Biometria

Distribuição normal

(Leitura complementar ao capítulo 4)

Sumário:

Características

A distribuição normal tem como características fundamentais a média e o desvio padrão.

Para os interessados por Ciências Biológicas é a mais importante das *distribuições contínuas* pois muitas variáveis aleatórias de ocorrência natural ou de processos práticos obedecem esta distribuição.

Abraham de Moivre, um matemático francês exilado na Inglaterra, publicou a função densidade de probabilidade da distribuição normal com média μ e variância σ^2 (ou, de forma equivalente, desvio padrão σ) em 1733:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad \text{em que } -\infty < x < \infty$$

É importante lembrar que os parâmetros populacionais μ e σ possuem os seguintes significados:

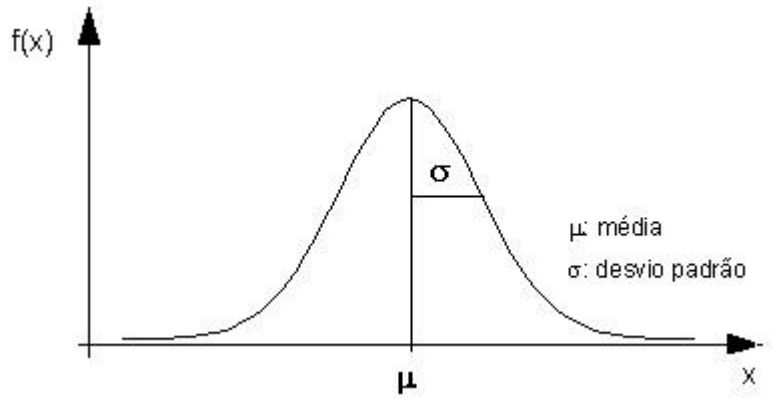
μ = média populacional: indica a posição central da distribuição

σ = desvio padrão populacional: refere-se à dispersão da distribuição

Se uma variável aleatória *x* tem distribuição normal com média μ e variância σ^2, diz-se que

$x \sim N(\mu, \sigma^2)$

A figura a seguir mostra uma curva normal típica, com seus parâmetros descritos graficamente.



A curva normal tem *forma de sino*, ou seja, é *unimodal* e *simétrica*, e o seu valor de máxima frequência, a moda coincide com o valor da média e da mediana.

A *média* é o *centro* da curva.

A distribuição de valores maiores que a média ($x + \mu > 0$) e a dos valores menores que a média ($x + \mu < 0$) é perfeitamente *simétrica*, ou seja, se passarmos uma linha exatamente pelo centro da curva teremos duas metades, sendo que cada uma delas é a imagem especular da outra.

As extremidades da curva se estendem de forma indefinida ao longo de sua base (o eixo das abcissas) sem jamais tocá-la. Portanto, o campo de variação da distribuição normal se estende de - infinito a + infinito.

Assim sendo, a curva apresenta *uma área central em torno da média*, onde se localizam os *pontos de maior frequência* e também possui áreas menores, progressivamente mais próximas de ambas as extremidades, em que são encontrados valores muito baixos de *x* (à esquerda) ou escores muito altos (à direita), ambos presentes em baixas frequências.

Como em qualquer função de densidade de probabilidade a área sob a curva normal é *1*, sendo a frequência total igual a 100%. Assim, a curva normal é uma distribuição que possibilita *determinar probabilidades* associadas a todos os pontos da linha de base.

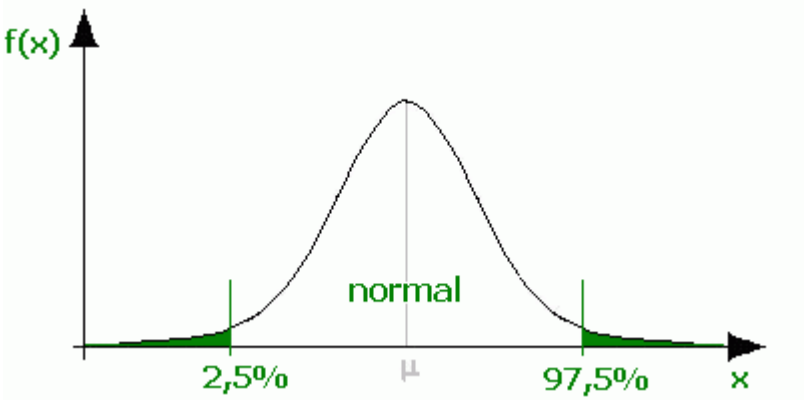
Portanto, tomando-se quaisquer dois valores pode-se determinar a proporção de área sob a curva entre esses dois valores. E essa área é o próprio valor da frequência da característica que ela determina.

Normal e anormal

A palavra *normal* tem um significado coloquial bastante indeterminado, mas tem um significado estatístico bem preciso.

O valor de uma variável tem *ocorrência normal* quando está entre *95%* da área sob a curva em forma de sino, que tem a variável frequência no eixo dos Y, cujas extremidades ocupam 2,5% cada.

Ou seja, algum valor é considerado normal se está na em qualquer ponto entre 0,025 e 0,975 (2,5 e 97,5%) da área sob a curva.



Portanto, há dois tipos de "**anorma**l". Todos os valores encontrados na área que está entre 0 a 2,5% correspondem a um tipo. E todos os que estão no final da curva, ou seja, entre 97,5 e 100% se refiram ao outro tipo.

