Capítulo

3

A Curva Normal

Luiz Pasquali

1 – A História da Curva Normal

A curva normal, também conhecida como a curva em forma de sino, tem uma história bastante longa e está ligada à história da descoberta das probabilidades em matemática, no século XVII, que surgiram para resolver inicialmente questões de apostas de jogos de azar (veja Bernstein, 1997). O responsável mais direto da curva normal foi Abraham de Moivre, matemático francês exilado na Inglaterra, que a definiu em 1730, dando seqüência aos trabalhos de Jacob Bernoulli (teorema ou lei dos grandes números) e de seu sobrinho Nicolaus Bernoulli, matemáticos suíços. Publicou seus trabalhos em 1733 na obra *The doctrine of chances*. A descoberta teve logo grande sucesso e grandes nomes estão ligados à curva normal, tais como, Laplace que em 1783 a utilizou para descrever a distribuição dos erros, e Gauss que em 1809 a empregou para analisar dados astronômicos. Inclusive, a curva normal é chamada de distribuição de Gauss. Hoje em dia, a curva normal é um ganho fundamental em ciência, porque (1) a normalidade ocorre naturalmente em muitas, senão todas as medidas de situações físicas, biológicas e sociais e (2) é fundamental para a inferência estatística.

A lei dos grandes números de Bernoulli diz o seguinte: numa situação de eventos casualóides, onde as alternativas são independentes, obter coroa em lances de uma moeda de cara e coroa, tem a probabilidade matemática exata de 50% (porque são somente dois eventos possíveis: cara ou coroa), mas na prática esta probabilidade de 50% é apenas aproximada. E essa aproximação é tanto mais exata quanto maior forem as tentativas que você fizer de lançar a moeda,

chegando a quase atingir os exatos 50% se você lançar a moeda infinitas vezes. Isto é, quanto mais lances você fizer, menor será o desvio em relação à média de 50% que o resultado irá produzir. Isso quer dizer que os erros (desvios) serão menores e menores na medida em que sobe o número de lances. Desvios grandes são raros e desvios pequenos frequentes; quanto menores os desvios, mais frequentes eles serão, de sorte que, aumentando as tentativas (os lances), aumenta o número de desvios pequenos, sobrepujando cada vez mais os desvios grandes, de tal sorte que, no limite, haverá quase somente desvios pequenos, sendo o desvio 0 o menor deles e, por consequência, o mais frequente.

Moivre assumiu essa idéia de Bernoulli e disse: erros grandes são mais raros que erros pequenos. Assim, quanto menores os erros, mais freqüentes eles serão e quanto maiores, menos freqüentes. Dessa forma, os erros se distribuem eqüitativamente em torno de um ponto modal, a média, formando uma curva simétrica com pico na média e caindo rapidamente para as caudas à esquerda (erros que subestimam a média) e à direita (erros que superestimam a média). Além disso, essa curva simétrica permitiu a Moivre calcular uma medida de dispersão das observações em torno da média, medida esta que hoje em dia é conhecida como o desvio padrão (DP). Moivre chamou essa curva de normal, porque a média dela representa a norma, isto é, as coisas todas deviam ser como a média; de sorte que tudo que se desvia dessa média é considerado erro, donde a equivalência entre desvio e erro. Moivre defendeu essa idéia sob o conceito do homem médio ou mediano, idéia que provocou brigas homéricas na história da curva normal. Esta idéia do homem médio insinua, por exemplo, que todos os homens deveriam ter a mesma altura, o mesmo peso, a mesma inteligência etc., isto é, todos eles deveriam ser medianos; os desvios dessa norma podem ser considerados “aberrações” da natureza! Se você não introduzir concepções filosóficas, esse modo de pensar de Moivre é muito útil e prático para entender o que seja e para que serve a curva normal.

Quetelet, matemático belga do século XIX, fez uma “orgia de medições” (Bernstein, 1997: 158) sobre eventos do homem (tais como, natalidade, mortalidade, alcoolismo, insanidade, medidas antropométricas etc.), resultando no *Tratado sobre o homem e o desenvolvimento de suas faculdades* (1835), onde afirma que tudo no homem e no mundo se distribui segundo a curva normal (Stigler, 1986). Embora essa afirmação de Quetelet tenha tido reações contrárias, ela evocou pesquisas sem fim sobre esta história da distribuição normal dos eventos, chegando hoje em dia a ser mantida a idéia de que, praticamente, todos os eventos se distribuem assim. Daí, a hegemonia da curva normal nas análises estatísticas em pesquisas científicas.

