O erro da pesquisa é de 3% - o que significa isto?

A Matemática das pesquisas eleitorais

José Paulo Carneiro & Moacyr Alvim

Introdução

Sempre que se aproxima uma eleição, os meios de comunicação passam a publicar diariamente, ou quase, pesquisas por amostragem que estimam as proporções de votos dos diversos candidatos, de acordo com as intenções dos eleitores naquele momento. Estas publicações são em geral acompanhadas da informação do *tamanho da amostra* ("foram entrevistados *x* eleitores") e de uma frase do tipo: "o erro da pesquisa é de 3%, para mais ou para menos". Nessas ocasiões, os professores de Matemática são frequentemente perguntados pelos alunos e por familiares ou amigos curiosos, sobre o significado desta frase. É o que pretendemos esclarecer.

No seu *site*, um conhecido instituto de pesquisa informa que o seu cálculo de erro amostral é feito no contexto de um "modelo de amostragem aleatório simples" (ver ao final o Apêndice 1). E os outros institutos também costumam adotar o mesmo procedimento. Por isto analisaremos este tipo de amostragem, para entender estas frases.

Amostra aleatória simples

Suponha que o universo a ser pesquisado tenha *N* unidades e que uma certa variável *X* assuma, nessas unidades, os valores $X_1, \cdots, X_N$. Deseja-se selecionar uma amostra de tamanho *n* (com $n < N$), de modo que ela seja *aleatória simples* (isto é, todas as unidades têm a mesma probabilidade 1/*N* de serem selecionadas) e *sem reposição*, isto é, nenhuma unidade pode ser selecionada mais de uma vez na mesma amostra.

O número de amostras possíveis é $k = C_N^n = \dfrac{N(N-1)\cdots(N-n+1)}{n!} = \dfrac{N!}{n!(N-n)!}$.

A média amostral

Uma vez selecionada tal amostra (usando uma urna, ou uma tabela de números aleatórios, ou outro processo válido), podemos estimar a *média* (aritmética) da variável *X*, isto é: $\overline{X} = \dfrac{X_1 + \cdots + X_N}{N}$.

Para isto, tomamos a média aritmética $\overline{x}$ desta amostra como sendo um *estimador* da média $\overline{X}$ da variável em questão.

Por exemplo, se $\{y_1,\cdots,y_n\}$ for tal amostra (onde os y_i naturalmente são alguns dos X_j, sem repetição), então $\overline{x}$, nesta amostra, assume o valor $\dfrac{y_1+\cdots+y_n}{n}$.

Vamos estudar agora a distribuição da *média amostral*, isto é, vamos ver o que podemos saber sobre como varia $\overline{x}$ ao longo de todas as amostras possíveis.

Já que todos os subconjuntos do universo com *n* elementos têm a mesma probabilidade de serem selecionados, o *valor esperado da média das amostras* ao longo de todas as *k* amostras possíveis será a média aritmética de todas as médias das amostras. Este valor é representado por $E(\overline{x})$.

Para concretizar, suponha que as amostras sejam $\{y_{11},\cdots,y_{1n}\},\cdots,\{y_{k1},\cdots,y_{kn}\}$, com médias, respectivamente: $m_1=\dfrac{y_{11}+\cdots+y_{1n}}{n},\cdots,m_k=\dfrac{y_{k1}+\cdots+y_{kn}}{n}$.

Então:

$$E(\overline{x})=\frac{m_1+\cdots+m_k}{k}=\frac{1}{k}\left(\frac{y_{11}+\cdots+y_{1n}}{n}+\cdots+\frac{y_{k1}+\cdots+y_{kn}}{n}\right)$$

$$=\frac{1}{kn}\left[(y_{11}+\cdots+y_{1n})+\cdots+(y_{k1}+\cdots+y_{kn})\right].$$

Na soma que está entre colchetes, todas as parcelas são valores de X_j. Quantas vezes aparece X_1 nesta soma? Tantas quantas sejam as amostras que contêm X_1, ou seja, C_{N-1}^{n-1}. O mesmo se passa com os outros X_j. Portanto, a soma entre colchetes é igual a:

$$C_{N-1}^{n-1}\left(X_1+\cdots+X_N\right)=C_{N-1}^{n-1}\cdot N\overline{X}\,.$$

Substituindo, levando em conta que $k=C_N^n$, ficamos com: $E(\overline{x})=\dfrac{N\overline{X}\;C_{N-1}^{n-1}}{n\,C_N^n}$.

