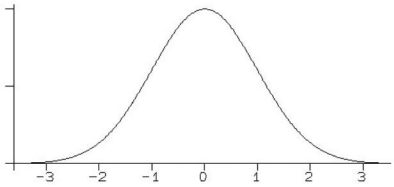
3.4.4. Modelo normal

Este capítulo apresenta o modelo *normal* ou *gaussiano*, certamente o modelo probabilístico mais importante da Estatística. Seu gráfico tem a forma de um sino (Fig. **1A**), forma que passou a ser associada com "Estatística" na cabeça das pessoas, e costuma por isto ser usada em propaganda ou logotipos como o da ABE (Fig. **1B**).



A. Curva normal padrão

B. Logotipo da ABE

Figura 1. Curva normal

Na seção 3.4.4.1 iremos discutir o que são modelos matemáticos e para que servem; este assunto já foi mencionado em seções anteriores, mas é tão importante, especialmente em relação ao modelo normal, que vale a pena voltar a ele. Na seção 3.4.4.2 mostramos as características do modelo normal e como ele é usado; na 3.4.4.3, algumas de suas propriedades, e porque ele é tão útil na Estatística; na 3.4.4.4, como identificar e ajustar um modelo normal para uma dada variável; por fim, na 3.4.4.5, como o modelo normal pode ser usado como aproximação para outros modelos.

3.4.4.1. Introdução aos modelos matemáticos

(i) Porque usar modelos matemáticos

Toda ciência se baseia na seleção de modelos que representam de forma simplificada as características das variáveis de interesse, e depois na utilização das propriedades destes modelos para fazer as contas e deduzir as conseqüências que nos interessam.

Por exemplo, usamos uma esfera como modelo da forma da Terra. A esfera só tem um parâmetro, o raio r ; se conseguimos uma estimativa deste raio, podemos calcular quantidades como a superfície S da Terra, seu volume V, sua massa (se conhecemos o volume e a densidade média), ou a área T de sua seção transversal, usando as equações conhecidas:

$$V = \frac{4}{3}\pi r^3 \qquad S = \pi r^2 \qquad T = \pi r^2$$

Um navegador no passado que usasse a esfera como modelo para a Terra conseguiria estimar sua posição no mar a partir da posição das estrelas; estas pareceriam estar em posições diferentes, dependendo de em qual parte da esfera o navio estivesse. A estrela Polar, por exemplo, seria visível se ele estivesse na parte de cima da esfera (hemisfério Norte), e formaria um ângulo de 45º com o horizonte se ele estivesse na latitude 45; seria invisível se ele estivesse na parte de baixo da esfera (hemisfério Sul). O navegador que partisse de um modelo que considerasse a Terra plana não poderia se orientar a partir das estrelas.

Toda conclusão ou previsão na ciência é feita a partir de modelos; contudo, é importante nos lembrarmos sempre de que modelos são *abstrações* matemáticas, que não existem no mundo real, e que os cientistas trabalham a partir da *pressuposição* de que um tal modelo seja o melhor para um dado problema. Quanto mais próximo estiver o modelo da realidade (no exemplo, quanto mais "esférica" for realmente a Terra), mais precisos serão os resultados dos cálculos.

Um exemplo de problema equivalente em Estatística poderia ser: se pesquisadores querem estudar o peso e a altura de crianças de dois anos de idade, para determinar em que faixas de valores se encontram a maioria das crianças saudáveis, irão considerar que *peso* e *altura* são variáveis aleatórias e escolher modelos para elas. A partir destes modelos, poderão conhecer como estas variáveis se distribuem: que faixas de valores têm maior probabilidade de ocorrência, que faixas têm menor probabilidade, e quais são seus parâmetros (mé-dia, desvio-padrão). Esta informação será útil para fazer os gráficos de crescimento que os pediatras usam para avaliar se uma criança está se desenvolvendo normalmente.

(ii) Como escolher o modelo

Na maioria das variáveis aleatórias discretas que vimos (seção 3.3), as características do problema guiavam a escolha dos modelos. Por exemplo, se a variável é gerada pela repetição de "tentativas" binárias que podem resultar em *sucesso* ou *fracasso*, modelos diferentes devem ser usados se o número de repetições for ou não constante, se a probabilidade de sucesso a cada tentativa depender ou não do resultado da tentativa anterior, se o número de sucessos desejados for ou não pré-definido, etc.

Nas VAs contínuas, porém, em geral não são as características do problema que definem a escolha do modelo, mas sim considerações sobre as (prováveis) características da população. Imaginamos como deve a distribuição da variável nesta população, depois

escolhemos um modelo, dentro de repertório de modelos que estão nos livros de Estatística, que pareça melhor descrever aquelas características. Para ilustrar isto, voltemos ao exemplo da forma da Terra. Os primeiros astrônomos que usaram a esfera para descrever a forma da Terra escolheram entre os modelos geométricos disponíveis (Fig. **2**) aqueles que mais pareciam coerentes com a informação de que dispunham na época (muito antes que as naves espaciais tivessem fotografado o planeta de todos os ângulos): a forma arredondado do mar, visto do alto do mastro de um navio, a sombra circular da Terra projetada sobre a lua, a forma circular da lua e do sol (vistas a olho nu). O modelo que pareceu mais adequado, é claro, foi a esfera. (A razão porque a Terra e os outros planetas têm esta forma foi explicada pela teoria da gravitação de Newton; mas isto só aconteceu vários séculos mais tarde).