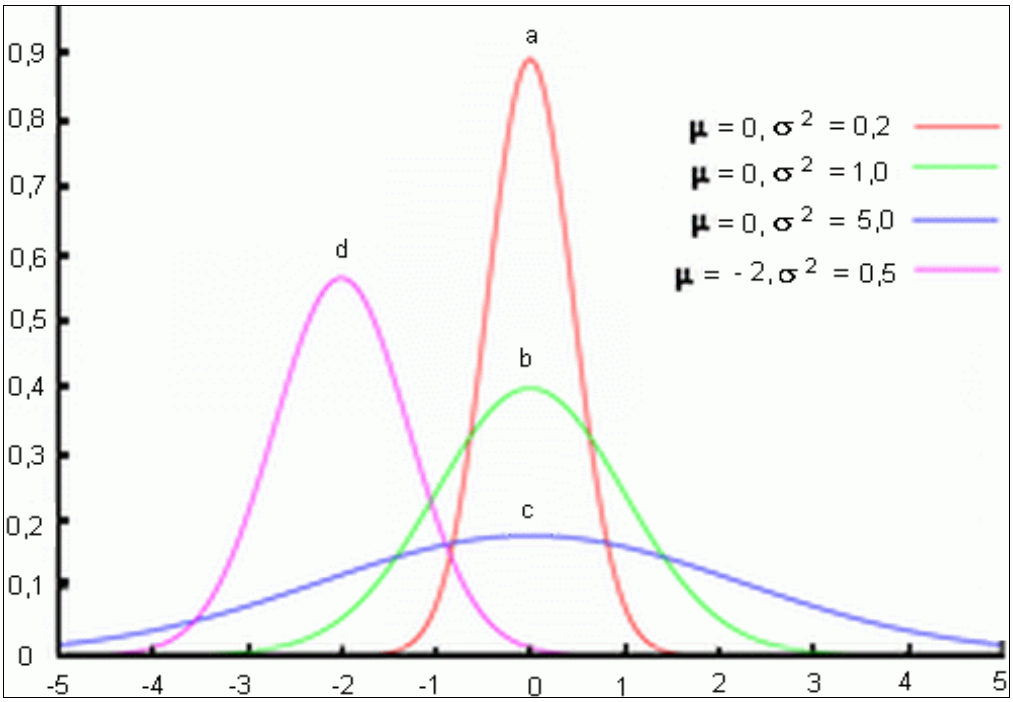
Uma pergunta pra pensar: É sempre ruim ser "anormal"?

É muito importante entender como a curva é afetada pelos valores numéricos de μ e σ.

Assim, como se vê na figura seguinte, em que x corresponde ao número de desvios padrão e Y demonstra a frequência, quanto maior a média, mais à direita está a curva.

Note-se que, se diferentes amostras apresentarem o *mesmo valor de média* μ e *diferentes valores de desvios padrão* σ, a distribuição que tiver o maior desvio padrão se apresentará mais achatada (c), com maior dispersão em torno da média. Aquela que tiver o menor desvio padrão apresentará o maior valor de frequência e acentuada concentração de indivíduos em valores próximos à média (a).

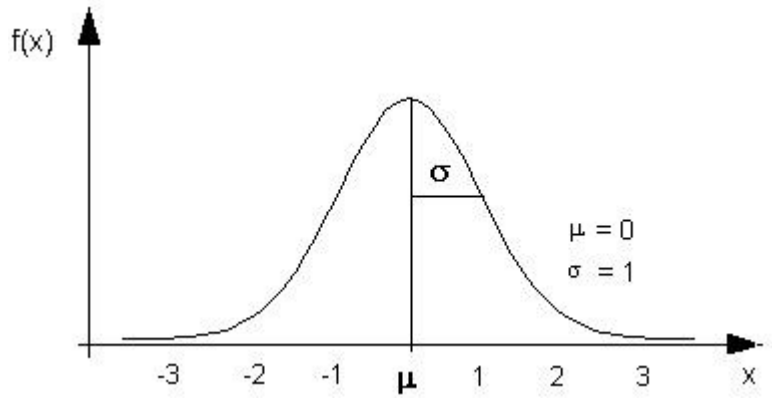
Já, distribuições normais com *valores de médias diferentes* e o *mesmo valor de desvio padrão* possuem a mesma dispersão, mas diferem quanto à localização no eixo dos X.



Distribuição Normal Padrão

Todas as curvas normais representativas de distribuições de frequências podem ser transformadas em uma curva normal padrão, usando-se o desvio padrão (σ) como unidade de medida indicativa dos desvios dos valores da variável em estudo (x), em relação à média (μ).

A *Distribuição Normal Padrão* é caracterizada pela média (μ) igual a *zero* e desvio padrão (σ) igual a *1*.



A figura anterior mostra também que o desvio-padrão controla o grau para o qual a distribuição se "espalha" para ambos os lados da curva. Percebe-se que aproximadamente toda a probabilidade está dentro de ± 3 σ a partir da média.

Se a variável *x* tem distribuição normal, pode ser transformada para uma forma padrão, denominada *Z*, (ou, como comumente se diz, pode ser padronizada) subtraindo-se sua média e dividindo-se pelo seu desvio padrão:

$$z = (x - \mu) / \sigma$$

Quando se estima os coeficientes, usa-se a seguinte notação:

$$z = (x - \bar{x}) / s$$

A equação da curva de **z** é:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} \quad \text{em que } -\infty < z < \infty$$

É importante lembrar que a *área sob a curva* pode ser entendida como uma *medida de sua probabilidade* e que a área sob a curva normal é igual a 1 (100%).

Assim, a variável *x* cuja distribuição é $N(\mu, \sigma^2)$ é transformada na *forma padronizada* **z** cuja distribuição é N(0,1). Essa é a distribuição normal padrão, que já está *tabelada*, pois os parâmetros da população (desvio padrão e média) são conhecidos.

Então, se forem tomados dois valores específicos, pode-se 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:

entre	*é igual a*
μ ± 1 σ	68,26% (1)
μ ± 2 σ	95,44% (2)
μ ± 3 σ	99,74% (3)

Z - dados tabelados

Como se chegou a esses valores?

Para responder essa pergunta é necessário conhecer a distribuição de *z*, que já está tabelada.

Note-se que a Tabela de z determina a *área* a partir do número de desvios-padrão, os quais são lidos assim:

a , b c	a = número inteiro lido na primeira coluna b = número decimal lido na primeira coluna c = número centesimal lido na primeira linha

O valor de *z* será encontrado na *intersecção* entre a coluna e a linha, sendo *adimensional*.

Verificando a tabela, percebe-se que para os valores negativos de *z* as áreas são obtidas por *simetria*, ou seja, existe o mesmo conjunto de valores, com sinal negativo, para o lado esquerdo da média, pois a tabela é *especular*.

Os valores de *z* permitem delimitar a área sob a curva, pois, como no eixo *Y* do gráfico está a *frequência* da variável, a *área sob a curva tem o mesmo valor da probabilidade de ocorrência* daquela característica.

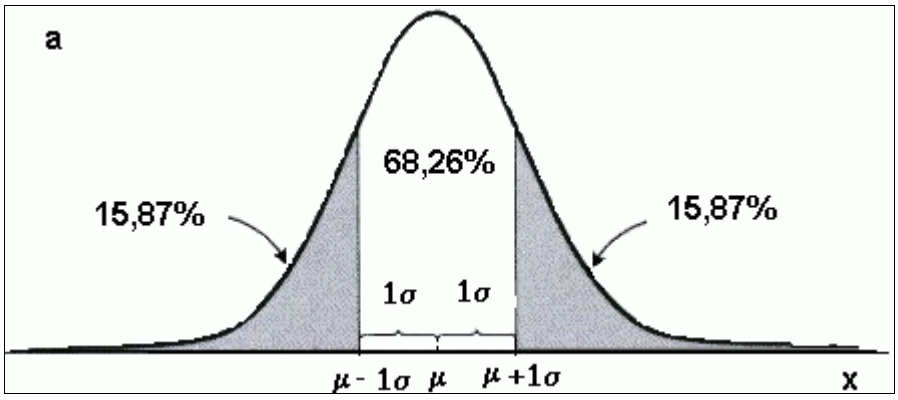
Exemplo 1

Qual é a área sob a curva normal contida entre *z* = 0 e *z* = 1?

