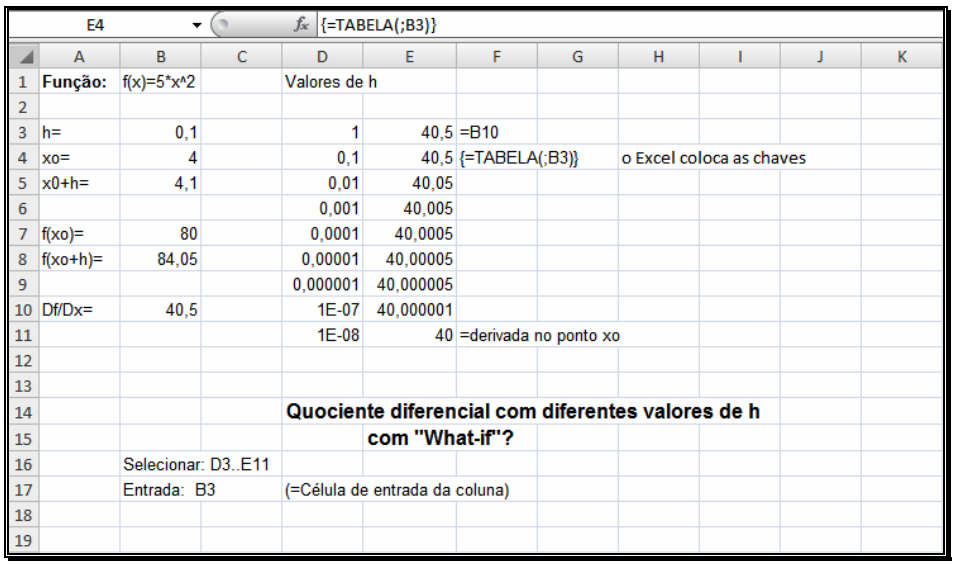


O valor na célula B8 (1,48) será então substituído pelos valores do intervalo B24:B94. Excel coloca em todas as células de C24 até C94 a fórmula matricial {=TABELA(;B8)}. O gráfico foi construído com os valores nas colunas B24:C94.

Foi isso um exemplo de um analise "what-if": o que passaria, se a espessura não for 1,4 mas ...?

(Com o mesmo método podemos demonstrar que o quociente de diferencias se aproxima ao valor limite, ou seja à derivada da função dada.

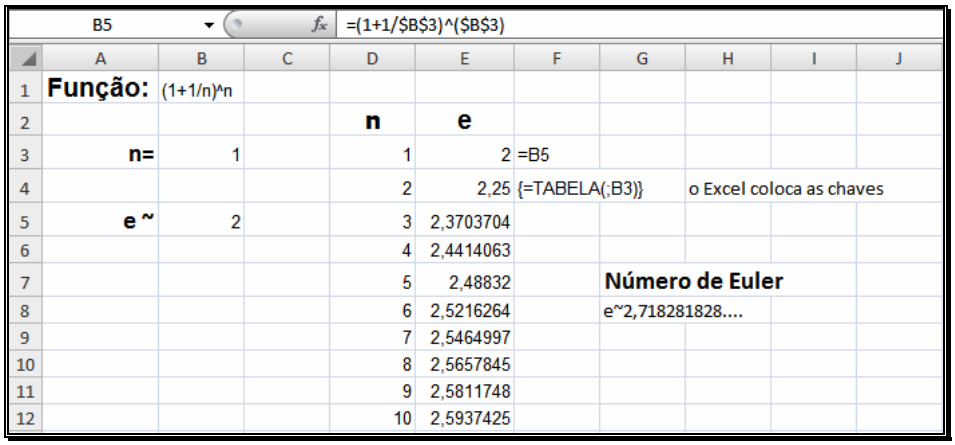
Veja a seguinte planilha na qual determinamos os valores do quociente diferencial da função $f(x)=5x^2$ para valores de h cada vez menores. D4: =D3/10 até E11. Excel coloca sucessivamente todos os valores de h na célula B3 e copia o conteúdo de B10 para E3:E11.



E4 {=TABELA(;B3)}

	A	B	C	D	E	F	G	H	I	J	K
1	**Função:**	f(x)=5*x^2		Valores de h							
2											
3	h=	0,1		1	40,5	=B10					
4	xo=	4		0,1	40,5	{=TABELA(;B3)}		o Excel coloca as chaves			
5	x0+h=	4,1		0,01	40,05						
6				0,001	40,005						
7	f(xo)=	80		0,0001	40,0005						
8	f(xo+h)=	84,05		0,00001	40,00005						
9				0,000001	40,000005						
10	Df/Dx=	40,5		1E-07	40,000001						
11				1E-08	40	=derivada no ponto xo					
12											
13											
14				**Quociente diferencial com diferentes valores de h**							
15					**com "What-if"?**						
16		Selecionar: D3..E11									
17		Entrada: B3		(=Célula de entrada da coluna)							
18											
19											

Outro exemplo é a avaliação de uma seqüência, por exemplo a famosa fórmula de Euler que já estudamos nas páginas 48 e 119. A planilha seguinte é fácil de entender:



B5 =(1+1/$B$3)^($B$3)

	A	B	C	D	E	F	G	H	I	J
1	**Função:**	(1+1/n)^n								
2				**n**	**e**					
3	**n=**	1		1	2	=B5				
4				2	2,25	{=TABELA(;B3)}		o Excel coloca as chaves		
5	**e ~**	2		3	2,3703704					
6				4	2,4414063					
7				5	2,48832		**Número de Euler**			
8				6	2,5216264		e~2,718281828....			
9				7	2,5464997					
10				8	2,5657845					
11				9	2,5811748					
12				10	2,5937425					

Aproveitemos deste exemplo, para introduzir a técnica de trabalhar com **arquivos seqüenciais**.

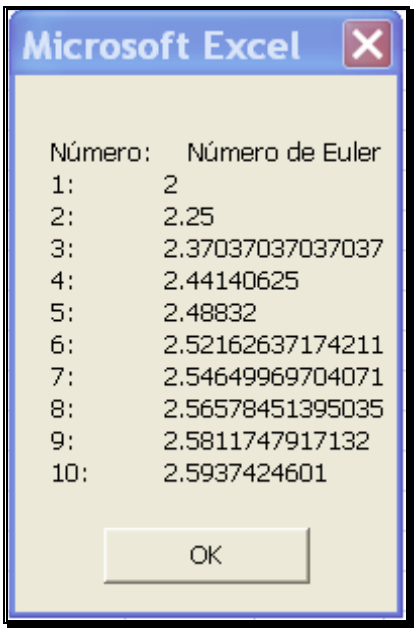
A primeira sub-rotina cria um arquivo seqüencial com números da forma $(1+1/n)^n$ que devem tender ao infinito para valores de n crescendo cada vez

mais. A segunda sub-rotina lê os números e mostra num MsgBox os números criados.

```
Sub CriarNumerosEuler()

Dim entrada As Double
Dim contador As Double
Dim Numero As Double
entrada = InputBox("Quantos números deseja criar?")

Open "c:\Numeros_de_Euler.txt" For Output As #1
Write #1, "Número", "   Número de Euler"

For contador = 1 To entrada
   Numero = (1 + 1 / contador) ^ contador
   Write #1, contador; Numero
Next

Close #1
MsgBox "Foram criados " & entrada & " números."

End Sub
```