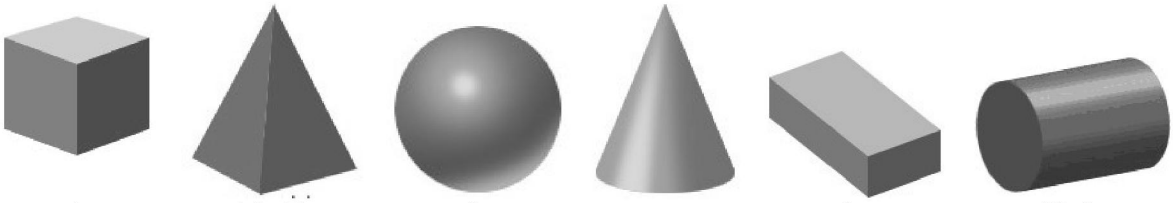
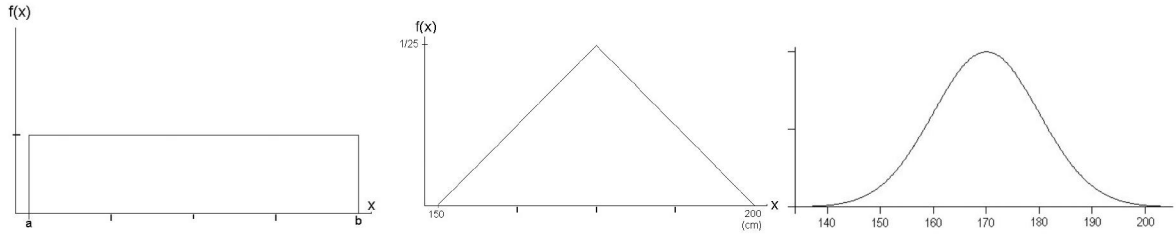


Figura 2. Modelos geométricos sólidos

Suponhas agora que nos interessa estudar e modelar, por exemplo, a variável *altura* na população de *homens brasileiros*, e queremos escolher um modelo para sua distribuição. Alguns modelos são mostrado na Fig. **3**.



1. Modelo uniforme 2. Modelo "triangular" 3. Modelo normal

Figura 3. Três modelos de distribuição de VAC

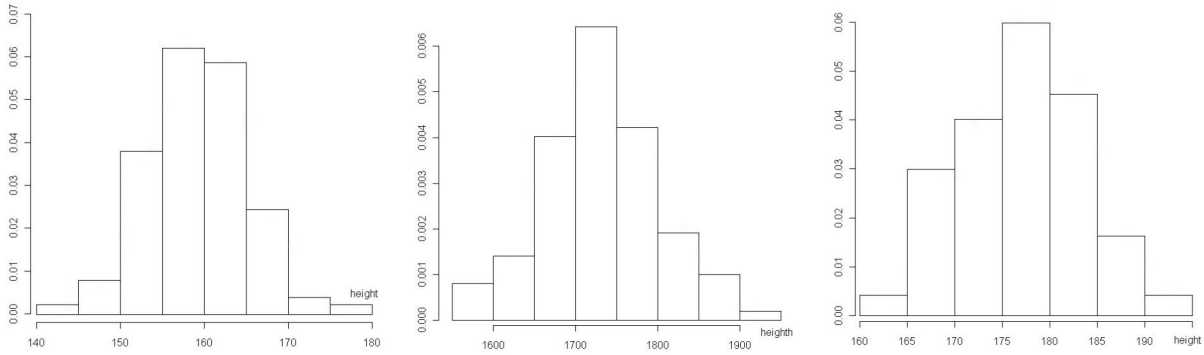
É evidente que um modelo de distribuição uniforme com parâmetros a=150 cm e b= 200 cm (Fig. **3A**) não parece muito apropriado. Primeiro, porque ele impõe limites aos valores da variável, e sugere que não existem homens com mais de 200 cm ou menos de 150 cm; contudo, sabemos por experiência empírica que existem homens com tais alturas, embora sejam raros. Além disso, este modelo diz que P(190<X<200) = P(170<X<180), por exemplo, o que também não é confirmado por nossa experiência; esperamos que haja muito mais homens na faixa 170<X<180 do que na faixa 190<X<200.

Um modelo baseado num triângulo isósceles (Fig. **3B**) definido pelas equações:

$$f(x) = \begin{cases} (x-150)/625 & para\ \ 150 < x < 175 \\ (200-x)/625 & para\ \ 175 < x < 200 \\ 0 & para\ \ x < 150\ ou\ x > 200 \end{cases}$$

é um pouco melhor, pois mostra que homens com alturas perto da média (175 cm) são mais comuns do que aqueles muito altos ou muito baixos, afastados da média; contudo, este modelo também impõe limites à variável, e pressupõe que não existam homens com mais de 200 ou menos de 150 cm. Um modelo mais realista seria um que atribua maiores probabilidades aos valores em torno da média, não imponha limites na altura (nem à esquerda, nem à direita), e dê probabilidades decrescentes para intervalos que se afastem do centro da distribuição. Alturas na faixa 200-210 cm, por exemplo, não seria impossíveis, mas teriam probabilidade muito pequena; o intervalo 210-220 cm teria probabilidade menor ainda, etc. (Por curiosidade: o homem mais alto já registrado foi o americano Robert Wadlow, que chegou até aos 274 cm). O modelo que desejamos deveria ter um gráfico parecido com o da Fig. **3C**; este gráfico é o do modelo *normal*, assunto deste capitulo

A análise é relativamente fácil neste problema porque temos experiência no dia-a-dia com este tipo de variável; sabemos quais faixas de alturas são razoavelmente comuns, e quais são mais raras; sabemos que há mais homens com 175 cm de altura do que homens com 200 cm, etc. Se estamos porém interessados numa variável como o "nível de bilirrubina nos sangue", provavelmente não temos nenhuma idéia *a priori* de como será sua distribuição (a bilirrubina é uma substância encontrada no plasma sanguíneo). Considerando que a maior parte das variáveis biológicas têm distribuições unimodais e razoavelmente simétricas, talvez possamos simplesmente pressupor que o nível de bilirrubina tenha distribuição aproximadamente normal. Veremos depois (seção 4.5) que esta aproximação é geralmente suficiente para a maior parte dos trabalhos que exigem inferência estatística, especialmente se as amostras forem razoavelmente grandes.