Procura-se o valor 1 na primeira coluna da tabela e o valor da coluna 0,00. O valor da intersecção é de 0,3413, ou seja, 34,13%.

Entretanto, lembrando que a curva normal é *simétrica*, sabe-se que a área sob a curva normal contida entre *z* = 0 e *z* = -1 *também* é 34,13%. Portanto, a área referente a -1 < *z* < 1 vale a soma de ambas, ou seja, *68,26%*.

Recordando que o valor central corresponde a μ, pode-se traçar o seguinte gráfico, em que se percebe que, excetuando-se os valores centrais, sobram apenas 15,87% para cada lado da curva.



Exemplo 2

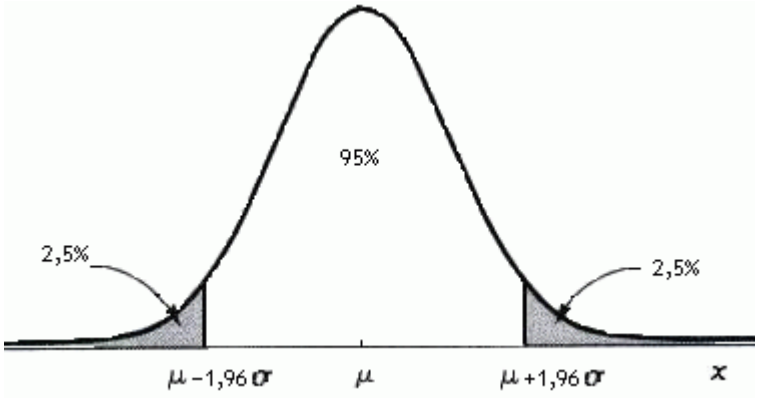
Assim sendo, considerando a *área sob a curva normal*, qual é a área correspondente a exatos 95% da curva?

z = 95% = 0, 95

0, 95 / 2 = 0,4750

Procurando esse valor (0,4750) na tabela de *z* chega-se a *1,96*.

Portanto, como o valor da área é o mesmo valor da probabilidade, se uma variável x tem distribuição normal, com média μ e desvio padrão σ a probabilidade de se sortear da população de valores de x um valor contido no intervalo μ ± 1,96 σ é igual a 95% (47,5% para cada lado da curva) e a probabilidade de se sortear da população de valores de x um valor *não* contido no intervalo μ ± 1,96 σ é igual a 5% (2,5% em cada extremo da curva).



(em que Média da população = μ e Desvio padrão da população = σ).

Resumo: Características da curva normal

a. O campo de variação é menos infinito < *x* < mais infinito

b. A distribuição normal de x é completamente determinada por dois parâmetros:

- Média da população = μ

- Desvio padrão da população = σ

c. A distribuição é simétrica em relação à média e os valores de média, moda e mediana são iguais. A área total sob a curva é igual a 1, ou 100%, com exatos 50% dos valores distribuídos à esquerda da média e 50% à sua direita

d. A área sob a curva normal contida

entre	*é igual a*
μ ± 1 σ	68,26% (1)
μ ± 2 σ	95,44% (2)
μ ± 3 σ	99,74% (3)

Exercícios - Exemplos do uso de z

1. Já foi visto como se chegou ao valor 68,26%. Como se chegou aos valores (2) 95,44% e (3) 99,74%?

Tente resolver!

Para ver uma resolução clique aqui.

2. Em uma população de indivíduos adultos de sexo masculino, cuja estatura média é 1,70 m e

desvio padrão é 0,08 m, qual é o intervalo de alturas em que 95% da população está compreendido?

Tente resolver!

Para ver uma resolução clique aqui.

3. Na mesma população, qual a probabilidade de um indivíduo apresentar estatura entre 1,60 e 1,82 m?

Tente resolver!

Para ver uma resolução clique aqui.

4. Qual a probabilidade de se encontrar 1 indivíduo com estatura menor que 1,58 m?

Tente resolver!

Para ver uma resolução clique aqui.

5. Sabendo-se que o índice de massa corpórea em uma população de pacientes com *diabetes mellitus* obedece uma distribuição normal e tem média = 27 kg/cm^2 e desvio-padrão = 3 kg/cm^2, qual a probabilidade de um indivíduo sorteado nessa população apresentar um índice de massa corpórea entre 26 kg/cm^2 e a μ?

Tente resolver!

Para ver uma resolução clique aqui.

6. Em mulheres, a quantidade de hemoglobina por 100 ml de sangue é uma variável aleatória com distribuição normal de média $\bar{x}$ = 16g e desvio padrão s = 1g. Calcular a probabilidade de uma mulher apresentar 16 a 18 g por 100 ml de hemoglobina no sangue.

Tente resolver!

(Resoluções acima em http://www.cultura.ufpa.br/dicas/biome/bionor3.htm)

Erro padrão da média e tamanho amostral

Se for retirado um certo número de amostras aleatórias de mesmo tamanho de uma população, não se deve esperar que todas as médias e desvios padrões amostrais sejam iguais. Encontra-se uma distribuição das médias amostrais.

População: Média = μ Desvio padrão = σ			
Amostra 1	Amostra 2	Amostra 3	Amostra 4
$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_4$
s_1	s_2	s_3	s_4

Intuitivamente percebe-se que o centro desta distribuição está próximo da média real da população.

Exemplo: Supondo as seguintes frequências cardíacas em 5 amostras, cada qual com 3 indivíduos, de uma população:

Amostra	1	2	3	4	5
Dados	68, 68, 71	68, 70, 72	67, 70, 73	67, 69, 69	68, 69, 70
Média ($\underline{x}_a$)	69,00	70,00	70,00	68,33	69,00

A média das médias é igual a:

= (69,33 + 70,00 + 70,00 + 68,33 + 69,00) / 5 = 69,27

Depois, calcula-se uma medida da dispersão das cinco médias amostrais: o desvio padrão das médias.

$$\text{Desvio padrão} = \sqrt{\Sigma (\underline{x}_a - \underline{x}) / (n-1)}$$

Ressalte-se que, nesse caso:

$\underline{x}_a$ = cada média amostral, $\underline{x}$ = média das amostras (69,27) e n = número de amostras.