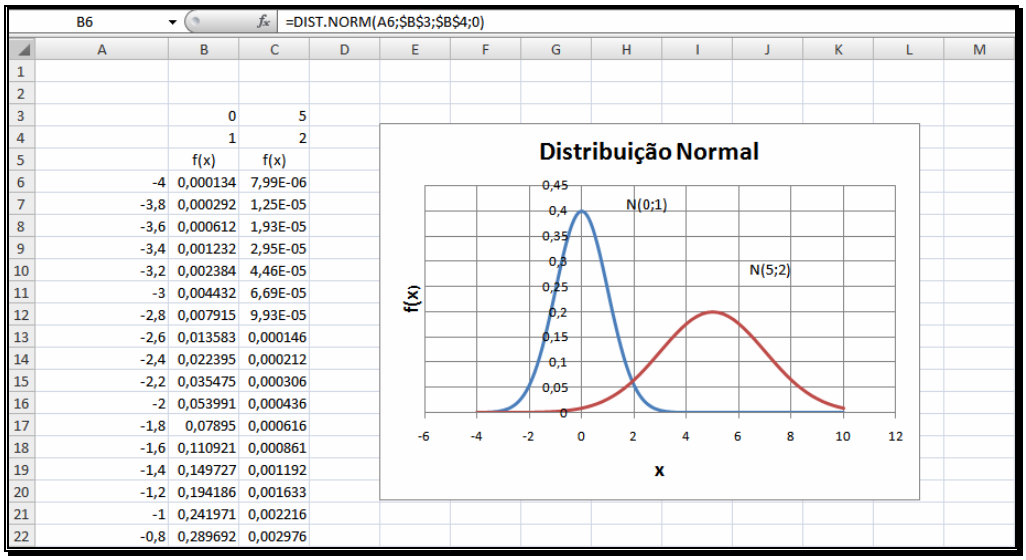
```
Sub LerNumerosEuler()
Dim Numero As String
Dim NumeroEuler As String
Dim Saida As String

Open "c:\Numeros_de_Euler.txt" For Input As #1
    Do While Not EOF(1)
        Input #1, Numero, NumeroEuler
        Saida = Saida & vbCr & Numero & ":" & vbTab & NumeroEuler
    Loop

Close #1

MsgBox Saida

End Sub
```

Aqui temos o "output" da sub-rotina LerNumerosEuler:



A convergência da seqüencia $e = \lim_{n\to\infty}(1+\frac{1}{n})^n$ é muito lenta. O valor de $(1+1/1000)^{1000}$ é 2,7169239.

Obviamente podemos traçar as curvas das funções F(x) ou f(x) sem TABELA, pois temos a função DIST.NORM. Vamos, então, e tracemos N(0,1) e N(5,2) fazendo uso desta última função:



	A	B	C
3		0	5
4		1	2
5		f(x)	f(x)
6	-4	0,000134	7,99E-06
7	-3,8	0,000292	1,25E-05
8	-3,6	0,000612	1,93E-05
9	-3,4	0,001232	2,95E-05
10	-3,2	0,002384	4,46E-05
11	-3	0,004432	6,69E-05
12	-2,8	0,007915	9,93E-05
13	-2,6	0,013583	0,000146
14	-2,4	0,022395	0,000212
15	-2,2	0,035475	0,000306
16	-2	0,053991	0,000436
17	-1,8	0,07895	0,000616
18	-1,6	0,110921	0,000861
19	-1,4	0,149727	0,001192
20	-1,2	0,194186	0,001633
21	-1	0,241971	0,002216
22	-0,8	0,289692	0,002976

Todos os valores vão da linha 6 até 76.
B6: =DIST.NORM(A6;$B$3;$B$4;0)
C6: =DIST.NORM(A6;$C$3;$C$4;0)

As curvas têm dois pontos de inflexão, simétricos em relação à média, que ocorrem quando x = μ+σ e x = μ-σ.

Graficamente, as curvas têm forma de sino, com concavidade voltada para baixo entre os pontos de inflexão da curva, e convexidade para aquém e além desses pontos. O máximo de uma curva têm as coordenadas $(\mu;\frac{1}{\sigma\sqrt{2\pi}})$. Assim, os máximos ficam em (0;0,399) para N(0;1) e em (5;0,1995) para N(5;2)

A inversão de Φ

Para determinar intervalos de confiança e para os testes de hipóteses, precisamos para um valor dado da função Φ(z) o valor z correspondente. Isso significa que devemos resolver a equação Φ(z) = 1-α com respeito a z. Não é possível fazer isso em forma analítica, mas existem vários métodos numéricos para esta tarefa. A função do Excel INV.NORM(γ;μ;σ) se baseia numa destes

métodos aproximativos e retorna o inverso da distribuição cumulativa normal para a média específica e o desvio padrão dados.

A seguinte planilha tem em F10 a função:

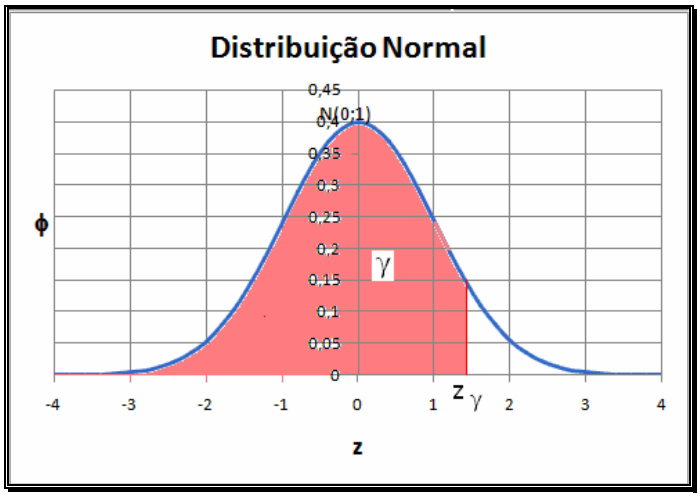
=INV.NORM(Se(B5=1;B13;0,5+B13/2);B9;B10) e em F11:
=INV.NORM(Se(B5=1;B13;0,5+B13/2);0;1)
Veja, também, as explicações para a distribuição_t mais à frente.

F10 =INV.NORM(SE(B5=1;B13;0,5+B13/2);B9;B10)

	A	B	C	D	E	F	G	H
1								
2								
3			**Inversão da função de distribuição**					
4								
5	unilateral?	1						
6	(sim = 1; não = 0)							
7								
8				Quantiles				
9	Média:	5,2						
10	Desvio padrão:	1,25		x para distr. N(u,s):		6,8019		
11								
12				z para distr. N(0,1):		1,2816		
13	Probabilidade:	0,9						
14								

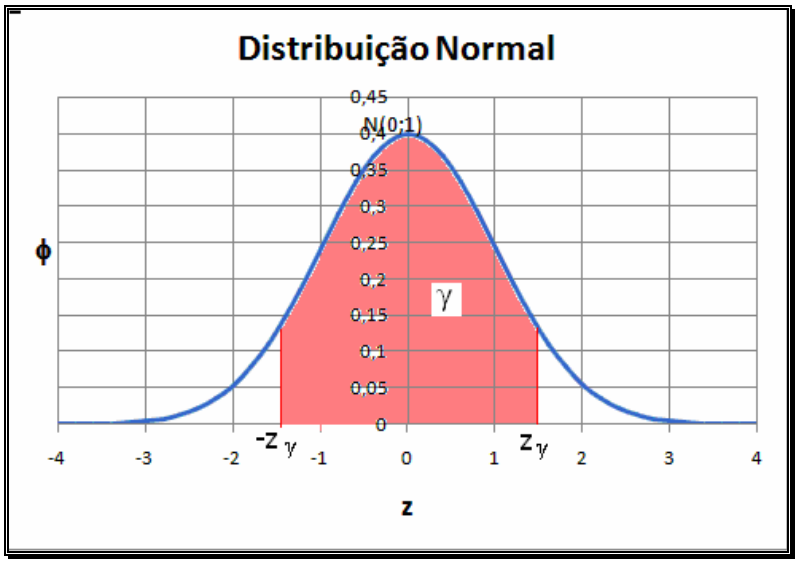
O número real z_γ no qual a função distribuição de probabilidade Φ corresponde ao valor γ na equação (integral) $P(X<=z_\gamma) = \Phi(z_\gamma) = \gamma$ é chamado de **γ-quantil** ou 100γ-percentil. Geometricamente, z_γ corresponde ao limite direito da área γ sob a curva da função Φ(z).

γ-quantil de Z unicaudal:



O γ-quantil de Z bicaudal é ilustrado pela seguinte figura.