A. 351 mulheres inglesas B. 199 homens ingleses C. 234 alunos da UFJF

Figura 4. Amostras de três populações – variável: altura

Para auxiliar na escolha do modelo, podemos também tirar uma amostra da população e nos basearmos na informação dada pelo seu histograma. (Um exemplo disto foi visto em relação ao modelo exponencial na seção 3.4.3). Uma amostra deve ser como uma miniatura da população; a forma de sua distribuição deve espelhar aproximadamente a forma da distribuição na população. Na Fig. **4** estão os histogramas das alturas em três amostras de pessoas adultas. Nas três amostras, a distribuição da altura é unimodal e próxima da simetria, e tem um formato que indica que o modelo normal pode ser apropriado para esta variável. Voltaremos a este exemplo na seção 3.4.4.4, depois de apresentarmos as características do modelo normal.

3.4.4.2. Características do modelo normal

(i) Função de densidade

A primeira versão da função de densidade de probabilidade deste modelo foi publicada por Gauss em 1809. A versão mais usada atualmente inclui modificação feitas por Pearson e Fisher, no início do século XX, que resultaram nesta expressão:

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

Este modelo parece bem complicado, e realmente é – todos os modelos de VAC parecem complicados para aqueles que não têm uma boa base em Matemática (ou seja, para a maioria dos pobres mortais como nós). Entre os modelos mais importantes de VAC, contudo, o modelo normal é comparativamente o mais simples; modelos mais complexos foram desenvolvidos a partir dele, como o de Student. Esta complexidade dificulta o trabalho de pesquisadores na Estatística Matemática (a área da Estatística que se ocupa com demonstrar teoremas, derivar propriedades dos modelos, ou criar novos modelos). Para quem usa a Estatística como ferramenta de pesquisa, esta complexidade não importa muito, já que todos os cálculos atualmente são feitos por meio de algum programa de computador,

Para nos familiarizarmos com o modelo, vejamos algumas de suas características mais evidentes. Primeiro, este modelo tem a curiosidade de ter sido o primeiro a reunir numa só fórmula as duas constantes mais importantes da Matemática: o número π (≅ 3,14), razão entre a circunferência e o diâmetro de um círculo, e o número *e* (≅ 2,7183), base dos logaritmos neperianos, bastante usado no Cálculo.

Segundo, as outras letras gregas na fórmula, μ (pronuncia-se *mi* ou *mu*) e σ (sigma), são os *parâmetros* do modelo; pode ser demonstrado que são iguais ao valor esperado e ao desvio-padrão da distribuição, respectivamente:

$$E(X) = \mu$$

$$\text{desvio-padrão}(X) = \sigma \;\rightarrow\; V(X) = \sigma^2$$

Para indicar que uma variável X tem distribuição normal com média μ e variância σ^2, usamos a notação:

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

É importante observar que nesta notação é usada a variância σ^2 do modelo, e não o desvio-padrão σ (embora o desvio-padrão seja mais importante, como veremos abaixo).

Terceiro, a variável *X* aparece no expoente dentro do parênteses, no termo:

$$\frac{x-\mu}{\sigma}$$

Este termo é chamado de *variável padronizada*, representado em geral pela letra *z*.

Uma vez que π é uma constante conhecida, e que μ e σ são os parâmetros que iremos escolher para o modelo, também já conhecidos, podemos substituir alguns termos da expressão. Chamando:

$$a = \frac{1}{\sigma\sqrt{2\pi}} \qquad b = \frac{1}{2} \qquad z = \frac{x-\mu}{\sigma}$$

O modelo fica reescrito como:

$$f(z) = ae^{-bz^2}$$

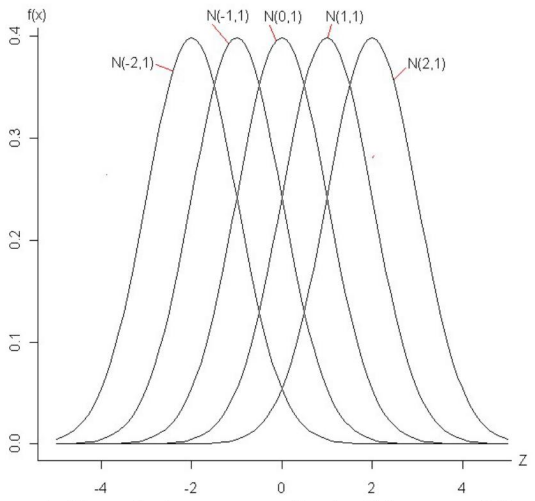
Feitas estas substituições, o modelo normal fica bastante parecido com o exponencial, com a diferença que variável (no expoente) está elevada à segunda potência