Substituindo os valores na equação:

Desvio padrão = raiz[$(69,00 - 69,27)^2 + 70,00 - 69,27)^2 + \ldots + (69,00 - 69,27)^2$] / 4 = 0,71

Notar que nenhuma das médias equivale ao valor encontrado. Assim, *sempre se comete erro* ao se calcular a média.

O procedimento descrito acima é um método empírico para definição do *erro padrão* da Média (EPM).

Matematicamente é possível calcular esse erro. O erro da média ou erro padrão da amostra ou, simplesmente erro padrão (s_x ou EPM) é dado por:

$$s_x = \sigma / \text{raiz } n \quad \text{ou} \quad s_x = s / \text{raiz } n$$

em que:

s = Desvio padrão da amostra (o desvio padrão da população não é conhecido)

σ = Desvio padrão da população

n = Tamanho da amostra

Conclui-se que:

- Existe uma relação *inversa* entre o *tamanho da amostra* e o *erro padrão*, ou seja, quando o tamanho da amostra aumenta o erro padrão diminui.
- O erro padrão da média diminui com a raiz quadrada do número n de medições realizadas. Portanto, realizar mais medidas melhora a determinação do valor médio como estimador da grandeza que se deseja conhecer.

Erro padrão só com 1 amostra

Nesse caso, os parâmetros da população (desvio padrão e média) são conhecidos.

$$z = (\bar{x} - \sigma) / \text{EPM} \text{ ou seja, } z = (\bar{x} - \sigma) / s_x$$

Exemplo:

Exemplo: Um médico receitou um medicamento vasodilatador (Nifedipina) para Hipertensão Arterial, mas ele suspeita que o medicamento está aumentando a frequência cardíaca dos pacientes. Sabedor que a população apresenta os seguintes valores: μ = 69,8 , σ = 1,86, coletou uma amostra aleatória de 50 pacientes e mediu as suas frequências cardíacas, obtendo a média de 70,5. Ele estava correto?

Estabelece-se as hipóteses, com α = 5%

$\bar{x} - \mu = 0$ $\quad H_0: \bar{x} \leq \mu$)

$\bar{x} - \mu \neq 0$ $\quad H_a: \bar{x} \geq \mu$)

Calcula-se o erro da média:

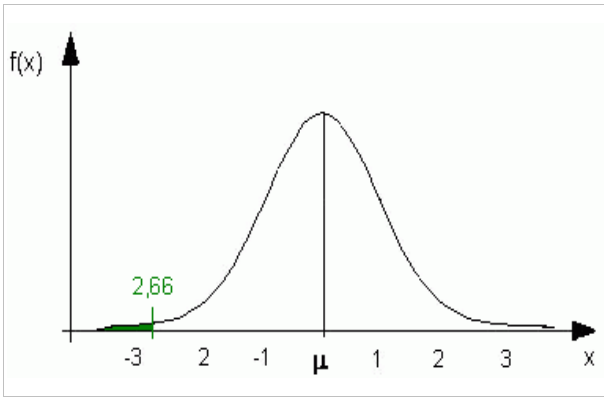
$s_x = \mu / \text{raiz } n = 1,86 / \text{raiz } 50 = 1,86 / 7,0710 = 0,2630$

Calcula-se z

$z = (\bar{x} - \mu) / s_x = (69,8 - 70,5) / 0,2630 = -0,7 / 0,2630 = -2,66$

Consultando o valor -2,66 na na Tabela de z obtém-se o valor 0,4961. Portanto:

z = 0,50 - 0,4961 = -0,0039 = 0,39%



Ou seja, existe uma probabilidade de aproximadamente 0,0039 (0,39%) de que seja obtida uma média *maior do que 70,5* ao acaso, quando são retiradas amostras aleatórias de tamanho 50 desta população.

Como essa probabilidade é menor que 5% ($p < 0,05$), rejeita-se H_0 e aceita-se H_1, concluindo-se que a suspeita do médico se confirmou e a nifedipina aumentou significativamente a frequência cardíaca.

Distribuição de t de Student

Em 1908, o estatístico inglês William Sealey Gosset, que assinava os seus trabalhos com o pseudônimo de "Student" descobriu essa distribuição. Mas seus trabalhos foram ignorados e redescobertos por Fisher só em 1924-25, apesar de terem enorme importância estatística.

O valor de *t* é a medida do desvio entre a média amostral $\bar{x}$, estimada a partir de uma amostra aleatória de tamanho n, e a média μ da população, usando o erro da média como unidade de medida:

$$t = (\bar{x} - \mu) / s_x$$

O parâmetro usado para descrever a distribuição t é o *número de graus de liberdade* que terá relação com o tamanho da amostra (n) .

Os dados sobre *t* também já se encontram tabelados. (Para ver a *tabela de t*, clique aqui).

A tabela é lida como a de Qui quadrado, ou seja, probabilidade (P) nas colunas e Graus de liberdade (G.L.) nas linhas, sendo o valor de t_c (t crítico) encontrado na *intersecção* entre a coluna de 5% e a linha correspondente ao número de graus de liberdade da amostra, sendo G.L. = n – 1.

Do mesmo modo que a tabela de *z*, a tabela de *t* é especular, ou seja, para os valores negativos de *t* existe esse mesmo conjunto de valores, mas com sinal negativo. Ou seja, a tabela de *t* é *bicaudal*.

Intervalo de confiança da média e limites fiduciais

Uma das aplicações importantes do conhecimento da distribuição de *t* é a possibilidade de, conhecendo-se

- a média amostral de uma variável x e

- o erro da média = $s_{\underline{x}} = s / \sqrt{n}$

poder estimar quais valores x poderá assumir dentro de um intervalo em torno da média μ.

Esse intervalo é denominado "Intervalo de confiança da média μ" e os valores que o delimitam

são os "limites fiduciais" ou "limites de confiança da média".

Supondo uma variável x, com distribuição normal, cuja média populacional μ não conhecemos e que, numa amostra casual de tamanho n, já se calculou x médio ($\bar{x}$) e o erro da média (s_x).

Se quisermos estabelecer o intervalo de confiança da média μ, com probabilidade de 95%, devemos verificar primeiramente, em uma tabela de t, qual é o valor de t, com n-1 graus de liberdade e 5% de probabilidade. Esse valor é chamado de **t crítico** ($\mathbf{t_c}$).

É importante lembrar que o valor de *t* amostral $t = (\bar{x} - \mu) / s_x$ deve estar no intervalo entre $- t_c$ e $+ t_c$ em 95% das amostras.