Aliás, assumir a distribuição normal em pesquisa está baseado em dois fundamentos (Hays, 1963: p. 242): (1) quando a distribuição da própria população de eventos é normal (como insiste Quetelet para todos os eventos) ou

(2) quando a distribuição da população não for normal, mas o número de casos for grande (teorema de Bernoulli ou o teorema do limite central). Essa história do limite central é extremamente complicada, mas os matemáticos chegaram a provar o teorema. Assim, qualquer que seja a distribuição dos seus dados, se você tiver um número grande de observações, você pode utilizar com tranquilidade a curva normal como uma aproximação adequada para a análise dos seus dados. Uma curiosidade: um N de 30 já é considerado um grande número se a distribuição da população for próxima do normal; um N bem maior será necessário se a distribuição da população não for normal, como, por exemplo, o QI de engenheiros, porque sujeitos com QI mediano e baixo dificilmente serão encontrados entre os engenheiros. Há, contudo, um "porém" em tudo isso: O teorema dos grandes números se aplica quando a amostra da pesquisa for aleatória! Veja essa história no capítulo sobre amostragem (cap. 5).

2 – A Curva Normal e a Curva Normal Padronizada

Os pesquisadores quando falam de curva normal, tipicamente entendem a curva normal padronizada, a qual é definida pela simetria e pela curtose. Mas a curva normal original é definida exclusivamente pela simetria, isto é, que as áreas sob a curva são idênticas em ambos os lados da média: a curva normal é unimodal (tem apenas um pico) e simétrica. Assim, todas as curvas da figura 3-1 são normais, porque têm um pico somene e são simétricas, embora os desvios sejam diferentes, provocando diferentes níveis de curtose.

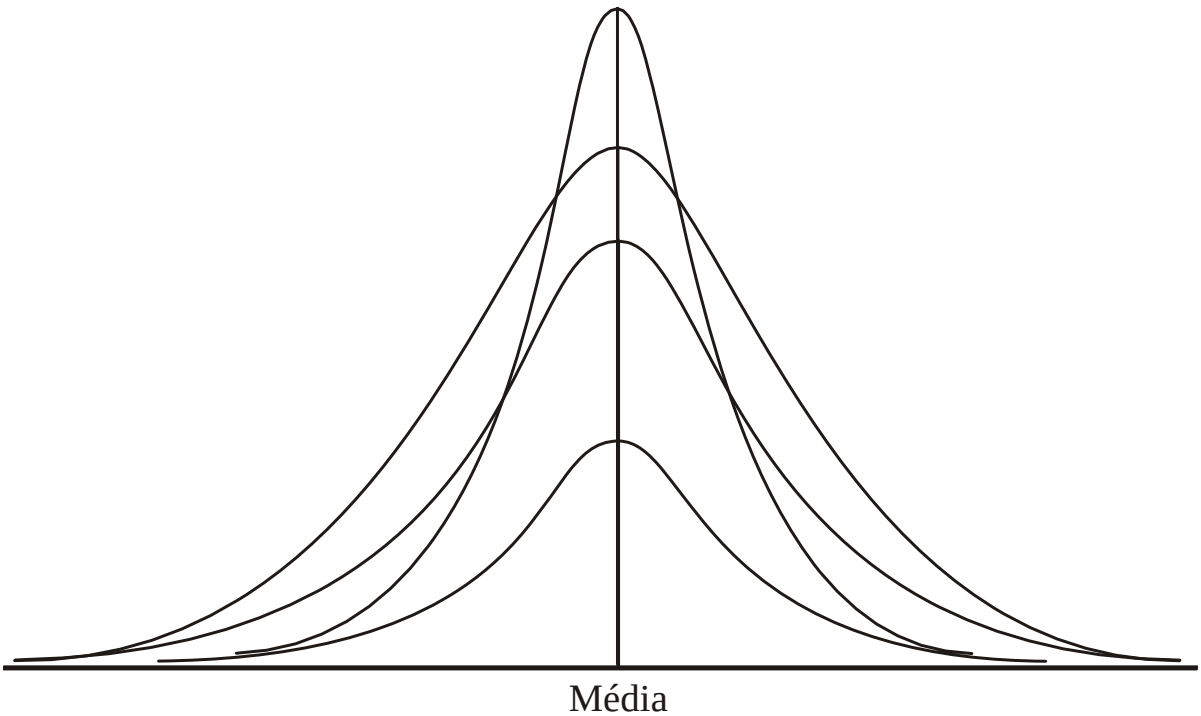

Figura 3-1. Distribuições normais

Mais ainda, curvas normais podem ter médias diferentes (figura 3-2a), desvios-padrão diferentes (figura 3-2b) ou ambas as coisas (figura3-2c).