γ-quantil de Z bicaudal:



A área sob a curva normal (na verdade abaixo de qualquer função de densidade de probabilidade) é 1. Então, para quaisquer dois valores específicos podemos determinar a proporção de área sob a curva entre esses dois valores. Para a distribuição Normal, a proporção de valores caindo dentro de um, dois, ou três desvios padrão da média são:

Intervalo	Probabilidade
μ±1σ	68,3%
μ±2σ	95,5%
μ±3σ	99,7%

Veja p. 8

Intervalo de confiança

Problemas sobre intervalos de confiança para a média μ desconhecida de certa população têm muitas vezes a forma do seguinte exemplo:

Recebemos uma quantidade grande, N, de baterias, das quais queremos saber em qual intervalo se encontra a média da variável $\overline{X}$ (=duração da bateria).

O método a usar recomenda avaliar uma amostra (tamanho n). Se o desvio padrão $\sigma_\mu = \sigma_x/\sqrt{n}$ for conhecido, sabemos que o intervalo aleatório

$$[\overline{X} - z\frac{\sigma_x}{\sqrt{n}}; \overline{X} + z\frac{\sigma_x}{\sqrt{n}}]$$

contém μ com a probabilidade de confiança γ = 2Φ(z)-1 (intervalo bicaudal). Assim, devemos determinar z pela inversão de Φ. Se σ_x não for conhecido, utilizamos a estimativa

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

Se o tamanho da amostra for n<30, devemos utilizar a distribuição **t**. O Excel tem para o intervalo de confiança a função INT.CONFIANÇA.

Na planilha a seguir utilizamos INV.NORM para a inversão de Φ e INT.CONFIANÇA(alfa;desv_padrao;tamanho) para o intervalo bicaudal de confiança. α é o nível de significância utilizado para calcular o nível de confiança. O nível de confiança é igual a 1-α, ou 100·(1-α)%, ou, em outras palavras, um alfa de 0,05 indica um nível de confiança de 95%. Chama-se γ = 1-α também de coeficiente de confiança.

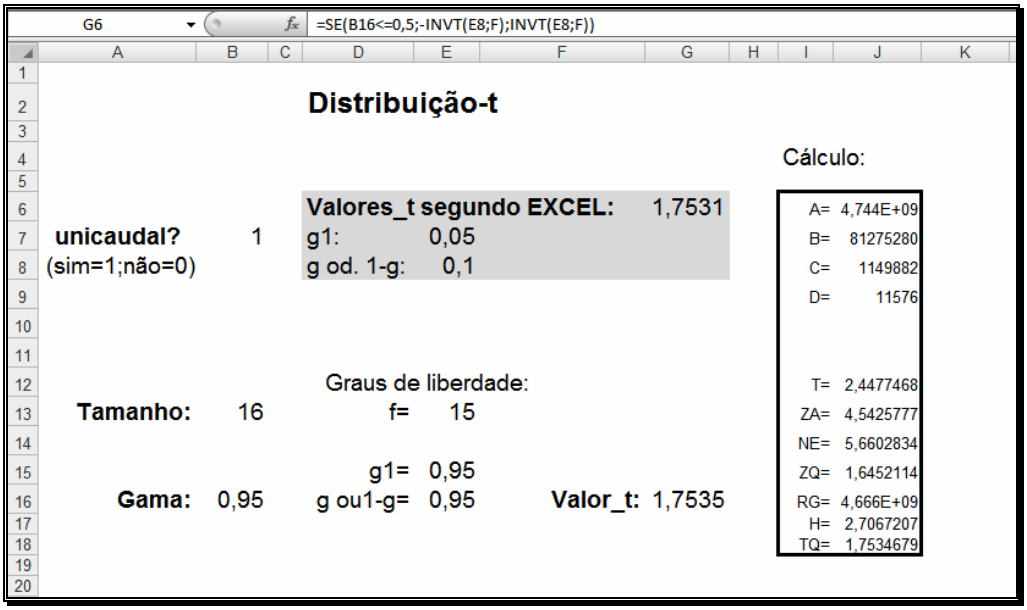
F18 fx =INT.CONFIANÇA(1-B16;B15;B13)

	A	B	C	D	E	F	G	H	I	J
1										
2					**Intervalo de confiança para a média**					
3					para amostras grandes					
4					(Distribuição Normal)					
5										
6										
7	unicaudal_acima?	0	=não							
8	unicaudal_abaixo?	0								
9	bicaudal?	1	=sim	23,257	**<=u<=**	23,743				
10										
11	Amostra:									
12										
13	tamanho:	100								
14	média:	23,5			0,95		1,645	= valor_z		
15	desvio padrão:	1,48								
16	nível de confiança:	0,9			0,95		0,243	= erro amostral		
17										
18				=INT.CONFIANÇA:		0,243				
19										

Amostras de tamanho pequeno

No ano 1908, W.S.Gosset propus a distribuição "Student", também chamada de distribuição-t, que, no caso de amostras de tamanho pequeno, substitui a distribuição Normal. (Student é um pseudônimo de William Sealy Gosset, que não podia publicar artigos usando seu próprio nome.)

Para calcular os intervalos de confiança, precisamos dos assim chamados valores_t (t_quantiles), ou seja, precisamos da solução da equação integral $\Phi_s(t_{1-\alpha;f}) = 1-\alpha$. Na seguinte planilha, calculamos os valores_t de duas maneiras. Primeiro, utilizando a função INV(p;f) do Excel que retorna o inverso da distribuição_t. Segundo, utilizamos um dos algoritmos desenvolvidos para a inversão da função de distribuição Student.



Entradas para Excel:

G6: =SE(B16<=0,5;-INVT(E8;F);INVT(E8;F)); (nomeei E13 de F)
E7: =SE(B7=1;1-B16;0,5+B16/2)
E8: =SE(E7<=0,5;2*E7;2*(1-E7))

A fórmula que foi usada no intervalo J6:J18 é

$$t \approx (au+bu^3+cu^5+du^7+eu^9)/(92160f^4)$$

As constantes são definidas da seguinte forma:

$a = 92160f^4+23040f^3+2880f^2-3600f-945$
$b = 23-40f^3+15360f^2+4080f-1920$
$c = 4800f^2+4560f+1482$
$d = 720f+776$; $e = 79$
u = quantil da distribuição N(0;1)

O cálculo de u nas células J12:J18 baseia-se na seguinte fórmula

$z \approx t-(a+bt+ct^2)/(1+dt+et^2+ft^3)$ com $t = \sqrt{(-2\ln(1-\gamma))}$

a = 2,515517; b = 0,802853; c = 0,010328
d = 1,432788; e = 0,189269; f = 0,001308

Obtemos os z-quantiles dos valores 0 < γ <= 0,5 com $z_\gamma = -z_{1-\gamma}$. Com estes z_γ-quantiles da distribuição N(0,1) determinamos em seguida os x_γ-quantiles da distribuição N(μ,σ) usando $x_\gamma = \mu+\sigma z_\gamma$.

Seguem aqui as entradas para o cálculo de t:

J6 (=A): =92160*F^4+23040*F^3+2880*F^2-3600*F-945
J7 (=B): =23040*F^3+15360*F^2+4080*F-1920
J8 (=C): =4800*F^2+4560*F+1482
J9 (=D): =720*F+776

Segue o cálculo do quantil da distribuição N(0;1)

J12 (T): =RAIZ(-2*LN(1-Q))
J13 (ZA): =2,515517+T*(0,802853+0,010328*T)
J14 (NE): =1+T*(1,432788+T*(0,189269+0,001308*T))
J15 (ZQ): =T-ZA/NE
J16 (RG): =92160*F^4
J17 (H): =ZQ^2
J18 (TQ): =ZQ*(A+H*(B+H*(_C+H*(D+79*H))))/RG

G16: =SE(B16<=0,5;-TQ;TQ)
E15: =SE(B7=1;B16;0,5+B16/2)
E16: =SE(E15<=0,5;1-E15;E15)

Os resultados do Excel e os das fórmulas diferem na quarta casa decimal. A implementação das fórmulas é complicada e o uso da fórmula INVT é, obviamente, preferível à implementação das fórmulas. Por outro lado, é interessante saber o que se esconde por detrás de INVT.