Portanto, pode-se dizer que existe uma probabilidade de 95% de encontrar:

$$- t_c \leq (\bar{x} - \mu) / s_x \leq + t_c$$

Se multiplicarmos todos os termos da expressão por s_x :

$$- t_c s_x \leq (\bar{x} - \mu) \leq + t_c s_x$$

Se transpusermos $\bar{x}$:

$$- \bar{x} - (t_c s_x) \leq \mu \leq - \bar{x} + (t_c s_x)$$

Mudando os sinais:

$$\bar{x} + (t_c s_x) \geq \mu \geq \bar{x} - (t_c s_x)$$

Invertendo os termos:

$$\bar{x} - (t_c s_x) \leq \mu \leq \bar{x} + (t_c sx)$$

Essa última expressão indica que antes de tomar uma amostra para estudo existe uma possibilidade de 95% do intervalo $\bar{x} \pm (t_c s_x)$ conter a média μ.

Exemplo:

1. Foi tomada a distância interpupilar de 131 mulheres adultas e obteve-se $\bar{x}$ = 59,2 mm e s = 2,75mm

$s_x = s / \sqrt{n} = 2,75 / \sqrt{131} = 0,2402$ mm

Para estimar o intervalo de confiança de 95% da média da distribuição da distância interpupilar nessa amostra, consulta-se a tabela de t com com n-1 graus de liberdade (131 - 1 = 130) e 5% de probabilidade.

Como 130 >120 (último valor na coluna1) pode-se ler o valor de t crítico na linha de infinito (∞) e na coluna de 0,05.

O t encontrado é 1,96. Calcula-se, então:

$$\bar{x} - (t_c s_x) \leq \mu \leq \bar{x} + (t_c s_x)$$

$$59,2 - (1,96 \times 0,24) \leq \mu \leq 59,2 + (1,96 \times 0,24), \text{ obtendo-se:}$$

$$58,73\text{mm} \leq \mu \leq 59,67\text{mm}$$

ou seja, a *média populacional*, calculada a partir de uma única amostra, deve estar entre os limites fiduciais *58,73* e *59,67* mm, um espaço menor que 1 mm (0,94 mm)

2. Suponha que os dados são os mesmos, exceto o tamanho amostral.

a. Qual seria o intervalo fiducial se *n* fosse 231? b. 61? c. 31? d. 21? e. 11? f. 6?

n =	**231**	**131**	**61**	**31**	**21**	**11**	**6**
média =	59,2	59,2	59,2	59,2	59,2	59,2	59,2
s =	2,75	2,75	2,75	2,75	2,75	2,75	2,75
tc =	1,960	1,960	2,000	2,042	2,086	2,228	2,571
s_x = s / raiz n	0,1809	0,2403	0,3521	0,4939	0,6001	0,8292	1,1227
tc.s_x =	0,3546	0,4709	0,7042	1,0086	1,2518	1,8474	2,8864
$\bar{x}$ -(tc.s_x) =	58,85	58,73	58,50	58,19	57,95	57,35	56,31
$\bar{x}$ -(tc.s_x) =	59,55	59,67	59,90	60,21	60,45	61,05	62,09
intervalo fiducial	0,71	0,94	1,41	2,02	2,50	3,69	5,77

Conclui-se que conforme o tamanho amostral *diminui* os limites fiduciais estão cada vez mais *distantes*. Assim, com amostras pequenas não se chega a uma boa ideia sobre o valor da média populacional.

Distribuições binomial e normal

Os dados biológicos muitas vezes apresentam-se graficamente como curvas com distribuição normal ou binomial.

É importante notar que a distribuição binomial se aproxima da distribuição normal à medida que o número de experimentos aumenta. E deve-se notar que curvas que obedecem binomiais, especialmente após GL = 30, são extremamente semelhantes às normais.

Assim, quando uma amostra tem n > 30 uma curva binomial tende a se assemelhar a uma curva normal. No caso de n = 31 a distribuição $(p + q)^{31}$ terá os seguintes valores:

Se *p* = *q* = 0,5	Se *p* = 0,75 e *q* = 0,25
μ = 15,5 e s = 2,78	μ = 7,75 e s = 2,41
95% da distribuição está entre 10,05 e 20,95	*95%* da distribuição está entre 3,02 e 12,47

Quando uma amostra tem n > 30, uma das consequências da aproximação da curva binomial à normal é que a média e o desvio padrão da distribuição binomial podem ser usados para por à prova:

H. Nula: a proporção observada (**o**) de 1 entre 2 acontecimentos alternativos não se desvia significativamente da proporção teórica esperada (μ). H_0: **o** = μ

H. Alternativa: o desvia-se significativamente de μ . H_a: **o** ≠ μ

Nesse caso, z = (**o** - μ) / **o**

O valor de z é comparado com o valor de tc:

Z < - t_c rejeita-se a hipótese nula **o** ≠ μ	- t_c:	- t_c < z < + t_c: aceita-se a hipótese nula **o = μ**	+ t_c	z > + t_c: rejeita-se a hipótese nula. **o** ≠ μ

Amostras com n > 30

Exemplo 1.

Um ortopedista ao estudar 52 filhos de casais que incluem 1 cônjuge com uma anomalia óssea verificou que 20 dos filhos apresentam a mesma anomalia encontrada em 1 de seus pais.

Hipótese H_0: é uma herança dominante, autossômica e monogênica, ou seja, p = q = 0,5

O número esperado de anômalos é

μ = nq, μ = 52 x 0,5 = 26

O desvio padrão é s = √ n p q = √ 52 x 0,5 x 05 = 3,606

O número observado de anômalos é = 20

z = (20 - 26) / 3,606 = -1,664

gl = 52 -1 = 51, t_c = 2,00

Lembrando do critério:

$-t_c$ $+t_c$

_____ -2,00 __|_____ 0 _____+2,00_____

-1,664

Como $-t_c < z < +t_c$ pode-se *aceitar* H_0.

Exemplo 2.

E *se* o ortopedista tivesse encontrado não 20, mas apenas 17 filhos com a mesma anomalia dos pais?

z = (17 - 26) / 3,606 = -2,496

Se apenas 17 filhos fossem anômalos, como $z > t_c$ poder-se-ia *rejeitar* H_0.

$-t_c$ $+t_c$

_____*!*_____ -2,00 __|_____ 0 _____+2,00_____

-2,496

Se apenas 17 filhos fossem anômalos, como $z < t_c$ poder-se-ia *rejeitar* H_0.

Amostras com n < 30

Mesmo em amostras com *n bem menor que 30 indivíduos* pode-se usar métodos aplicáveis à distribuição normal.