Porém $\dfrac{C_{N-1}^{n-1}}{C_N^n}=\dfrac{(N-1)!}{(n-1)!(N-n)!}\Big/\dfrac{N!}{n!(N-n)!}=\dfrac{(N-1)!n!}{N!(n-1)!}=\dfrac{n}{N}$.

Logo: $E(\overline{x})=\overline{X}$.

Isto significa que o valor esperado do estimador $\overline{x}$ é a própria média $\overline{X}$ da variável no universo. Por isto, diz-se que este é um *estimador não tendencioso*.

O desvio padrão amostral

Com a mesma nomenclatura do parágrafo anterior, a *variância* da variável *X* é, por definição, $V(X)=\frac{\left(X_1-\overline{X}\right)^2+\cdots+\left(X_N-\overline{X}\right)^2}{N}$, ou seja, a média dos desvios quadráticos de *X* em relação a sua média. O *desvio padrão* $s(X)$ da variável *X* é a raiz quadrada da variância, isto é: $s(X)=\sqrt{V(X)}$.

Uma outra expressão útil da variância decorre do seguinte desenvolvimento:

$$V(X)=\frac{\left(X_1-\overline{X}\right)^2+\cdots+\left(X_N-\overline{X}\right)^2}{N} \qquad =\frac{X_1^2-2\overline{X}X_1+\overline{X}^2+\cdots+X_N^2-2\overline{X}X_N+\overline{X}^2}{N}$$

$$=\frac{X_1^2+\cdots+X_N^2-2\overline{X}(X_1+\cdots X_N)+N\overline{X}^2}{N}=\frac{X_1^2+\cdots+X_N^2-2\overline{X}\cdot N\overline{X}+N\overline{X}^2}{N}=$$

$$\frac{X_1^2+\cdots+X_N^2}{N}-\overline{X}^2 .$$

A fórmula $V(X)=\frac{X_1^2+\cdots+X_N^2}{N}-\overline{X}^2$ é usualmente verbalizada assim: “a variância é igual à média dos quadrados menos o quadrado da média”.

A *variância amostral*, isto é, a variância do estimador $\overline{x}$ ao longo de todas as amostras, é dada por $V(\overline{x})=\frac{\left(m_1-\overline{X}\right)^2+\cdots+\left(m_k-\overline{X}\right)^2}{k}$, e o *desvio padrão amostral* é $s(\overline{x})=\sqrt{V(\overline{x})}$. O desvio padrão amostral é a principal medida do erro amostral, como veremos. No Apêndice 2, deduz-se a seguinte importante fórmula, que fornece a variância amostral:

$$V(\overline{x})=\left(\frac{1-\frac{n}{N}}{1-\frac{1}{N}}\right)\frac{V(X)}{n}$$

Observe que o fator $f=\frac{1-\frac{n}{N}}{1-\frac{1}{N}}$ tende a 1 quando *N* tende a infinito. Portanto, para uma população infinita, teríamos $V(\overline{x})=\frac{V(X)}{n}$. Por isto, *f* é chamado *fator de correção para população finita*. Além disto, f já é muito próximo de 1 para valores grandes de *N* e valores razoáveis de *n*. Por exemplo, em uma pesquisa eleitoral, o universo é o total de eleitores, atualmente em cerca de 135 milhões. Neste caso, para uma amostra de 2 mil

eleitores, $f = 0{,}999985$, com 6 decimais. Por este motivo, para pesquisas eleitorais, adota-se simplesmente a fórmula aproximada: $V(\overline{x}) = \dfrac{V(X)}{n}$.

Segue que o desvio padrão amostral é:

$$s(\overline{x}) = \frac{s(X)}{\sqrt{n}}$$

Esta fórmula é muito importante e tem vários significados e conseqüências. Por exemplo:

1) Para um tamanho fixo de amostra, o desvio padrão amostral é diretamente proporcional ao desvio padrão (no universo) da variável a ser pesquisada. Por exemplo, se a variável *A* é 2 vezes mais dispersa (em termo de desvio padrão) do que a variável *B*, então o desvio padrão amostral da variável *A* será o dobro do desvio padrão amostral da variável *B*.
2) Para uma mesma variável (portanto $s(X)$ está fixo), o erro amostral é inversamente proporcional à raiz quadrada do tamanho da amostra *n*. Por exemplo, se quadruplicarmos o tamanho da amostra, o erro se reduz à metade (e não à quarta parte, como se poderia pensar). Isto mostra que aumentar demais o tamanho da amostra não necessariamente melhora tanto a precisão da estimativa.