exponencial: $f(x) = \alpha e^{-\alpha x}$

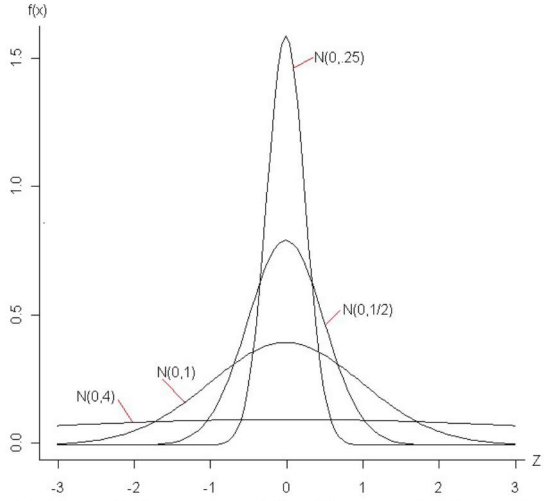
normal: $f(z) = ae^{-bz^2}$

(ii) Parâmetros do modelo normal

Este modelo gera uma "família" de distribuições, cujos gráficos têm a mesma forma: são distribuições unimodais e simétricas que lembram o perfil de um sino, com caudas que se estendem assintoticamente ao infinito. O que pode variar de um modelo normal para outro é a *posição* da curva ao longo do eixo horizontal, e a "largura" da curva (a *dispersão* da distribuição); estas características são determinadas pelos parâmetros.



A. Normais de mesma variância, diferentes médias



B. Normais de mesma média, diferentes desvios-padrões

Figura 5. Distribuições normais com diferentes médias e desvios-padrões

O parâmetro μ (chamado de *parâmetro de localização* ou de *posição*) serve para indicar onde está a média da distribuição; se alterarmos este parâmetro, alteramos a localização da curva ao longo do eixo. A Fig. **5A** mostra cinco modelos normais que têm todos o mesmo desvio-padrão σ=1, mas médias variando de μ=–2 a μ=+2.

O parâmetro σ (chamado de *parâmetro de escala*) indica a dispersão da distribuição; se alteramos este parâmetro, tornamos a curva mais estreita ou mais larga. A Fig. **5B** mostra quatro modelos normais que têm a mesma média μ=0, mas desvios-padrões variando de σ=0,25 a σ=4. Por fim, a Fig. **6** compara três distribuições que têm μ e σ diferentes.

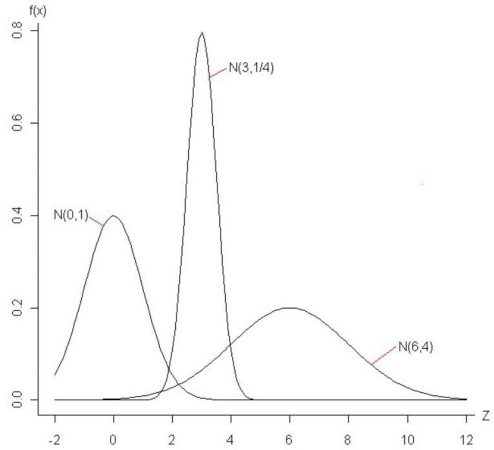


Figura 6. Distribuições normais, diferentes médias e desvios-padrões

(iii) Variável normal padronizada

O modelo normal é baseado numa função exponencial (seção 3.4.3), modificado por algumas constantes, e com a variável original *X* sofrendo uma *translação* e um *reescalonamento*. *Transladar* uma variável significa adicionar ou subtrair uma constante a esta variável; no caso, subtrair a média:

$$x - \mu$$

A translação faz a valor da variável ser deslocado ao longo do eixo. *Reescalonar* uma variável significa multiplicar ou dividir a variável por uma constante; no caso, a variável transladada é dividida pelo desvio-padrão:

$$\frac{x-\mu}{\sigma}$$

O resultado destas duas operações é chamado de *variável padronizada*, e é geralmente representado pela letra *z*:

$$z = \frac{x-\mu}{\sigma}$$

Note que a variável padronizada *Z* nada mais é do que a distância entre o valor da variável *X* e o centro μ da curva, medida em desvios-padrões σ.

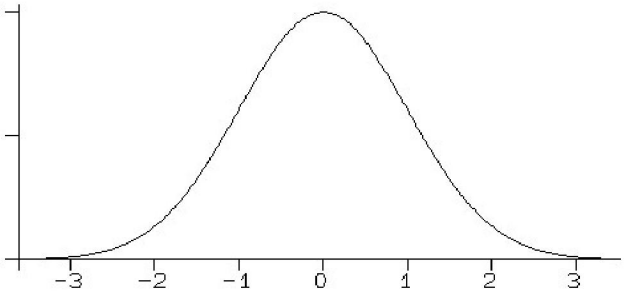


Figura 7. Curva normal padrão

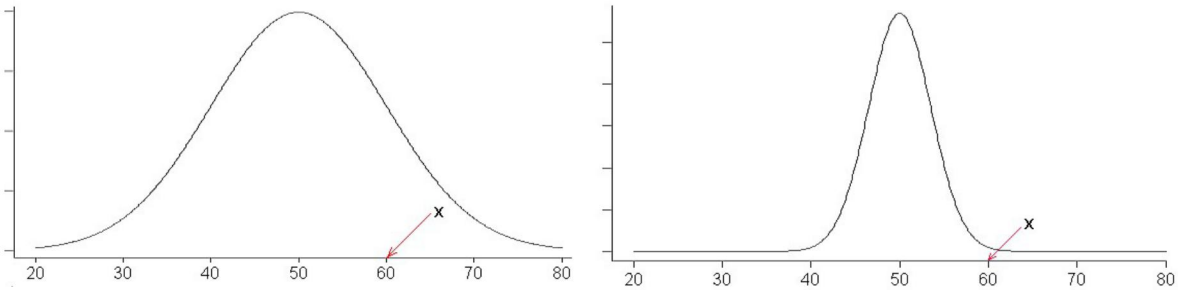
A padronização de uma variável não altera a *forma* da distribuição. A variável transformada Z continua tendo distribuição que pertence a família das gaussianas, mas com parâmetros diferentes: esta distribuição é chamada de *normal padrão*, e tem μ=0 e σ=1. O modelo normal padrão (Fig. 7) foi o que Gauss propôs em sua primeira publicação.