Exemplo 1:

Considerando uma certa anomalia que tem probabilidade de 0,5 de se manifestar em filhos de casais que incluem 1 cônjuge afetado. Analisando irmandades de diferentes tamanhos geradas por esses casais, qual a probabilidade de encontrarmos pelo menos 7 anômalos nas irmandades com 12 irmãos?

Resolução 1

- Usando o Triângulo de Pascal

Para se determinar os coeficientes da equação, monta-se o Triângulo até atingir o expoente desejado no binômio de Newton:

1	0
1 1	1
1 2 1	2
1 3 3 1	3
1 4 6 4 1	4
1 5 10 10 5 1	5
1 6 15 20 15 6 1	6
1 7 21 35 35 21 7 1	7
1 8 28 56 70 56 28 8 1	8
1 9 36 84 126 126 84 36 9 1	9
1 10 45 120 210 252 210 120 45 10 1	10
1 11 55 165 330 462 462 330 165 55 11 1	11
1 12 66 220 495 792 924 792 495 220 66 12 1	12

Portanto, a equação será:

$1p^{12}\ q^0 + 12\ p^{11}\ q^1 + 66p^{10}q^2 + 220p^9\ q^3 + 495p^8\ q^4 + 792p^7\ q^5 + 924p^6\ q^6 + 792p^5\ q^7 + 495\ p^4\ q^8 + 220p^3q^9 + 66p^2\ q^{10} + 12p^1\ q^{11} + 1p^0\ q^{12}$

Sendo p = normalidade e q = anomalia, como o problema pede "pelo menos 7 anômalos nas irmandades com 12 irmãos" nos interessa apenas essa parte da equação:

$792p^5\ q^7 + 495\ p^4\ q^8 + 220p^3q^9 + 66p^2\ q^{10} + 12p^1\ q^{11} + 1p^0\ q^{12}$

Somando-se seus coeficientes (792 + 495 + 220 + 66 + 12 + 1= 1586), temos 1586 indivíduos para 4096 no total das irmandades.

1586 / 4096 = 0,3872, portanto, P = 38,7%

Ou seja, a probabilidade de se encontrar "pelo menos 7 anômalos nas irmandades com 12 irmãos" é igual a 38,72%.

Resolução 2

- Usando *as características da curva normal*

$\mu = nq = 12 . 0,5 = 6$

s = *raiz* npq = *raiz* 12 . 0,5 . 0,5 = 1,73

O valor da média pode ser tomado como um centro de classe do intervalo 5,5 a 6,5.

$z = (x - \mu) / (\sigma) = (6,5 - 6) / 1,73 = 0,29$

Consultando a tabela de *z*, vê-se que o valor correspondente a 0,29 é 0,1141, o que indica que

a área ocupada a partir de 5,5 é 0,5000 - 0,1141 = 0,3869, ou seja, que tem uma probabilidade de 38,7%

Portanto, nota-se que apesar de estarmos tratando de outra distribuição (binomial) as fórmulas referentes à distribuição normal podem ser usadas pois a diferença encontrada nos resultados é insignificante, (38,72% e 38,69%) é insignificante, praticamente desprezível.

Exemplo 2:

Qual a probabilidade de encontrarmos irmandades com 4 indivíduos normais e 8 anômalos?

Resolução 1

- Usando o Triângulo de Pascal

Verificar no triângulo montado.

O valor desejado é 495 p^4q^8. Substituindo p e q por 0,5:

495 $0,5^4$ $0,5^8$ = 0,121 ou 12,1%

Resolução 2

- Usando as características da curva normal

A área sob a curva na classe correspondente a 8 (com limites 7,5 e 8,5) deve ser calculada

Lembrando que z = (*x* - μ) / σ, calcula-se:

z_1 = (limite min - μ) /σ e z_2 = (limite max - μ) /σ

z_1 = 7,5 - 6 / 1,73 = 0,87 e z_2 = 8,5 - 6 / 1,73 = 1,45

Verificando na Tabela de z:

0,87 corresponde a 0,3078 e 1,45 corresponde a 0,4265

A diferença entre essas áreas dá a a área limitada por 0,87 e 1,45, ou seja,

0,4265 - 0,3078 = 0,1187 = 0,119

0,119 = aproximadamente 12%

Novamente percebe-se que apesar de ser um caso de distribuição (binomial) as fórmulas referentes à distribuição normal podem ser usadas pois a diferença encontrada nos resultados é insignificante, praticamente desprezível.

Tamanho da amostra

Em uma amostragem não probabilística, o tamanho amostral é estabelecido sem nenhuma base de sustentação técnica. Comumente corresponde a 10% ou 15% da população alvo.

Já, em uma amostragem probabilística, o tamanho da amostra é função:

- do(s) parâmetro(s) a estimar,
- do nível de confiança desejável,
- do erro tolerável *ou* índice de precisão escolhidos,
- do grau de dispersão da população,
- pode, ainda, depender do tamanho da população e de outros parâmetros específicos.

Basicamente, o tamanho da amostra depende da precisão desejada, conforme o arbítrio do pesquisador. Assim, é intuitivo perceber que o tamanho depende do erro aleatório mencionado acima.

Há uma relação inversa entre o erro e o tamanho da amostra. Amostras “grandes” estão associadas a erros “pequenos” e amostras “pequenas” a erros “grandes”. Assim, deve-se procurar uma compatibilidade entre o tamanho amostral e o erro que se “tolera” cometer em um estudo.

Se soubermos o valor do desvio padrão da variável que está sendo estudada podemos ter uma ideia de qual deve ser um bom tamanho amostral, pois

O *erro tolerável* (E) é :

Erro da média $= \sigma_{\underline{x}} = \sigma / \sqrt{n}$, com intervalo de confiança $\underline{x} \pm 1{,}96\ s_x$

em que n = tamanho amostral.

O *erro tolerável* (E) é :

$$E = 1{,}96\sigma / \sqrt{n}$$

Elevando ao quadrado, obtém-se:

$$E^2 = 1{,}96^2\ \sigma^2 / n$$

o que permite escrever:

$$n = 1{,}96^2\sigma^2 / E^2$$

Exemplo 1:

Foi feita uma dosagem bioquímica de um certo composto em uma amostra de 36 indivíduos e obteve-se $\bar{x}$ = 300 mg e s = 15 mg. Qual é um bom tamanho para essa amostra (n)?