No entanto, cabe perguntar: como calcular o erro amostral por esta fórmula, se ele depende do desvio padrão da variável no universo, o qual é desconhecido? Há diversas maneiras de tentar contornar este problema, sempre tentando usar algum conhecimento sobre o universo.

Amostragem de proporções

Nas pesquisas eleitorais, queremos saber, por exemplo, a proporção dos eleitores que têm intenção de votar num determinado candidato. Vamos ver que isto se reduz a estimar uma média. Quando queremos estimar qual a *proporção* de uma população de tamanho *N*, que possui uma certa característica, criamos uma variável *X*, que vale 1 quando o indivíduo tem esta característica, e vale 0, em caso contrário. Neste caso, a soma $X_1 + \cdots + X_N$ traduz o número de pessoas que possuem a característica, enquanto a média $\overline{X} = \dfrac{X_1 + \cdots + X_N}{N} = P$ é justamente a proporção (a ser estimada) de pessoas que possuem a característica em questão. Já que *P* é a média da variável *X*, podemos aplicar o que aprendemos nos parágrafos anteriores sobre médias. Em uma amostra aleatória simples sem reposição, um estimador para *P* é a proporção *p* de pessoas da amostra que declaram seu voto em *A* (isto é, *p* é aqui o nosso $\overline{x}$).

Para estimar o erro amostral, vamos primeiro calcular a variância (no universo) de *X*, que é: $V(X) = \dfrac{X_1^2 + \cdots + X_N^2}{N} - \overline{X}^2$. Já sabemos que $\overline{X} = P$. Por outro lado, como *X* só

assume os valores 0 e 1, então $X_j^2 = X_j$, para cada j de 1 a N. Portanto: $V(X) = \frac{X_1 + \cdots + X_N}{N} - \overline{X}^2 = \overline{X} - \overline{X}^2 = P - P^2 = P(1-P)$.

Finalmente, aplicando a fórmula $V(\overline{x}) = \frac{V(X)}{n}$ (para tamanhos grandes de universo), vem que $V(p) = \frac{P(1-P)}{n}$.

Logo, o desvio padrão amostral para proporções é:

$$s(p) = \frac{\sqrt{P(1-P)}}{\sqrt{n}} \ .$$

Por exemplo, para estimar uma proporção de 40% (no universo) com uma amostra aleatória simples de 1.000 pessoas, o desvio padrão amostral é de $\frac{\sqrt{0,4 \cdot 0,6}}{\sqrt{1000}} \approx 0,0155 = 1,55\%$.

Desvio padrão máximo para proporções

A expressão $P(1-P) = P - P^2$ é uma forma quadrática.

Exercício: Mostre que o valor máximo que $P(1-P)$ pode assumir é $1/4$, o que ocorre quando $P = 1/2 = 0,5 = 50\%$.

Conseqüência: Tomando a raiz quadrada, conclui-se que o desvio padrão amostral máximo das proporções é $\frac{\sqrt{1/4}}{\sqrt{n}} = \frac{1}{2\sqrt{n}}$, o qual ocorre para a proporção de 50%.

Os institutos de pesquisa, em geral, fornecem a sua informação de erro amostral, tendo em vista o erro máximo (veja, novamente, o Apêndice 1)

O papel da curva normal

Como foi sugerido pelo experimento inicial do curso, numa amostra aleatória simples, desde que o tamanho do universo seja suficientemente "grande" (um conceito relativo em Matemática), a distribuição das médias de todas as possíveis amostras é aproximadamente igual à de uma curva normal, com média e desvio padrão iguais, respectivamente, à média e ao desvio padrão amostrais.

Por outro lado, é sabido (da teoria da curva normal) que, se uma variável aleatória for distribuída segundo uma distribuição normal de média *m* e desvio padrão *s*, então a probabilidade de que esta variável assuma valores entre $m-s$ e $m+s$ é de aproximadamente 68%, e a probabilidade de que esta variável assuma valores entre $m-2s$ e $m+2s$ é de aproximadamente 96%. Também muito usado é o intervalo entre

$m-1,96s$ e $m+1,96s$, que cobre aproximadamente 95%. Sobre as propriedades da curva normal, ver Apêndice 3.