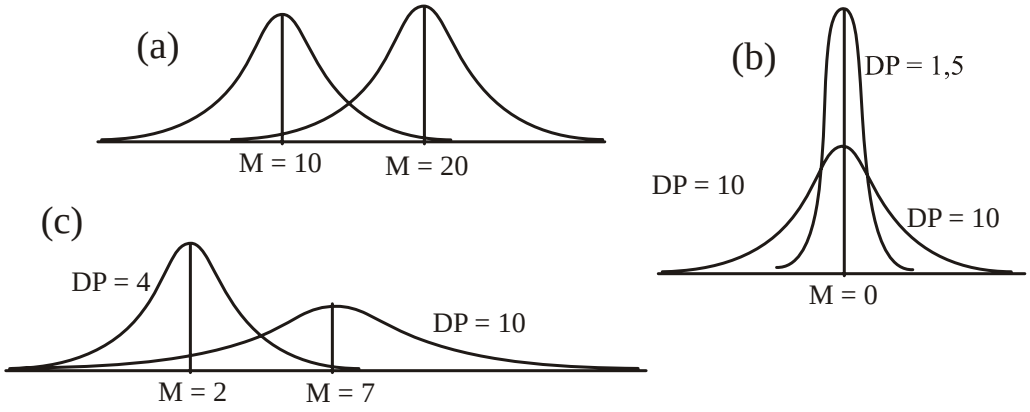

Figura 3-2. Distribuições normais com diferentes médias e desvios-padrão

Isso acontece porque a curva normal trabalha diretamente com os escores originais X e os seus parâmetros da distribuição, a saber, a média e o desvio-padrão (que são os dois parâmetros fundamentais da curva normal), conforme se vê na sua fórmula:

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X-M)^2/2\sigma^2} \qquad (3.1)$$

Nessa fórmula complicada, a parte mais importante é o expoente:

$$-\frac{(X-M)^2}{2\sigma^2} \qquad (3.2)$$

e nele se vê que quem comanda as ações são os dados empíricos X e os parâmetros de sua distribuição (M e DP). Agora, tanto os X quanto os parâmetros de uma distribuição variam de pesquisa para pesquisa e, assim, as curvas normais que resultam serão diferentes. Entretanto, você se lembra do capítulo anterior que tendo esses dados, isto é, o X, a Média e o DP, eu posso transformar qualquer escala em escores-padrão *z*, tornando todas as escalas idênticas e diretamente comparáveis. Pois é, aqui posso fazer a mesma coisa: em lugar de trabalhar com os escores brutos X, posso transformá-los em escores *z*. Agora, a distribuição da curva normal que resulta com escores padronizados é a famosa curva normal padronizada, aquela que todo o mundo entende quando se fala simplesmente da curva normal. Inclusive, a fórmula desta curva normal padronzada aparece como mais simples (para os estatísticos), ou seja:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad (3.3)$$

A vantagem dessa curva normal padronizada consiste em que alguns parâmetros já estão automaticamente definidos para qualquer escala de medida que você utilizar, quais seja, a média é sempre 0 e a variância é sempre 1. Além disso, existem tabelas construídas para essa curva que mostram quanto por cento da população se encontra dentro de cada faixa de *z*, como veremos a seguir, tabelas estas que você encontra em qualquer livro de estatística.

A curva normal padronizada é definida pela simetria e pela curtose; ela é chamada de mesocúrtica. Vejamos essa história da curtose.

A curtose da curva normal se refere à altura do pico da curva, o qual acontece na média da distribuição: se o pico é muito elevado, a curva é chamada de leptocúrtica; se o pico é achatado, a curva se chama platicúrtica e se for mediano, a curva será mesocúrtica, sendo esta última, a característica da curva normal padronizada. Veja a figura 3-3 para visualizar a curtose das curvas normais.