É bom saber que para amostras grandes (n > 30) a distribuição_t se aproxima a uma distribuição Normal.

Intervalo de confiança para a distribuição "t"

Temos uma amostra *pequena* com $\overline{x}$ e *s* calculados (n<30). Queremos saber em que intervalo podemos esperar a média μ. O intervalo buscado podemos escrever como $\overline{x} - a_{\overline{x}} < \mu < \overline{x} + a_{\overline{x}}$ onde $a_{\overline{x}}$ é o erro da estimativa da média da população (erro de amostragem). $a_{\overline{x}}$ pode ser estimado através da seguinte

expressão $a_{\bar{x}} = \frac{s}{\sqrt{n}} t_{1-\alpha;\, f}$, no caso de um intervalo de confiança unicaudal. No caso dum intervalo bicaudal, temos de usar α/2 em vez de α. Se se tirar uma amostra (n) de uma população (N) pequena, precisa-se introduzir um fator de correção k = $\sqrt{\frac{N-n}{N-1}}$.

Em base nestas aclarações, criamos uma planilha do Excel.

E15: =SE(B9=0;1-B16;0,5+B16/2)
E16: =SE(E15<=0,5;2*E15;2*(1-E15))
E7: =SE(B7=1;B14-G19;"")
E8: =SE(B8=1;B14+G19;"")
E9: =SE(B9=1;B14-G19;""); E13 = F
F7: =SE(B7=1;"<=μ";"")
D8: =SE(B8=1;" μ<=";"")
G9: =SE(B9=1;B14+G19;"")
G16: =SE(B16<=0,5;-INVT(E16;F);INVT(E16;F))
G19: =B15*G16/RAIZ(B13) (erro de amostragem)

	A	B	C	D	E	F	G	H	I	J
1										
2				Intervalo de confiança para μ						
3				para amostras pequenas						
4				(Distribuição t)						
5										
6	sim=1;não=0									
7	unicaudal-acima?	0								
8	unicaudal-abaixo?	0								
9	bicaudal?	1			9,69163	<=μ<=	11,2684			
10										
11	Amostra:									
12										
13	Tamanho:	10		f=	9	=graus de liberdade (n-1)				
14	Média:	10,48								
15	Desvio padrão:	1,36		Q2=	0,95					
16	Coef. de confiança:	0,9		Q1=	0,1	Valor_t:	1,83311			
17										
18						Erro da				
19						amostra:	0,78837			
20										

Exemplo:

Dez mensurações (=amostra) são feitas para a resistência de um certo tipo de fio, fornecendo os valores X_1, ...,X_{10}. Suponha-se que $\overline{X}$ =10,48 ohms e σ = 1,36 ohms. Vamos supor que X tenha distribuição N(μ,σ) e que desejemos obter um intervalo de confiança para μ, com coeficiente de confiança γ = 0,90. Portanto, α = 0,10.

A planilha "Distribuição-t" determina que o valor-t é 1,833. Conseqüentemente, o intervalo de confiança procurado será:

$(10{,}48 - 10^{-0.5}(1{,}83)(1{,}36);\ 10{,}48 + 10^{-0.5}(1{,}83)(1{,}36)) = (9{,}69;\ 11{,}27)$

Este intervalo corresponde ao resultado determinado pela última planilha.

Ao afirmar que (9,69;11,27) constitui um intervalo de confiança de 90% para μ, não estaremos dizendo que 90% das vezes a média amostral cairá naquele intervalo. A próxima vez que tiramos uma amostra aleatória, $\overline{X}$ presumível-mente será diferente e, por isso, os extremos do intervalo de confiança serão diferentes. O que estamos dizendo é que 90% das vezes, μ estará contido no intervalo $(\overline{X} - 1{,}83\sigma/\sqrt{n},\ \overline{X} + 1{,}83\sigma/\sqrt{n})$.

Testes de Hipóteses

Nesta seção, encontraremos outra maneira de tratar o problema de fazer uma afirmação sobre um parâmetro desconhecido. Consideremos o seguinte **exemplo**:

Um fabricante declara que a duração da vida X das N = 3000 baterias enviadas é pelo menos 230 horas (*hipótese de nulidade*). O fabricante e o comprador das baterias são decididos a testar a hipótese de nulidade H_0: μ≥230 contra a *hipótese alternativa* H_a: μ < 230. Ao mesmo tempo querem determinar um intervalo de confiança para a média μ desconhecida (sabe-se que a média aritmética $\overline{X}$ dos valores de uma amostra de tamanho n constitui uma "boa" estimativa de μ). Eles analisam uma amostra de n = 50 baterias e encontram para μ uma estimativa de 223 horas; a estimativa do desvio padrão é s = 21 horas.

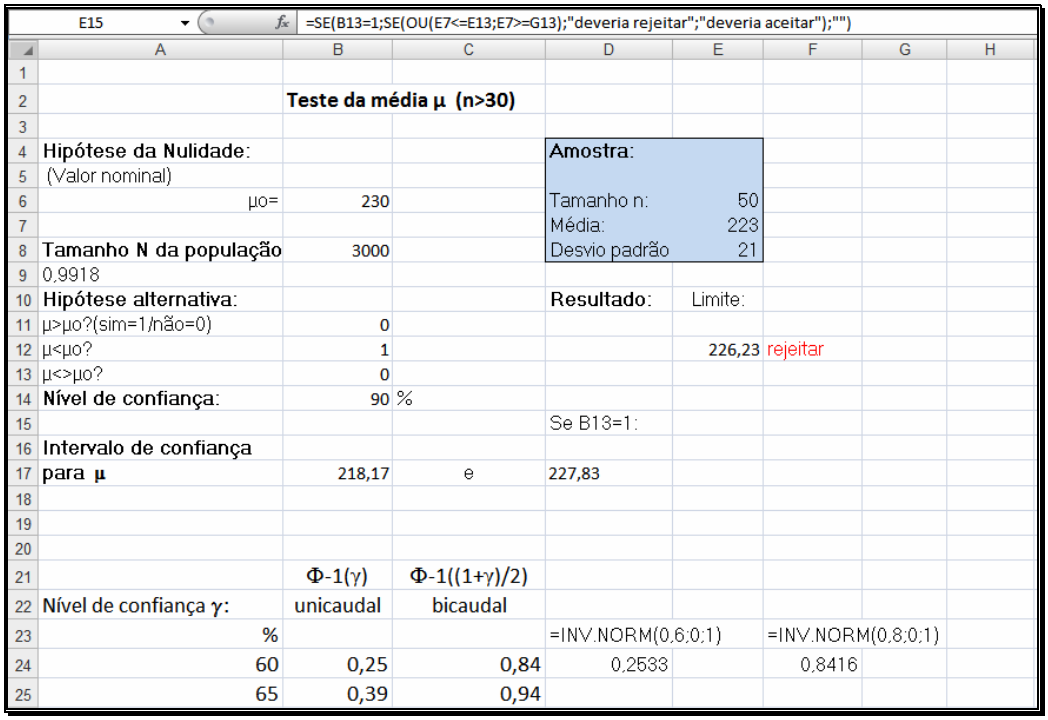
Para variar a metodologia, buscamos os valores de $z = \Phi^{-1}(\gamma \text{ ou } (1+\gamma)/2)$ numa pequena tabela que colocamos no bloco A24:C32

21		Φ-1(γ)	Φ-1((1+γ)/2)				
22	Nível de confiança γ:	unicaudal	bicaudal				
23	%			=INV.NORM(0,6;0;1)		=INV.NORM(0,8;0;1)	
24	60	0,25	0,84	0,2533		0,8416	
25	65	0,39	0,94				
26	70	0,52	1,04				
27	75	0,67	1,15				
28	80	0,84	1,28				
29	85	1,04	1,44				
30	90	1,28	1,64				
31	95	1,64	1,96				
32	99	2,326	2,575	1,28	unicaudal		
33	Erro amostral:						
34	3,7702	1		1,64	bicaudal		
35	Fator de correção:	0,991797					

Podemos encontrar os valores nesta tabela numa tábua da distribuição Normal ou por meio da função INV.NORM(γ;0;1).