Exemplo aplicado às pesquisas eleitorais

Suponha que um Instituto de Pesquisa tenha realizado uma amostragem aleatória simples de âmbito nacional para estimar proporções de intenção de votos, com uma amostra de 2.000 eleitores. Então, o desvio padrão amostral máximo é $s = \frac{1}{2\sqrt{1500}} \approx 0,013 = 1,3\%$. Como $2s = 2,6\%$, então o Instituto poderá dizer que "o erro da pesquisa é de 2,6%".

Com isto, confiando no caráter normal da distribuição amostral, ele espera garantir que somente em 4% de todas as amostras possíveis, uma proporção (no universo) de 50% poderia aparecer na amostra como mais de 52,3% ou menos do que 47,7%.

Uma informação mais detalhada seria uma tabela do tipo:

Proporção	**Erro amostral**
(%)	**(%)**
10	1,5
20	2,1
30	2,4
40	2,5
50	2,6

onde os valores da segunda coluna correspondem a $2s = \sqrt{\frac{P(1-P)}{n}}$.

Note que os valores da 1ª coluna referem-se ao universo.

Comentário final sobre as pesquisas eleitorais

Na prática, é inviável economicamente fazer uma pesquisa eleitoral de âmbito nacional (e mesmo estadual ou municipal, para municípios grandes) usando amostra aleatória simples. O que se faz comumente é selecionar a amostra em dois estágios, selecionando primeiro uma amostra de municípios (são cerca de 5.700 no Brasil). Nesta amostra, os municípios não são selecionados com igual probabilidade, e sim com probabilidade proporcional à sua população. Dentro de cada município selecionado, a idéia é fazer

uma *amostragem estratificada*, isto é o universo é dividido em *estratos* supostamente homogêneos em relação à variável pesquisada. Este procedimento tende a reduzir o desvio padrão amostral. No caso das pesquisas eleitorais, a estratificação é feita por renda, gerando os estratos denominados "classe A", "classe B", etc. Uma maneira de fazer isto é usar informações, por exemplo, do último Censo Demográfico do IBGE. Uma maneira muito mais barata, mas bem menos precisa, é a chamada "amostragem por quotas". Nesta, o instituto determina previamente quantos eleitores vão ser pesquisados em cada estrato e sai "caçando" os eleitores nas ruas, coletando sua intenção de votos e também a sua informação de renda. A partir daí, completa as suas "quotas". Neste último sistema, é praticamente impossível calcular o erro amostral. Uma amostra por estágios estratificada, se for bem feita, permite o cálculo do erro amostral, mas este seria bastante complexo. Como se viu, na prática, os institutos de pesquisa, para efeito de erro amostral, fazem de conta que a amostra é aleatória simples.

Ilustração prática

Para ilustrar praticamente estes conceitos durante o curso, foi proposto primeiramente estimar a altura média dos participantes do curso, que eram 12. Além de calcular, numa planilha eletrônica, a média e o desvio padrão do universo, o tamanho pequeno do universo permitiu observar todas as amostras, a média amostral e o desvio padrão amostral. Na oportunidade, foi verificada a veracidade das fórmulas deduzidas. Foram feitas também experiências fictícias com universos maiores. Foi explorado o fato de que o aspecto dos histogramas se aproximava do aspecto de uma curva normal (ver adiante). Também foi feito um experimento com proporções (ver o parágrafo seguinte). Tudo isto consta da planilha anexa, denominada *Experimentos Amostrais*.

Apêndice 1

Informação dada no site do IBOPE - Acesso em 24/03/2011

http://www.ibope.com.br/calandraWeb/BDarquivos/sobre_pesquisas/pesquisa_eleitoral.html

Margem de erro

Por se tratar de estatísticas e não números absolutos, toda pesquisa apresenta uma margem de erro que depende do **tamanho da amostra** estudada e dos resultados obtidos. Isso ocorre porque não é entrevistado todo o universo da população, mas apenas uma parte representativa deste. Trabalhando dessa maneira, há sempre um **erro amostral** conhecido e calculado especificamente para cada pesquisa eleitoral.

Para uma mesma amostra, quanto maior a **homogeneidade da população** pesquisada, menor será o erro amostral e vice-versa. Por isso, não existe um erro amostral único e fechado para a pesquisa como um todo, pois em cada informação fornecida pela pesquisa há um erro correspondente.

No caso das pesquisas eleitorais, esses erros são geralmente desiguais para os diversos candidatos em função da **distribuição geográfica** do eleitorado de cada um deles. A margem de erro comumente divulgada refere-se a uma **estimativa de erro máximo**, considerando-se um **modelo de amostragem** aleatório simples. Dessa maneira, os resultados de uma pesquisa devem ser interpretados dentro de um intervalo que estabeleça limites à estimativa obtida: o chamado **intervalo de confiança**.