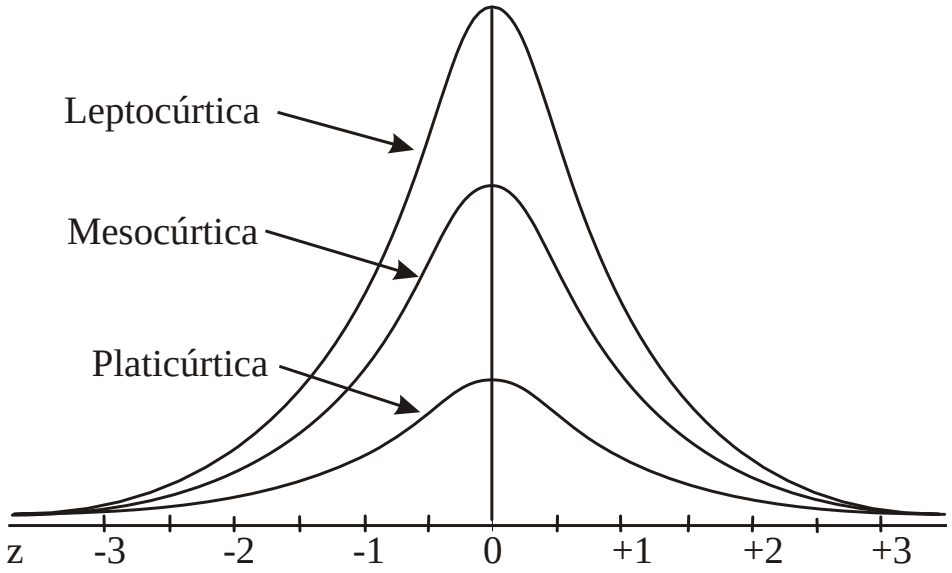

Figura 3-3. As distribuições da curva normal

Em pesquisas, quando se fala de curva normal, sem maiores detalhes, normalmente se está falando ou assumindo a curva normal padronizada, isto é, a curva normal mesocúrtica.

3 – As Áreas Sob a Curva Normal

Quanto maior for o expoente da fórmula da curva normal, qualquer delas (inclusive a padronizada – veja fórmulas 3.1 e 3.3), mais rapidamente a curva vai caindo para a abscissa; mas ele nunca chegará a zero. De sorte que as caudas da curva vão até o infinito; elas são assíntotas. Assim, a curva normal cobre uma área que vai do -∞ a +∞. As áreas sob a curva são dividas pelo desvio-padrão em torno da média. Quando você trabalha com a curva normal padronizada, a média é 0 e o desvio-padrão é 1. Quando não for a padronizada, então você tem que calcular a média e o DP da distribuição e trabalhar com os dois parâmetros. Você vê, então, que trabalhar com a curva normal padronizada facilita enormemente a vida da gente. De qualquer forma, o que define as áreas sob a curva são os DP, ou os *z* no caso da curva normal padronizada. E, para cada DP ou *z* em torno da média, corresponde uma proporção bem definida de casos da população que caem dentro deles. Veja, por exemplo, o caso com a curva normal padronizada na figura 3-4.

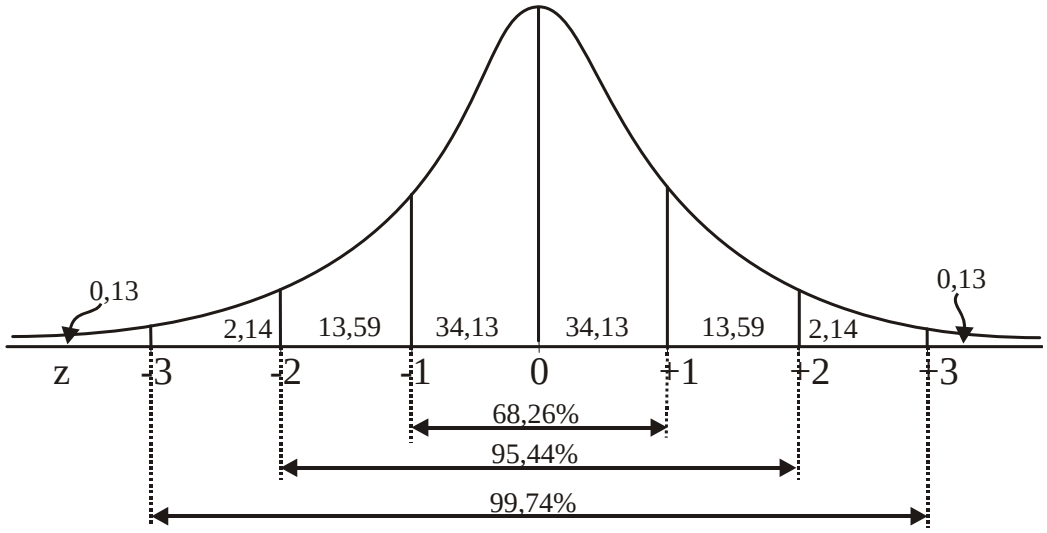

Figura 3-4. Áreas da curva normal e percentagem de casos

Embora a curva normal vá até o infinito (positivo e negativo), você vê que a quase totalidade dos casos cai entre -3 e +3 DP (ou *z*); de fato, 99,74% dos casos.