Embora a curva seja teoricamente ilimitada, tanto à direita quanto à esquerda, é possível perceber pelo gráfico que a maior parte de sua área está entre os valores de z=–2 e z=+2 (veremos que a este intervalo corresponde uma probabilidade de 0,9545), e que praticamente toda ela está entre os valores de z=–3 e z=3 (probabilidade de 0,9973).

(iv) Exemplo do uso da variável padronizada

Para ilustrar a utilidade da padronização da variável, usaremos como exemplo as notas do SAT (*Scholastic Aptitude Test*), um exame de admissão para várias universidades americanas (algo como o vestibular no Brasil). Numa prova deste tipo, o que interessa não é saber quanto o aluno sabe daquelas matérias; o que interessa é saber se ele sabe *mais* ou *menos* do que os outros alunos, porque todos estão competindo pelas mesmas vagas.

Suponha que o valor da prova seja 100, e um aluno tirou nota 60. Isto é um bom resultado ou não, em relação ao resto da turma? Não é possível saber, se não conhecemos a média e o desvio-padrão da distribuição das notas. Se a média foi μ=50, a nota do aluno parece ter sido boa; pelo menos, foi melhor do que a média. Podemos porém dizer que esta nota foi excelente? Isto vai depender de qual foi o desvio-padrão da distribuição. O gráfico da Fig. **8A** mostra a distribuição das notas se σ=10. Neste gráfico, a nota X=60 é boa, mas não é excepcional; há uma grande proporção de alunos que tiraram notas melhores do que esta. O gráfico da Fig. **8B** mostra uma distribuição na qual σ=3,5. Neste gráfico, uma nota de X=60 aparece como excelente; muito poucos alunos tiraram notas melhores do que esta.



A. Distribuição normal, μ=50, σ=10 B. Distribuição normal, μ=50, σ=3,5

Figura 8. Distribuições normais de mesma média e diferentes desvios-padrões

Não podemos portanto comparar diretamente as notas se as distribuições têm médias ou variâncias diferentes; para que a comparação seja possível, precisamos primeiro *padronizar* estas notas. O valor padronizado indica se a nota do aluno foi maior do que a média (Z> 0) ou menor do que a média (Z< 0); indica além disso qual foi a distância entre a nota do aluno e a média, usando o desvio-padrão como unidade (ou seja, quantos desvios-padrões a nota ficou acima ou abaixo da média).

Se o desvio-padrão da distribuição das notas foi σ=10, um nota x=60 equivale a um valor padronizado:

$$z = \frac{x-\mu}{\sigma} = \frac{60-50}{10} = 1{,}00$$

Se estas notas têm uma distribuição normal (e as provas do SAT são elaboradas de forma que a distribuição das notas se aproxime o máximo de uma normal), podemos ver na curva normal padrão onde esta nota se localiza, em relação ao resto da distribuição. A Fig. **9A** mostra a distribuição da variável padronizada (Z) e da variável original (X); a forma da distribuição é a mesma, o que muda é apenas a escala do eixo horizontal.

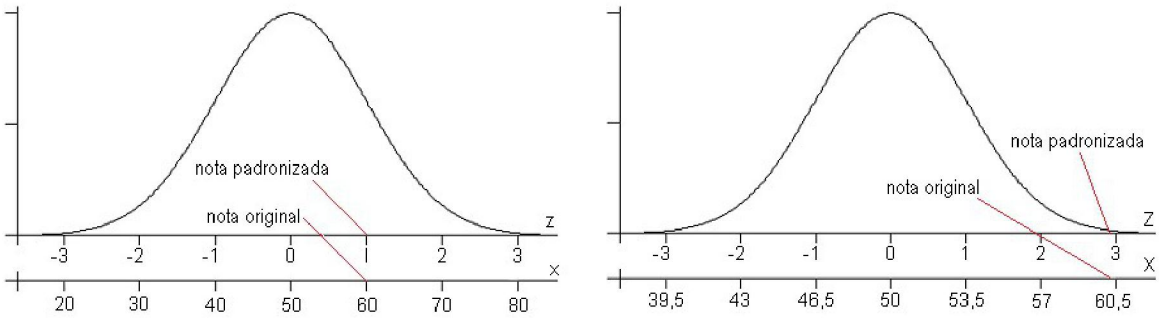


Figura 9. Notas no SAT na variável original e na variável padronizada

A julgar pela posição do valor padronizado no gráfico, esta nota foi boa, mas não excepcional. Se porém o desvio-padrão foi σ=3,5, um nota x=60 equivale a um valor padronizado:

$$z = \frac{x-\mu}{\sigma} = \frac{60-50}{3,5} = 2,86$$

Esta nota é excelente, e o aluno parece ter distanciado muito dos outros concorrentes (Fig. **9B**). Para avaliar numericamente quão boas são estas notas, temos que calcular as probabilidades de um aluno tirar notas acima ou abaixo destes valores. (veremos este cálculo na seção seguinte).