Entradas:

D32: =PROCV(B14;A24:C32;B34+1)
A34: = $D$32*E8/RAIZ(E6)*B35 (multiplicação com o fator B35: =RAIZ((B8-E6)/(B8-1)))
B34: =SE(B13=1;2;1)
D34: = PROCV(B14;A24:C32;3)
E11: =SE(B11=1;B6+A34;"")
F11: =SE(B11=1;SE(E$7>=E11;"rejeitar";"aceitar");"")
E12: =SE(B12=1;B6-A34;"")
F12: =SE(B12=1;SE(E$7<=E12;"rejeitar";"aceitar");"")
E13: =SE(B13=1;B6-$D$32*E8/RAIZ(E6);"")
G13: =SE(B13=1;B6+$D$32*E8/RAIZ(E6);"")
E15: =SE(B13=1;SE(OU(E7<=E13;E7>=G13);"deveria rejeitar";"deveria aceitar");"")
B17: =E7-$D$34*E8/RAIZ(E6)*B35
D17: =E7+$D$34*E8/RAIZ(E6)*B35



E15 =SE(B13=1;SE(OU(E7<=E13;E7>=G13);"deveria rejeitar";"deveria aceitar");"")

	A	B	C	D	E	F	G	H
1								
2		**Teste da média μ (n>30)**						
3								
4	**Hipótese da Nulidade:**			**Amostra:**				
5	(Valor nominal)							
6	μo=	230		Tamanho n:	50			
7				Média:	223			
8	**Tamanho N da população**	3000		Desvio padrão	21			
9	0,9918							
10	**Hipótese alternativa:**			**Resultado:**	Limite:			
11	μ>μo?(sim=1/não=0)	0						
12	μ<μo?	1			226,23	rejeitar		
13	μ<>μo?	0						
14	**Nível de confiança:**	90	%					
15				Se B13=1:				
16	**Intervalo de confiança**							
17	**para μ**	218,17	e	227,83				
18								
19								
20								
21		Φ-1(γ)	Φ-1((1+γ)/2)					
22	Nível de confiança γ:	unicaudal	bicaudal					
23	%			=INV.NORM(0,6;0;1)		=INV.NORM(0,8;0;1)		
24	60	0,25	0,84	0,2533		0,8416		
25	65	0,39	0,94					

Comparação de duas Médias

Dois instrumentos (multímetros) são usados para medir a intensidade da corrente elétrica. O instrumento 1 produziu com 8 medições $\overline{x}_1$= 1,486, o instrumento 2 deu com 13 medições $\overline{x}_2$ = 1,492. Os desvios padrões dos instrumentos foram s_1 = 0,026 e s_2 = 0,021. (A amostra com o maior desvio recebe o índice 1.) A pergunta que se impõe é: As leituras de ambos os instrumentos são significativamente diferentes ou pode-se dizer que as médias μ_1 e μ_2 das populações subjacentes são idênticas?

Para testar isso, devemos saber se as duas populações têm as mesmas variâncias (Teste-F). (Isso é o caso se o quociente s^2_1/s^2_2 é menor do que o valor correspondente $F_{1-\alpha;\, f_1,\, f_2}$ da distribuição F que se obtém para α = 0,05 por meio de =INVF(0,05;7;12) (= 2,91). Este valor é maior do que $(s_1/s_2)^2$ = 1,53.)

Como quantidade de teste y, utilizamos a diferença d = $\overline{x}_1$-$\overline{x}_2$: $y = \frac{d}{s}\sqrt{\frac{n_1 n_2}{n_1+n_2}}$. A variância total vem dada por $s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$ (= pooled variance = variância amostral combinada).

Temos como hipótese da Nulidade H_0: $\mu_1 = \mu_2$ e como hipóteses alternativas

$$H_a : \mu_1<\mu_2;\ \mu_1>\mu_2;\ \mu_1\#\mu_2$$

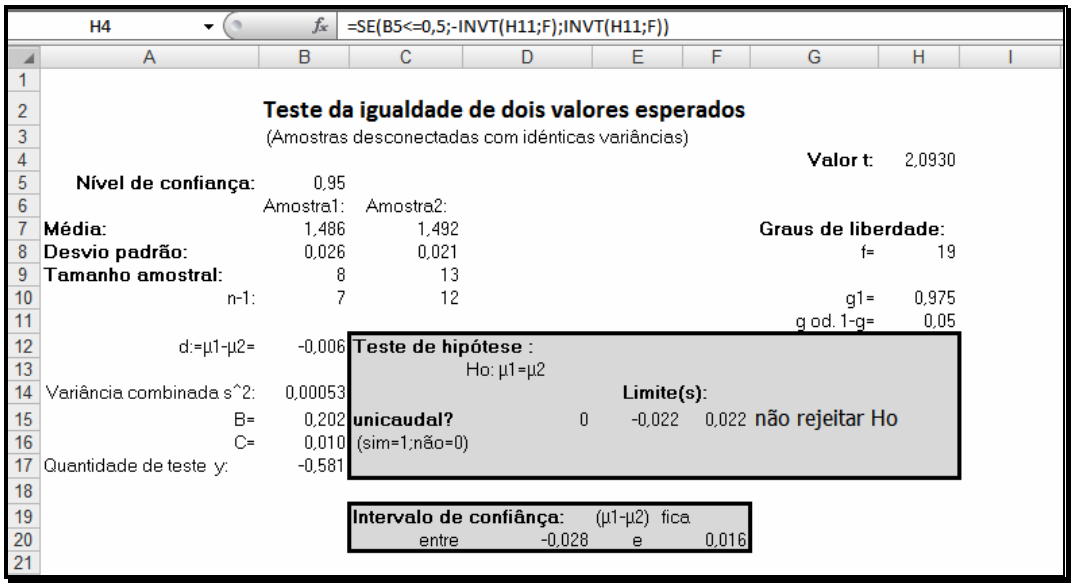
No caso $\mu_1>\mu_2$ rejeitamos H_0, se $y > t_{1-a;f}$. (Obtemos o valor de t com nossa planilha da distribuição t.) Se escolhermos $\mu_1<\mu_2$, teremos como critério de rejeição de H_0: $y < -t_{1-\alpha;f}$.

Geralmente, escolhe-se $\mu_1\#\mu_2$ e rejeita-se H_0, se $|y| > t_{1-\alpha/2;f}$.

O intervalo de confiança de d é **(d-t·d/y; d+t·d/y)**.

Entradas:

B14: =((B10*B8^2+C10*C8^2)/H8); B15: =(1/B9+1/C9)
B16: =RAIZ(B14*B15); B17: =B12/B16 (=y)
E15: =SE(D15=1;H4*B16;-H4*B16)
F15: =SE(D15=0;H4*B16;""); H8: =B9+C9-2 (graus de liberdade)
G15: =SE(D15=1;SE(B12>E15;"μ1 é maior do que μ2";"μ2 é maior do que μ1");SE(OU(B12<E15;B12>F15);"rejeitar H_0";"não rejeitar Ho"))
D20: =SE(D15=0;B12-H4*B12/B17;"")
F20: =SE(D15=0;B12+H4*B12/B17;"")
H4: =SE(B5<=0,5;-INVT(H11;F);INVT(H11;F))
H10: =SE(D15=1;1-B5;0,5+B5/2)
H11: =SE(H10<=0,5;2*H10;2*(1-H10))



H4 | =SE(B5<=0,5;-INVT(H11;F);INVT(H11;F))

Teste da igualdade de dois valores esperados

(Amostras desconectadas com idênticas variâncias)

				Valor t:	2,0930
Nível de confiança:	0,95				
	Amostra1:	Amostra2:			
Média:	1,486	1,492		**Graus de liberdade:**	
Desvio padrão:	0,026	0,021		f=	19
Tamanho amostral:	8	13			
n-1:	7	12		g1=	0,975
				g od. 1-g=	0,05
d:=μ1-μ2=	-0,006	**Teste de hipótese :**			
		Ho: μ1=μ2			
Variância combinada s^2:	0,00053			**Limite(s):**	
B=	0,202	**unicaudal?**	0	-0,022	0,022 não rejeitar Ho
C=	0,010	(sim=1;não=0)			
Quantidade de teste y:	-0,581				

Intervalo de confiança:	(μ1-μ2)	fica		
entre	-0,028	e	0,016	

Conclusão: O teste não pode rejeitar a H_0, porque a diferença d = -0,006 está dentro do intervalo (-0,022; 0,022). Com um nível de confiança de 90% podemos supor que as duas populações saíram da mesma população-mãe. Ao nível de significância de 5%, a leitura do instrumento 1 não é significativamente diferente da leitura do instrumento 2.