Aceitando-se que s é um bom estimador para σ

σ = 15 mg e $\sigma_{\underline{x}} = \sigma / \sqrt{n} = 15 / \sqrt{36} = 2{,}5$ mg

E = 1,96 σ = 1,96 x 2,5 = 4,9 mg = precisão da estimativa

Ou seja, a média tem 95% de chance de estar entre 300 $\pm$ 4,96,(entre 295,1 e 304,9 mg).

Entretanto, se o pesquisador quiser *aumentar essa precisão* de modo que o intervalo de confiança da média fique entre 298 e 302, E será igual a 2.

Então:

$n = 1{,}96^2\sigma^2 / E^2 = 1{,}96^2\ 15^2 / 2^2 = 216{,}09 = 216$ indivíduos

Como já há 36 pessoas na amostra, faltam 216 - 36 = 180

Assim, para conseguir que o erro passe de 4,9 para 2 o pesquisador precisaria de mais 180 indivíduos.

Obs. Se a distribuição da amostra for *binomial* (e não normal) deve-se usar essas fórmulas:

$E = 1{,}96 \sqrt{pq / n}$ e $n = 1{,}96^2\, pq / E^2$

Momentos, Assimetria e Curtose

Momentos

1o. momento	2o. momento	3o. momento	4o. momento
r = 1	r = 2	r = 3	r = 4
$\sum x / n$	$\sum x^2 / n$	$\sum x^3 / n$	$\Sigma x^4 / n$

Momentos centrados na média

1o. momento centrado na média	2o. momento centrado na média	3o. momento centrado na média	4o. momento centrado na média
m_1	m_2	m_3	m_4
$\sum (x - \bar{x}) / n$	$\sum (x - \bar{x})^2 / n$	$\sum (x - \bar{x})^3 / n$	$\sum(x - \bar{x})^4 / n$

Em relação ao primeiro momento, sabe-se que é nulo, pois, $\sum (x - \bar{x}) / n = 0$

O segundo momento $\sum (x - \bar{x})^2 / n$ é muito parecido com a variância $\sum (x - \bar{x})^3 / n$. O desenvolvimento dessas fórmulas permite, usando os dados individuais, chegar em:

$m_2 = \sum x^2 / n - \bar{x}^2$

$m_3 = \sum x^3 / n - (3 \bar{x} \sum x^2)/ n + 2 \bar{x}$

$m_4 = \sum x^4 / n - (4 \bar{x} \sum x^3)/ n + (6 \bar{x}^2 \sum x^2) / n – 3 \bar{x}^4$

Fórmulas para dados agrupados em classes

$\bar{x}$ = média

i = intervalo de classe

X = centros de classe

f = frequência absoluta

n = tamanho da amostra, chega-se a essas fórmulas:

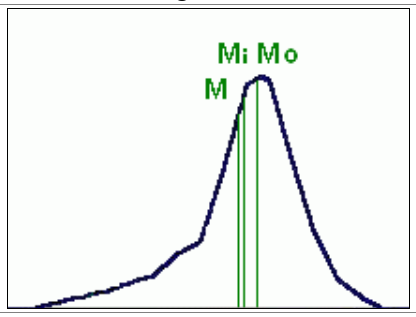
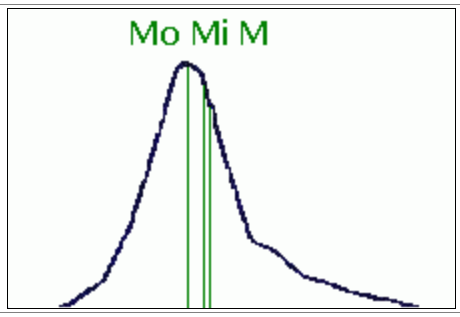
$m_2 = \{\sum fX^2 / n - fX^2] \} i^2$

$m_3 = \{\sum fX^3 / n - (3 fX \sum fX^2)/ n + [2 (\sum (fX)^3 / n^3] \} i^3$

$m_4 = \{\sum fX^4 / n - (4 fX \sum fX^3)/ n^2 + [6 (\sum fX)^2 \sum fX^2) / n^3 - [3 (\sum fX)^4 / n^4] \} i^4$

Simetria

O terceiro momento centrado na média é utilizado na investigação de *assimetria* nas distribuições. Nas distribuições unimodais essa investigação é muito interessante pois é necessário saber se existe assimetria positiva ou negativa, ou seja, se é significativo o alongamento de uma das caudas da distribuição (à direita ou à esquerda da média).

Assimetria Negativa: M < Mi < Mo	*Assimetria Positiva*: Mo > Mi > M
	

Para estudar a *assimetria* em distribuições unimodais Fisher propôs o coeficiente g_1

Coeficiente | **quantidade k**
$g_1 = k_3 / s^3$ | $k_3 = m_3\ n^2 / (n-1).(n-2)$

sendo que:

erro do coeficiente | **teste t**
$s_{g1} = \sqrt{}\ [(6n\ (n-1) / (n-2)\ (n+1)(n+3)]$ | $t = g_1 / s_{g1}$

Fórmulas simplificadas, usadas para amostras com grande tamanho

coeficiente | **quantidade k** | **erro do coeficiente**
$g_1 = m_3 / m_2 \sqrt{}\ m_2$ | $k_3 \sim m_3$ | $s_{g1} = \sqrt{}\ 6/n$

Para verificar se o valor de g_1 se desvia significativamente de zero calcula-se a razão entre g_1 e s_{g1} obtendo-se um t que deve ser comparado a um t crítico (t_c) com infinitos graus de liberdade ao nível de significância de 5% (t_c = ± 1,96).

Um valor de *t* calculado igual ou maior que *+1,960* indica que g_1 é significativamente maior que zero, ou seja, que a assimetria é positiva. Do mesmo modo, um valor de *t* calculado igual ou menor que *-1,960* indica que g_1 é significativamente menor que zero, ou seja, que a assimetria é negativa.

Curtose

O quarto momento centrado na média é utilizado na investigação de *curtose* nas distribuições. Calcula-se:

coeficiente | **Quantidade k**
$g_2 = k_4 / (s^3)^2$ | $k4 = [m_4\ n^2\ (n+1) - 3(n-1)^3\ (s^2)^2] / [(n-1).(n-2)\ (n-3)]$

sendo que:

Erro do coeficiente | **teste t**
$s_{g2} = \sqrt{}\ [(24n\ (n-1)2 / (n-3)(n-2)\ (n+3)(n+5)]$ | $t = g_2 / s_{g2}$

Fórmulas simplificadas, usadas para amostras com grande tamanho

coeficiente | **Quantidade k** | **erro do coeficiente**
$g_2 = m_4 / (m_2)^2 - 3$ | $k_4 = m_4 - 3\ (m_2)^2$ | $s_{g2} = \sqrt{}\ 24/n$

O teste *t* tem t_c = ± 1,96, sendo que um valor de *t* calculado igual ou maior que *+1,960* indica que g_2 é significativamente maior que zero, ou seja, que a distribuição é leptocúrtica. Do mesmo modo, um valor de *t* calculado igual ou menor que *-1,960* indica que g_2 é significativamente menor que zero, ou seja, que a distribuição é platicúrtica.