Na verdade, as notas do SAT sofrem ainda outra transformação, antes de serem publicadas, para evitar que alunos tenham notas negativas (o que a maioria das pessoas não entenderia). A nota final é definida como uma variável Y:

$$Y = 500 + 100Z$$

A média das notas da turma corresponde a um valor Y=500. As notas padronizadas z=1,0 e z=2,86, do exemplo acima, corresponderiam a notas Y=600 e Y=786, respectivamente. As notas extremas z=–3 e z=+3 correspondem a Y=200 e Y=800 (notas abaixo de z=–3 ou acima de z=+3 são arredondadas para os valores 200 e 800). Este tipo de critério foi usado também em algumas universidades brasileiras (por exemplo, a UFJF nos anos 1980), mas foi em geral abandonado. (Note que a nota Y é uma translação e reescalonamento da nota original, e também terá distribuição normal, com média μ=500 e desvio-padrão σ=100).

(v) Cálculo das probabilidades usando a tabela da curva normal padrão

Como visto na seção 3.4.1.1, para calcular as probabilidades de um intervalo de uma VAC temos que calcular a área sob a curva da função de densidade *f(x)* naquele intervalo. Isto é feito, teoricamente, por meio da integração da função no intervalo desejado. Se queremos calcular a probabilidade de um aluno tirar nota maior que 60, na prova do exemplo acima, precisamos calcular:

$$P(X > 60) = \int_{60}^{\infty} f(x)dx$$

Graficamente, isto equivale a determinar a área sob a curva para notas maiores que 60, como destacada no gráfico da Fig. **10** (este gráfico foi feito supondo σ=10).

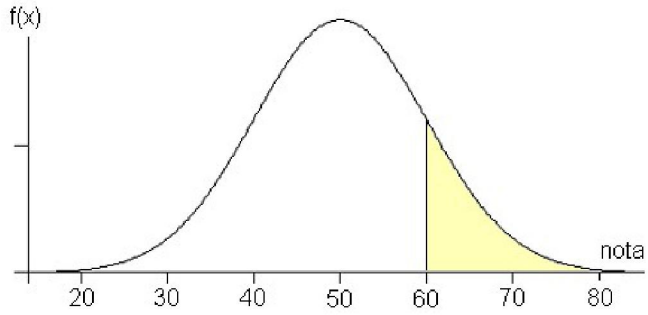
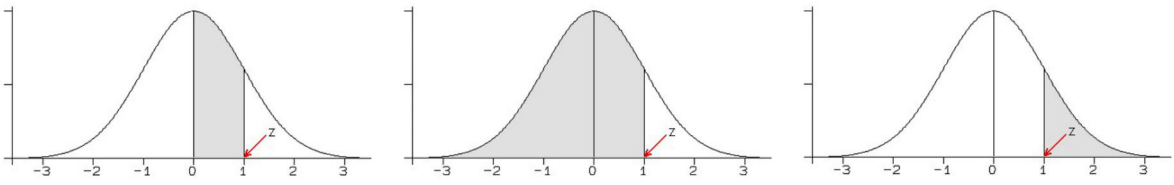


Figura 10. Nota no SAT e área a ser calculada

A dificuldade é que a *f(x)* de uma distribuição normal não tem uma integral conhecida; para obter seu valor é preciso usar técnicas de integração numérica (que é o que os computadores fazem), ou tabelas que dão estas áreas de forma aproximada. Como existem infinitas curvas normais possíveis (com diferentes parâmetros μ e σ), as tabelas são feitas sempre para a distribuição normal padronizada (μ=0, σ=1); se quisermos calcular as probabilidades de uma variável X que tenha um modelo normal diferente, temos que padronizar X e usar depois a tabela da distribuição padrão.

Existem várias formas de tabelas para a distribuição normal padrão. Em algumas delas a tabela dá a área entre o valor de z encontrado e o centro da distribuição, isto é, P(*0*<*Z*<*z*), como na Fig. **11A**. Em outras, a tabela dá a área abaixo deste z, P(*Z*<*z*), ou acima de z, P(*Z*>*z*), como nas Figs. **11B** e **11C**, respectivamente.



P(0<Z<z) **P(Z<z)** **P(Z>z)**

Figura 11. Áreas calculadas por diferentes tipos de tabelas de normal padrão

Usaremos como exemplo uma tabela como o da Fig. **11A**. Uma parte desta tabela é reproduzida na Tab. **1**; a tabela completa está na seção 3.6.3.

Dado um valor *z*, para calcularmos a probabilidade P(*0* < *Z* < *z*) – isto é, a probabilidade de a variável Z estar entre *0* e o valor *z* –, procuramos o algarismo dos inteiros e da primeira decimal de *z* (1,0) na primeira coluna, e em seguida o algarismo da segunda decimal (0,00) na primeira linha. A célula que está no cruzamento entre esta linha e esta coluna contém a probabilidade desejada.

Voltando ao exemplo do SAT. Se o valor de Z encontrado foi igual z=1,00, a área entre este valor e o centro da distribuição (Fig. **12A**) será igual 0,3413, dada pelo cruzamento da coluna e da linha destacadas em vermelho na Tab. **1**.

P(*0* < *Z* < 1,00) = 0,3413.