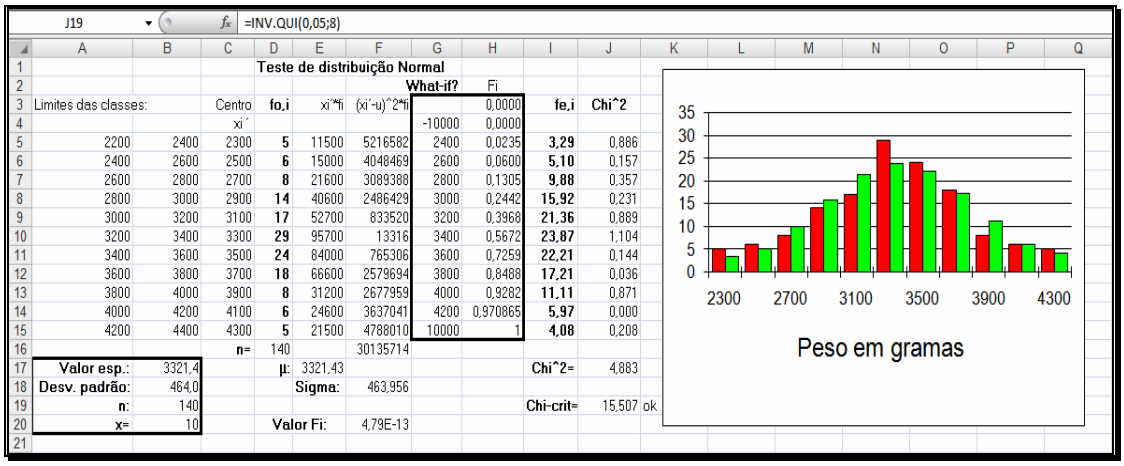
Teste Qui-Quadrado (χ^2; χ = letra grega chi)

Deseja-se verificar a afirmação de que o peso de meninas recém-nascidas segue a distribuição Normal. Numa clinica foram pesadas n = 140 meninas e seus pesos distribuídos sobre 11 classes (blocos, bins) cada um de 200g.

Precisamos das seguintes informações:

1. Os centros x_i' dos intervalos e as freqüências absolutas observadas $f_{o,i}$.
2. Fórmula para o cálculo do valor esperado para dados classificados em k classes e n observações: $\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i' f_i$
3. Fórmula para a variância amostral: $s^2 = \frac{1}{n}\sum_{i=1}^{k} (x_i' - \overline{x})^2 f_i$
4. Fórmula para χ^2 : $\chi^2 = \sum_{i=1}^{k} \frac{(f_{o,i} - f_{e,i})^2}{f_{e,i}}$; f_e = freqüência esperada
5. A função Φ ("Fi"): =DIST.NORM(x;média;desv_padrão;1)
6. A função {=TABELA(;Bx)} do menu *Dados*, veja "Distribuição Normal"
7. A função =INV.QUI(α;f) para determinar o valor crítico de χ (Qui). f = 11 – 3 = 8 (número de classes – condições = número de graus de liberdade).

As colunas A, B e D contêm os valores observados.

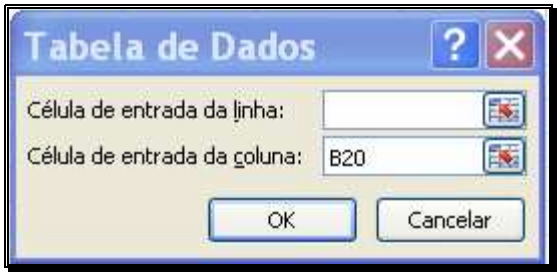


J19 =INV.QUI(0,05;8)

	A	B	C	D	E	F	G	H	I	J
1				Teste de distribuição Normal						
2							What-if?	Fi		
3	Limites das classes:		Centro	fo,i	xi'*fi	(xi'-u)^2*fi		0,0000	fe,i	Chi^2
4			xi'				-10000	0,0000		
5	2200	2400	2300	5	11500	5216582	2400	0,0235	3,29	0,886
6	2400	2600	2500	6	15000	4048469	2600	0,0600	5,10	0,157
7	2600	2800	2700	8	21600	3089388	2800	0,1305	9,88	0,357
8	2800	3000	2900	14	40600	2486429	3000	0,2442	15,92	0,231
9	3000	3200	3100	17	52700	833520	3200	0,3968	21,36	0,889
10	3200	3400	3300	29	95700	13316	3400	0,5672	23,87	1,104
11	3400	3600	3500	24	84000	765306	3600	0,7259	22,21	0,144
12	3600	3800	3700	18	66600	2579694	3800	0,8488	17,21	0,036
13	3800	4000	3900	8	31200	2677959	4000	0,9282	11,11	0,871
14	4000	4200	4100	6	24600	3637041	4200	0,970865	5,97	0,000
15	4200	4400	4300	5	21500	4788010	10000	1	4,08	0,208
16			n=	140		30135714				
17	Valor esp.:	3321,4		μ:	3321,43				Chi^2=	4,883
18	Desv. padrão:	464,0			Sigma:	463,956				
19	n:	140							Chi-crit=	15,507 ok
20	x=	10		Valor Fi:		4,79E-13				
21										

Entradas:

C5: =(A5+B5)/2, copiar até C15; E5: =C5*D5, copiar
D16: =SOMA(D5:D15) (foi chamado de Numero)
E17: =SOMA(E5:E15)/D16 (=Mu)
F5: =(C5-E$17)^2*D5
F16: =(C5-E$17)^2*D5; F18: =RAIZ(F16/D16) (Sigma)
F20: =DIST.NORM(B20;B17;B18;1)
H3: =F20 (valor Fi)
Selecionar G3:H15 e escolher *Dados/Teste de Hipóteses/ Tabela de Dados*