O intervalo de confiança é sempre pré-estabelecido antes do início da pesquisa, de comum acordo entre o cliente e o IBOPE. Geralmente, fica em torno de 95%. Isso quer dizer que se uma pesquisa fosse realizada 100 vezes em 95 delas o resultado ficaria dentro da margem de erro.

Apêndice 2

Dedução da fórmula da variância amostral

A variância amostral, isto é, a variância do estimador $\overline{x}$ ao longo de todas as amostras, é dada por $V(\overline{x}) = \dfrac{\left(m_1 - \overline{X}\right)^2 + \cdots + \left(m_k - \overline{X}\right)^2}{k}$. Por um desenvolvimento análogo ao que foi feito para $V(X)$, verifica-se que $V(\overline{x}) = \dfrac{m_1^2 + \cdots + m_k^2}{k} - \overline{X}^2$.

Vamos calcular a soma $m_1^2 + \cdots + m_k^2 = \left(\dfrac{y_{11} + \cdots + y_{1n}}{n}\right)^2 + \cdots + \left(\dfrac{y_{k1} + \cdots + y_{kn}}{n}\right)^2$

$= \dfrac{1}{n^2}\left[(y_{11} + \cdots + y_{1n})^2 + \cdots + (y_{k1} + \cdots + y_{kn})^2\right]$.

A expressão entre colchetes será a soma dos quadrados mais a soma dos duplos produtos dos *y*'s. Mas os *y*'s são os próprios X_j que aparecem nas amostras correspondentes. Como cada X_j aparece em C_{N-1}^{n-1} amostras, então a soma dos quadrados será igual a $C_{N-1}^{n-1}\left(X_1^2 + \cdots + X_N^2\right)$.

Por outro lado, o produto $2X_1X_2$, por exemplo, aparecerá tantas vezes quantas forem as amostras que contiverem X_1 e X_2 ao mesmo tempo, ou seja, C_{N-2}^{n-2} vezes. O mesmo ocorrerá com qualquer outro duplo produto. Logo, a soma dos duplos produtos será $2\,C_{N-2}^{n-2}\left(X_1X_2 + \cdots + X_{N-1}X_{N-2}\right)$.

Levando ainda em consideração que $k = C_N^n$, segue que:

$$\frac{m_1^2 + \cdots + m_k^2}{k} = \frac{C_{N-1}^{n-1}\left(X_1^2 + \cdots + X_1^2\right) + 2\,C_{N-2}^{n-2}\left(X_1X_2 + \cdots + X_{N-1}X_{N-2}\right)^2}{n^2\,C_N^n}.$$

Para simplificar, vamos fazer $X_1^2 + \cdots + X_1^2 = Q$ e $2\left(X_1X_2 + \cdots + X_{N-1}X_{N-2}\right) = P$.

Levando em conta que $C_{N-1}^{n-1} = \frac{n}{N}C_N^n$ e que $C_{N-2}^{n-2} = \frac{n(n-1)}{N(N-1)}C_N^n$, vem que:

$$\frac{m_1^2 + \cdots + m_k^2}{k} = \frac{1}{n^2}\left(\frac{n}{N}Q + \frac{n(n-1)}{N(N-1)}P\right) = \frac{1}{nN}\left(Q + \frac{n-1}{N-1}P\right).$$

Logo: $V(\overline{x}) = \frac{1}{nN}\left(Q + \frac{n-1}{N-1}P\right) - \overline{X}^2$

Por outro lado: $N^2\overline{X}^2 = \left(X_1 + \cdots + X_N\right)^2 = Q + P$, donde segue que $P = N^2\overline{X}^2 - Q$.