Tabela 1. Probabilidades na distribuição normal

z	**0,00**	0,01	0,02	0,03	0,04	0,05	**0,06**	0,07	0,08	0,09
...	...	...	...	...	...	...	...	...	...	...
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	**0,3413**	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
...	...	...	...	...	...	...	...	...	...	...
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	**0,4979**	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
...	...	...	...	...	...	...	...	...	...	...

Se quisermos a probabilidade de um aluno tirar uma nota melhor do que esta, ou a proporção de alunos que tiraram notas melhores do que esta (representada pela área em vermelho no gráfico da Fig. **12B**), teremos que considerar a metade direita da curva, e subtrair dela a área em azul; o resultado é:

$$P(Z > 1{,}00) = 0{,}5 - P(0 < Z < 1{,}00) = 0{,}5 - 0{,}3413 = 0{,}1587$$

Portanto, 15,87 % dos alunos tiraram notas melhores do que esta; esta nota foi boa, mas não foi excepcional.

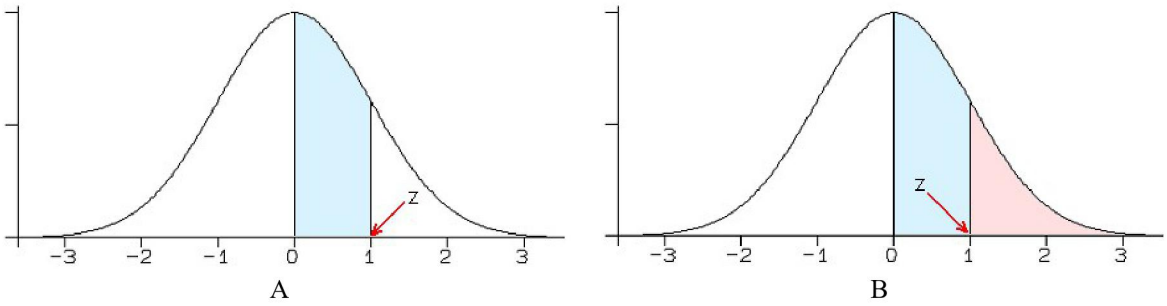


Figura 12. Curva normal padrão

Se o valor de Z encontrado foi igual z=2,86, a área entre este valor e o centro da distribuição é dada pelo cruzamento da linha e da coluna marcadas em azul na Tab. **1**.

$$P(0 < Z < 2{,}86) = 0{,}4979$$

A proporção de alunos que tiraram notas melhores do que esta será igual a;

$$P(Z > 2{,}86) = 0{,}5 - P(0 < Z < 2{,}86) = 0{,}5 - 0{,}4979 = 0{,}0021$$

Portanto, apenas 0,21 % dos alunos (cerca de 1 em 500) conseguiu nota melhor que esta; esta nota portanto foi excelente.

3.4.4.3. Por que o modelo normal é importante

O modelo normal é sem dúvida o mais importante da Estatística. Por que é tão importante? Basicamente, porque tem várias características que o tornam muito flexível, e permitem que ele sirva como base matemática de grande parte da Inferência Estatística (isto é, do trabalho estatístico feito com amostras).

(i) Porque tem propriedades matemáticas úteis

Mencionaremos a seguir duas destas propriedades:

> **(1)** A soma de variáveis independentes que tenham distribuições normais também é uma variável normal.

Vimos (seção 3.3.1.5) que a soma de duas variáveis X_1 e X_2 independentes resulta numa variável Y cujo valor esperado e variância podem ser calculados a partir dos valores esperados e variâncias de X_1 e X_2:

se $Y = X_1 + X_2 \rightarrow$ $\quad E(Y) = E(X_1) + E(X_2)$
$\quad V(Y) = V(X_1) + V(X_2)$

Estas propriedades servem para variáveis aleatórias que tenham qualquer distribuição; observe porém que não dizem nada sobre a *forma* da distribuição: a soma de duas variáveis que tenham uma certa distribuição geralmente é uma variável com distribuição diferente. Por exemplo, suponha que lançamos dois dados e somamos os números. O número mostrado por cada dado será uma variável com E(X)=3,5 e distribuição uniforme; a soma dos dois números será uma variável com E(X)=3,5+3,5=7, como prevê propriedade acima, mas sua distribuição não será uniforme (veja seção 3.1.3).

A distribuição normal é a única que tem esta propriedade: a soma de variáveis normais é sempre uma variável normal. Por que esta propriedade é importante? Porque podemos, a partir dela, tirar conclusões sobre a soma ou a média das variáveis numa amostra. Por exemplo, se o peso de um homem adulto é $X \sim N(\mu=75\text{ kg}, \sigma^2=100)$, o peso de um grupo aleatório de 6 homens será normal com parâmetros

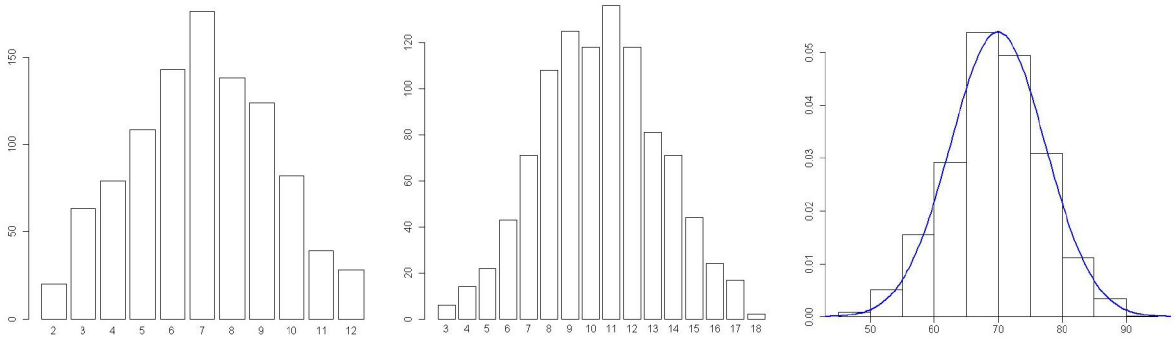
$$\mu_Y = 6\times 75 = 450 \text{ kg}$$
$$\sigma_Y^2 = 6\times 100 = 600 \rightarrow \sigma_Y = \sqrt{600} = 24{,}49 \text{ kg}$$