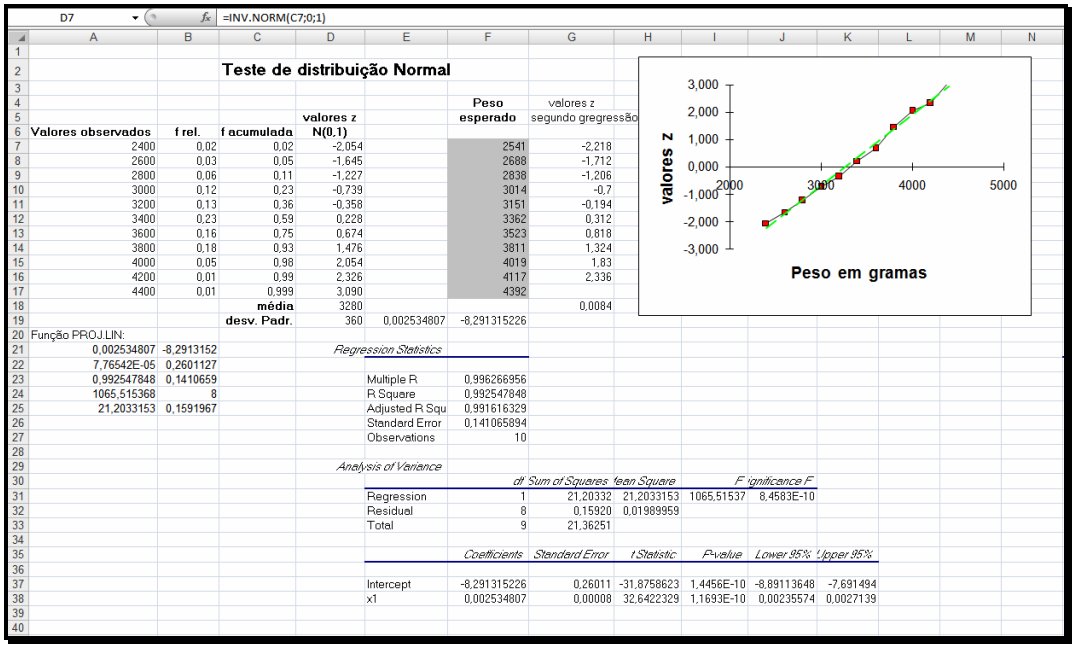


O valor de x em B20 será automaticamente substituído pelos valores em G4:G15. G4 e G15 foram ocupadas te tal forma que H4 dê o valor 0 e H15 1. x tem o valor 10 para dar em H3 também 0.

Na coluna I estão os valores esperados (calculados) da freqüência absoluta.
I5: =(H5-H4)*D16, copiar até I15
J5: =(D5-I5)^2/I5, copiar até J15
J17: =SOMA(J5:J15); valor de Qui^2

O Qui^2-crítico, para o nível de 5%, com f = 8, é 15,51 (=INV.QUI(0,05;8)). O valor observado de Qui^2 é então altamente significativo e há bom motivo para crer que o peso das meninas seja normalmente distribuído. Isso vê-se também no histograma onde os valores calculados (verdes) correspondem satisfatoriamente aos valores observados (vermelhos). (A região crítica é constituída de valores maiores de Qui^2-crítico.)

Quero terminar este exemplo com um análise mais direto do problema. Trata-se duma interpretação gráfica dos dados. Vamos considerar as freqüências acumuladas observadas como probabilidades acumuladas, P(Z≤z), de uma variável aleatória Z = (p – μ)/σ onde p é o peso das meninas recém-nascidas.



D7 =INV.NORM(C7;0;1)

Teste de distribuição Normal

	Valores observados	f rel.	f acumulada	valores z N(0,1)	Peso esperado	valores z segundo gregressão
7	2400	0,02	0,02	-2,054	2541	-2,218
8	2600	0,03	0,05	-1,645	2688	-1,712
9	2800	0,06	0,11	-1,227	2838	-1,206
10	3000	0,12	0,23	-0,739	3014	-0,7
11	3200	0,13	0,36	-0,358	3151	-0,194
12	3400	0,23	0,59	0,228	3362	0,312
13	3600	0,16	0,75	0,674	3523	0,818
14	3800	0,18	0,93	1,476	3811	1,324
15	4000	0,05	0,98	2,054	4019	1,83
16	4200	0,01	0,99	2,326	4117	2,336
17	4400	0,01	0,999	3,090	4392	
18			média	3280		0,0084
19			desv. Padr.	360	0,002534807	-8,291315226

Função PROJ.LIN:

0,002534807	-8,2913152				
7,76542E-05	0,2601127				
0,992547848	0,1410659				
1065,515368	8				
21,2033153	0,1591967				

Regression Statistics

Multiple R	0,996266956
R Square	0,992547848
Adjusted R Squ	0,991616329
Standard Error	0,141065894
Observations	10

Analysis of Variance

	df	Sum of Squares	Mean Square	F	Significance F
Regression	1	21,20332	21,2033153	1065,51537	8,4583E-10
Residual	8	0,15920	0,01989959		
Total	9	21,36251			

	Coefficients	Standard Error	t Statistic	P-value	Lower 95%	Upper 95%
Intercept	-8,291315226	0,26011	-31,8758623	1,4456E-10	-8,89113648	-7,691494
x1	0,002534807	0,00008	32,6422329	1,1693E-10	0,00235574	0,0027139

O gráfico dos valores z (que determinamos com nossa planilha "Inversão da função de distribuição") e do peso p deveria dar uma reta, pois

$$Z = \frac{p-\mu}{\sigma} = \frac{1}{\sigma}p - \frac{\mu}{\sigma}$$

é a equação de uma reta. A intercepção com o eixo-p vai dar o valor esperado μ e a inclinação dará 1/σ.

Na planilha vemos na coluna A os pesos observados e em B as freqüências relativas f_r. Em C temos as freqüências acumuladas: C7: =B7; C8= =B8+C7, copiar até C16. Em C17 colocamos 0,999. D7: =INV.NORM(C7;0;1), copiar até D17.
Antes de seguir adiante, fazemos o gráfico. Observamos que os pontos dos dados observados ficam perto duma reta. Isso nos deixa de pensar que, efetivamente, estamos frente a uma distribuição Normal. O corte da reta com o eixo de p corresponde, mais ou menos, a 3300g. Da inclinação da reta obtemos σ = 360g. São estes os valores experimentais que colocamos nas células D18 e D19. Na coluna F temos os valores esperados de p (F7: =INV.NORM (C7;D$18;D$19)

A planilha mostra também uma análise de regressão feita com "Análise de Dados". A reta da regressão é y = -8,29 + 0,00253x. Na coluna G ficam os valores calculados com esta equação.

É mais simples fazer este análise usando a função PROJ.LIN do Excel. É necessário selecionar duas células adjacentes, por exemplo E19 e F19. A fórmula = PROJ.LIN(D7:D16;A7:A16) é uma fórmula matricial e deve ser inserida pressionando Ctrl+Shift+Enter. Resultado: em E19 aparece o valor 0,00253 e em F19 temos -8,29

Analise de Dados com o módulo de regressão PROJ.LIN

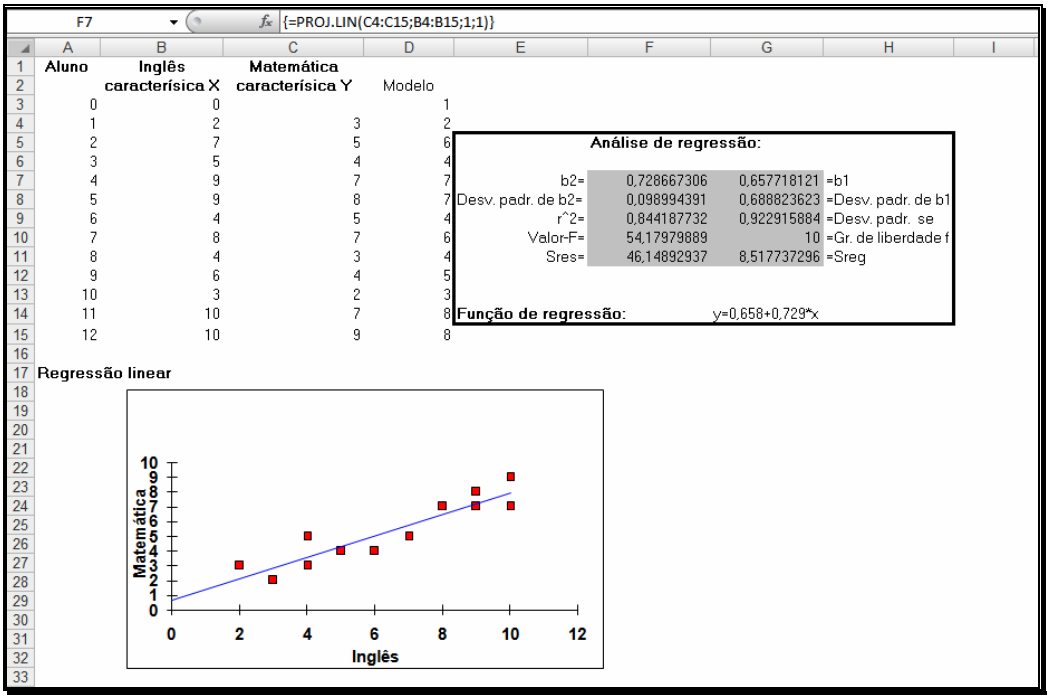
Utilizaremos a função PROJ.LIN quando buscamos relações entre duas ou mais variáveis. Na planilha a seguir analisamos a afirmação de certo professor de que alunos com boas notas em Inglês também são bons em Matemática. O professor quer comprovar esta hipótese com o seguinte material (hipotético):

Aluno	1	2	3	4	5	6	7	8	9	10	11	12
Inglês (X)	2	7	5	9	9	4	8	4	6	3	10	10
Matemática (Y)	3	5	4	7	8	5	7	3	4	2	7	9

Os algarismos na tabela são pontos entre 1 e 10.