Para facilitar os cálculos utilize uma planilha especial:

Distribuição normal - cálculo de Momentos 2, 3 e 4 em amostras grandes
Copie a planilha comprimida em formato xls ou em ods

http://www.cultura.ufpa.br/dicas/biome/biozip/momentos.zip

O coeficiente de variação C

Como já foi visto, o coeficiente de variação é uma medida da dispersão dos dados.

E é a razão entre o desvio padrão e a média amostral:

$$C = s / \bar{x}$$

Quando se transforma o desvio padrão em uma fração da média pode-se comparar amostras com desvios-padrão diferentes.

O teste t é feito, por meio da seguinte fórmula:

$$t = (C_a - C_b) / \ raiz\ (V_{Ca} + V_{Cb})$$

em que:

V_{Ca} = Variância da amostra a e V_{Cb} = Variância da amostra b

Graus de liberdade = $n_a + n_b - 4$, em que n_a e n_b são os tamanhos amostrais.

Se os coeficientes de variação forem menores que 0,30 (o que acontece quase sempre) pode-se calcular a variância do seguinte modo:

$$V_C = C^2 / 2n\ (1 + 2C^2)$$

Se os coeficientes de variação forem maiores que 0,30, calcula-se a variância assim:

$$V_C = \bar{x}\ [(m_4 - m_2^2) - 4\bar{x}\, m_2 m_3 + 4\bar{x}\, m_2^3] / 4n.\bar{x}^4$$

em que:

m_2, m_3 e m_4 = segundo, terceiro e quarto momentos centrados na média

$\bar{x}$ = média

n = tamanho da amostra

Exemplo:

Supondo 2 amostras onde foi coletada a altura de indivíduos. Ambas são constituídas por indivíduos caucasóides, de sexo masculino, de Campinas. Mas a primeira amostra recém nascidos e a segunda universitários, sendo que:

Amostra a. recém-nascidos, caucasóides, sexo masculino, de Campinas, em que:

$\bar{x}$ = 49,0; s = 2,55, n = 50

Amostra b: universitários, caucasóides, sexo masculino, de Campinas, em que:

$\bar{x}$ = 170,11 s = 8,38 n = 100

Portanto:

Amostra a*,* recém-nascidos: Ca = 2,55 / 49 = 0,052

Como o coeficiente de variação é menor que 0,30, usa-se:

$V_C = C^2 / 2n (1 + 2C^2) = (0,0522 / 2 . 50) (1 + 2 . 0,0522) = 0,000027$

Amostra b*,* universitários: Cb = 8,38 / 170,11 = 0,049

Como o coeficiente de variação é menor que 0,30, usa-se:

$V_C = C^2 / 2n (1 + 2C^2) = (0,0492 / 2 . 100) (1 + 2 . 0,0492) = 0,000012$

Teste t

$t = (C_a - C_b) / \sqrt{(V_{Ca} + V_{cb})}$

$t = (0,052 - 0,049) / \sqrt{(0,000027 + 0,000012)} = 0,500$

Graus de liberdade = 50 + 100 - 4 = 146, portanto 0,60 < P < 0,70.

Assim, os coeficientes de variação não diferem significativamente. Ou seja, apesar das amostras serem muito diferentes quanto à idade de seus indivíduos, a distribuição das alturas é semelhante em ambas.

Como desenhar uma Curva Normal?

Há uma maneira de conseguir desenhar a curva normal esperada para a população a partir dos dados amostrais.

Exemplo:

Ao estudar o nível de uma certa enzima nos hemolisados de 138 homens brasileiros adultos, jovens e sadios, verificou-se que a sua distribuição segundo a atividade dessa enzima era unimodal. Os dados amostrais a respeito dessa atividade (x 10^4) foram agrupados na tabela abaixo.

Com base nesses dados, criar um gráfico, em colunas, da distribuição observada, sob um gráfico, em linha, de sua curva normal.

min	***max***	***cen***	***f***	***min***	***max***	***cen***	***f***
18,00	22,00	20	0	58,00	62,00	60	15
22,00	26,00	24	2	62,00	66,00	64	9
26,00	30,00	28	1	66,00	70,00	68	8
30,00	34,00	32	3	70,00	74,00	72	7
34,00	38,00	36	8	74,00	78,00	76	3
38,00	42,00	40	11	78,00	82,00	80	1
42,00	46,00	44	14	82,00	86,00	84	2
46,00	50,00	48	15	86,00	90,00	88	0
50,00	54,00	52	20	90,00	94,00	92	0
54,00	58,00	56	18	94,00	98,00	96	1

Segue, abaixo, um método fácil para desenhar a curva normal:

a. Calcular a média amostral ($\bar{x}$)

b. Calcular o desvio padrão amostral (s)

c. Obter os pontos para a curva normal completando a tabela a seguir, usando uma tabela com a distribuição de Y.

d. Traçar um gráfico em colunas da distribuição

e. Sobrepor ao gráfico a curva normal

Os valores obtidos na última coluna devem ser usados para montar o gráfico.

Limites	Centro	x - $\underline{x}$	*z* = (x - $\underline{x}$) / s	y	y.n/s	100. [(yn)/s] / Σ (yn/s)
18-22						
22-26						
...						

Qual é o tipo do gráfico a ser criado?

Para facilitar os cálculos utilize uma planilha especial:

Distribuição normal - como traçar a curva normal
em amostras com até 25 classes.
Copie a planilha comprimida em formato xls ou em ods

http://www.cultura.ufpa.br/dicas/biome/biozip/distnor.zip

Acesse uma resolução clicando em http://www.cultura.ufpa.br/dicas/biome/bionor2.htm

———

Este "site", destinado prioritariamente aos alunos de Fátima Conti, pretende auxiliar quem esteja começando a se interessar por Bioestatística, computadores e programas, estando em permanente construção. Sugestões e comentários são bem vindos. Agradeço antecipadamente.

———

Endereço dessa página:

HTML: http://www.cultura.ufpa.br/dicas/biome/bionor.htm

PDF: http://www.cultura.ufpa.br/dicas/pdf/bionor.pdf

Última alteração: 4 nov 2009 (Solicito conferir datas. Pode haver atualização só em HTML)