4 – Utilizando as Tabelas da Curva Normal

Qualquer livro de estatística traz a tabela da curva normal, muitas vezes apropriadamente intitulada como *proporções da área sob a curva normal padronizada*. As informações contidas nessa tabela não são sempre idênticas

nos diferentes autores. Entretanto, duas informações sempre estão presentes e essas são as mais importantes, a saber, o *z* e a proporção de casos que caem na faixa que vai da média (0) até este *z*. Assim, se você conhece o *z*, você pode descobrir qual a proporção de casos que corresponde a ele ou, se você conhece a proporção de casos, você pode descobrir qual o *z* que lhe corresponde. Não tem nada de mágico nessa história; apenas, precisa um pouco de prática para realizar a tarefa adequadamente. Vamos dar alguns exemplos; fique olhando para a figura 3-5.

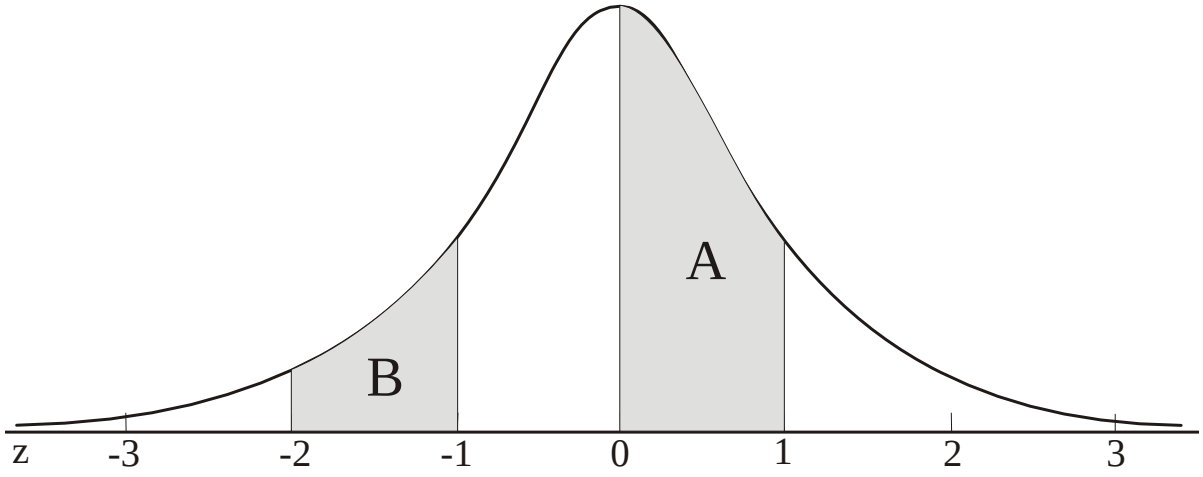

Figura 3-5. Proporções de casos sob a curva normal padronizada

Se quiser saber quanto por cento dos casos caem nas faixas A e B, faço o seguinte:

Para a faixa A: na coluna *z* da tabela da curva normal procuro o valor 1 (porque a faixa A vai de 0 a 1); ao lado, na coluna Área, acho a percentagem de casos (a proporção), que no caso diz 0,3413447, isto é, entre 0 (a média da distribuição) e 1z caem 34,13% dos casos.

Para a faixa B: aqui é um pouco mais complicado, porque a faixa cai entre -1z e -2z, e não mais entre 0 e algum *z*. Assim, devo, primeiramente, procurar a proporção que cai entre 0 e -2z (como a curva é simétrica, pode desconsiderar o sinal antes do 2z); em seguida, procuro a proporção que cai entre 0 e -1z; por fim, faço a diferença entre as duas proporções encontradas e surge a proporção da faixa B. Veja:

Entre 0 e 2z: proporção = 0,4772499
Entre 0 e 1z: proporção = 0,3413447
Diferença: 0,4772499 – 0,3413447 = 0,1359052
Assim, na faixa B caem 13,59% dos casos.

Nota: se você estiver trabalhando com uma escala não padronizada e quer saber quantos sujeitos estão abaixo ou acima de um escore qualquer da sua escala, basta primeiro transformar esse escore da sua escala em escore

padrão pela fórmula usual, isto é, z = (X–M)/DP, e procurar o resultado na tabela da curva normal padronizada, como explicado acima.