Busca-se, usando o Método dos Mínimos Quadrados, a reta de regressão $y = b_1 + b_2x$.

Deixa-se guiar pela seguinte planilha.



Selecione F7:G11 e aplique a função PROJ.LIN(C4:C15;B4:B15;1;1). Ela vai também retornar os dados estatísticos de regressão adicionais. Ao pressionar Ctrl+Shift+Enter, aplicamos a fórmula matricial ao bloco selecionado.

Primeiro, aparecem os coeficientes de regressão b_1 = 0,658; b_2 = 0,729. Debaixo seguem os desvios padrões de b_1 e b_2: o desvio padrão de b_1 fica em G8: 0,689, o de b_2 em F8: 0,099. (Entre estes desvios existe a seguinte relação $s_{b_1} = s_{b_2}\sqrt{\frac{1}{n}\sum x_i^2}$, veja o capítulo anterior, fórmula (6).)

Em F9 encontramos o coeficiente de determinação R^2 = 0,844. Isso significa que 84,4% da variação dos valores y (pontos em Matemática) podem ser explicados pela regressão. (Isso é considerável, se bem que, neste exemplo, puramente hipotético.)

Também podemos calcular os intervalos de confiança para os coeficientes (desconhecidos) β_1 e β_2 da verdadeira reta de regressão $\hat{y} = \beta_1 + \beta_2 x$.

$$b_1 - t \cdot s_{b1} \leq \beta_1 \leq b_1 + t \cdot s_{b1} \text{ e } b_2 - t \cdot s_{b2} \leq \beta_2 \leq b_2 + t \cdot s_{b2}$$

Para f = n - 2 = 10 e 1 – α = 0,95 temos t = 2,228. O intervalo de 95% para β_1 será: -0,877 ≤ β_1 ≤ 2,193.

Regressão linear múltipla

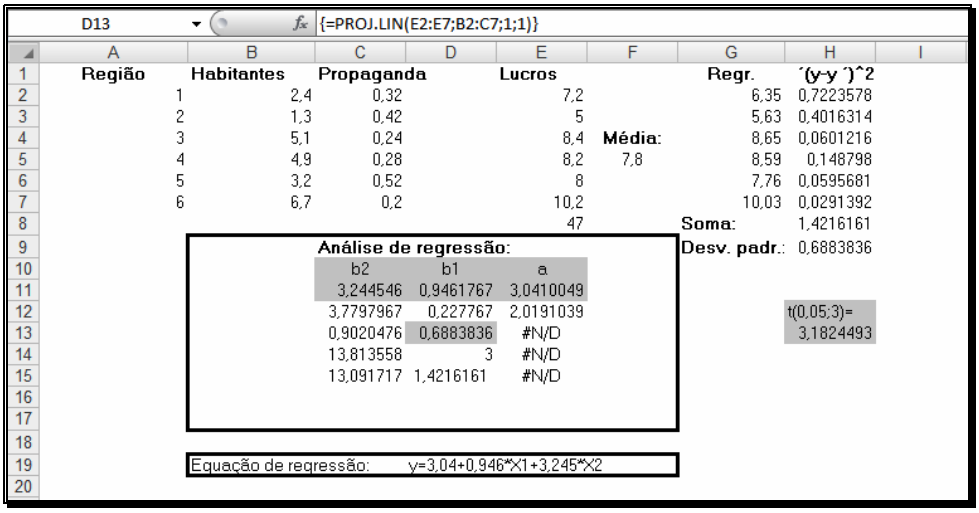
A função PROJ.LIN preste-se, também, para avaliar uma amostra com duas ou mais variáveis como ilustramos no seguinte exemplo.

A direção de uma companhia de cosméticas acha que a ganância y (por persona) do produto "Cheiro de Ouro" não só depende do número de habitantes x_1 da região das vendas, como também das despesas publicitárias x_2 gastas por persona. Os seguintes dados devem ser analisados para detectar uma possível relação.

Região	Habitantes x_1 (Milhões)	Propaganda x_2 ($/persona)	Lucros y (por persona)
1	2,4	0,32	7,2
2	1,3	0,42	5,0
3	5,1	0,24	8,4
4	4,9	0,28	8,2
5	3,2	0,52	8,0
6	6,7	0,2	10,2

Busca-se uma equação de regressão da forma $\hat{y} = a + b_1x_1 + b_2x_2$. $\hat{y}$ é um estimador para o lucro y. Os valores numéricos de $\hat{y}$ denominamos

estimativas. Neste exemplo, não estamos buscando uma *reta*, mas sim um *plano* de regressão.



D13 {=PROJ.LIN(E2:E7;B2:C7;1;1)}

	A	B	C	D	E	F	G	H	I
1	Região	Habitantes	Propaganda		Lucros		Regr.	´(y-y´)^2	
2	1	2,4	0,32		7,2		6,35	0,7223578	
3	2	1,3	0,42		5		5,63	0,4016314	
4	3	5,1	0,24		8,4	Média:	8,65	0,0601216	
5	4	4,9	0,28		8,2	7,8	8,59	0,148798	
6	5	3,2	0,52		8		7,76	0,0595681	
7	6	6,7	0,2		10,2		10,03	0,0291392	
8					47		Soma:	1,4216161	
9			Análise de regressão:				Desv. padr.:	0,6883836	
10			b2	b1	a				
11			3,244546	0,9461767	3,0410049				
12			3,7797967	0,227767	2,0191039			t(0,05;3)=	
13			0,9020476	0,6883836	#N/D			3,1824493	
14			13,813558	3	#N/D				
15			13,091717	1,4216161	#N/D				
16									
17									
18									
19		Equação de regressão:	y=3,04+0,946*X1+3,245*X2						
20									

A coluna G vai receber os valores que calculamos por meio da equação de regressão. Os valores da variável dependente y estão em E2:E7 (E8 contém a sua soma.) Selecione o intervalo C11:E15 para receber a fórmula matricial = PROJ.LIN(E2:E7;B2:C7;1;1), compare com o exemplo anterior.

O erro padrão de y fica em D13 e H8, compare com equação (3) do capítulo anterior. Em nosso caso, s = 0,6884 com $s^2 = \frac{\sum (y - \hat{y})^2}{n - k - 1}$; n = número das observações (6), k = número das variáveis independentes (2). O número dos graus de liberdade é f = n-k-1 = 3

s_2 é o desvio padrão de b_2 e o seu valor de 3,78 é muito grande. De $t = b_2/s_2$ = 3,245/3,78 = 0,858 < $t_{0,05;3}$ = 3,182 (= INVT(0,05;3)) segue que, para um nível de confiança de 95%, b_2 não é significativamente diferente de 0. Isso significa que a propaganda não teve êxito. De fato, obtemos, utilizando somente x_1, um erro padrão de 0,665 e a equação com $\hat{y} = a + b_1x_1 = 4,676 + 0,803 \cdot x_1$ é um modelo satisfatório para os lucros. Disso segue que foram gastas grandes quantidades de dinheiro para as propagandas sem resultar em aumentar os lucros.