E ainda, como visto acima, $V(X) = \frac{Q}{N} - \overline{X}^2$, donde segue que $Q = N\,V(X) + N\,\overline{X}^2$ e, conseqüentemente, $P = N^2\overline{X}^2 - N\,V(X) - N\,\overline{X}^2 = N\left((N-1)\overline{X}^2 - V(X)\right)$

Portanto: $Q + \frac{n-1}{N-1}P = N\,V(X) + N\,\overline{X}^2 + \frac{n-1}{N-1}N\left((N-1)\overline{X}^2 - V(X)\right)$, enquanto

$$\frac{1}{nN}\left(Q + \frac{n-1}{N-1}P\right) = \frac{V(X)}{n} + \frac{\overline{X}^2}{n} + \frac{n-1}{n}\overline{X}^2 - \frac{n-1}{n(N-1)}V(X) = \frac{N-n}{n(N-1)}V(X) + \overline{X}^2$$

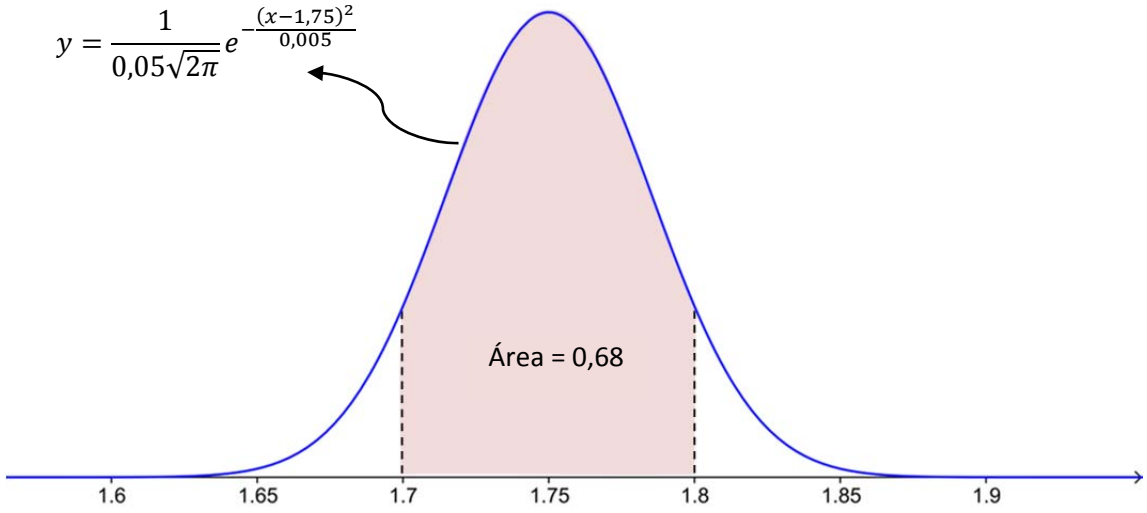
Finalmente:

$$V(\overline{x}) = \frac{N-n}{N-1}\frac{V(X)}{n} = \left(\frac{1-\frac{n}{N}}{1-\frac{1}{N}}\right)\frac{V(X)}{n}$$

Apêndice 3

A curva normal com média *m* e desvio padrão *s* tem expressão $y = \frac{1}{s\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2s^2}}$. O gráfico desta curva com m = 1,75 e s = 0,05 é exibido na figura abaixo. A curva normal é sempre simétrica com relação à média m e a área total sob a curva é igual a 1.

Dizemos que uma variável aleatória tem distribuição normal se a probabilidade do valor desta variável estar em um intervalo [a, b] for a área sob a curva normal no intervalo [a,b]. Por exemplo, suponha que a altura de certa população seja bem aproximada por uma distribuição normal com média m = 1,75 metros e desvio padrão s = 0,05 metros. Podemos então estimar o percentual da população que tem altura entre [1,70 1,80] calculando a área sob a curva $y = \frac{1}{0{,}05\sqrt{2\pi}}e^{-\frac{(x-1{,}75)^2}{0{,}005}}$ entre x=1,80 e x=1,90. Neste caso a área é 0,68 e, portanto, 68% da população tem altura no intervalo [1,70 e 1,80].



As áreas correspondentes a certos intervalos em torno da média são muito usadas: a área sob a curva no intervalo [m-s, m+s] é aproximadamente 68% da área total sob a curva (é o caso do exemplo acima). A área no intervalo [m-2s, m+2s] é aproximadamente 96%. E o intervalo em torno da média que corresponde a área de 95% é [m-1,96s, m+1,96s].

A curva normal é freqüentemente utilizada como modelo de distribuição de probabilidade de diversas medidas, de alturas de indivíduos até velocidades de moléculas de gás. No nosso contexto, a curva normal é importante por que fazemos uso do Teorema Central do Limite, segundo o qual, dada uma amostra aleatória simples, a média amostral tem distribuição de probabilidades bem aproximada pela curva normal, quando *n* é suficientemente grande. Portanto, usando amostras aleatórias simples, podemos usar a curva normal para avaliar as margens de erro.