Um engenheiro que trabalha com elevadores pode a partir daí calcular, por exemplo, a probabilidade de um grupo de 6 homens que entram no elevador tenham um peso total de mais de 500 kg. Além disso, podemos tirar conclusões sobre a média da variável na amostra: se a soma dos pesos dos homens na amostra é normal, a média destes pesos também será normal (a média é a soma dividida pelo número de homens, o que é um reescalonamento da soma; o reescalonamento não altera a forma da distribuição). Esta propriedade da média de uma amostra é fundamental para a Inferência (veja seção 4.5.3).

> **(2)** A soma de *n* variáveis que tenham uma mesma distribuição qualquer (não necessariamente normal) tem uma distribuição que tende para a distribuição normal, desde que *n* seja grande.

Esta propriedade é afirmada pelo *Teorema do Limite Central*, o teorema básico da Inferência Estatística (seção 4.5.3.1). Por causa dela, podemos tirar conclusões a partir de amostras e fazer testes estatísticos sobre as populações .

A Fig. **13** mostra amostras simuladas (n=1000) da soma dos números obtidos no lançamento de dados: na Fig. A. a soma de dois dados; na Fig. B, a soma de três dados; na Fig. C, a soma de 20 dados. É possível ver que, à medida que aumenta o número de dados (isto é, o número de variáveis somadas), a distribuição da soma se aproxima de uma normal.



A. Dois dados B. Três dados C. Vinte dados

Figura 13. Distribuição simulada da soma dos números mostrados por 2, 3 e 20 dados

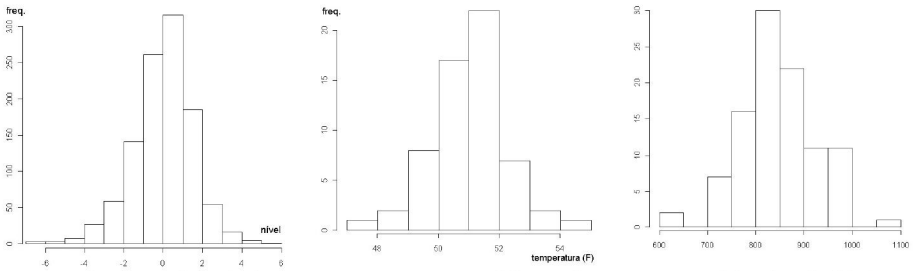
(ii) Porque serve de modelo para muitas variáveis da natureza e da tecnologia

Por causa das propriedades acima, o modelo normal pode ser usado para descrever a distribuição de uma variável que seja afetada pela adição de vários fatores independentes; se estes fatores têm a mesma distribuição, esta variável terá distribuição que tende para a normal. Variáveis assim são muito comuns nas ciências naturais e na tecnologia.

Um exemplo é a altura de pessoas adultas. A altura é determinada não apenas por fatores genéticos (os genes recebidos do pai e da mãe), mas também pela alimentação na infância, pela atividade física (ou falta dela), pelos acidentes ocorridos ou doenças graves que possam ter afetado o desenvolvimento, pelos hábitos de postura, etc. Podemos considerar que cada um destes fatores é uma variável aleatória e que a altura final de uma pessoa é resultado da soma de todos estes fatores; a altura terá portanto uma distribuição que tende para a normal.

Um outro exemplo é o da distribuição dos erros que ocorrem quando uma mesma quantidade é medida várias vezes com precisão, como ocorre por exemplo na Física. A primeira publicação do modelo normal, aliás, foi feita por Gauss em 1809 numa artigo sobre os erros em observações astronômicas. Gauss tinha analisado medições da posição do asteróide Ceres feitas por diversos observatórios europeus, e verificou que todas tinham erros e discordavam entre si. Estes erros podem ser devidos a vários fatores, como condições atmosféricas, qualidade do equipamento usado, características dos observadores, etc. Gauss considerou que a média aritmética das diversas medições deveria ser a melhor estimativa da posição real do asteróide, e notou que estas medições se distribuíam em torno da média de acordo com um gráfico que tinha o perfil de um sino. Usando as médias como estimativas das posições do asteróide no passado, foi capaz de prever sua posição alguns meses no futuro, previsão que foi confirmada pelas observações dos astrônomos.

Há um grande número de variáveis encontradas nas ciências que têm distribuições razoavelmente parecidas com a do modelo normal: distribuições unimodais, mais ou menos simétricas, com os intervalos mais prováveis ocorrendo perto da média. A Fig. **14** mostra três exemplos: (A) Variação anual do nível do Rio Negro, em Manaus; (B) Temperatura média anual na cidade de New Haven; (C) Medidas da velocidade da luz feitas por Michelson e Morley, em 1882.



A. Variação anual do nível do rio Negro, em Manaus

B. Temperatura média anual em New Haven, EUA (°F)

C. Medidas da velocidade da luz feitas em 1882.

Figura 14. Exemplos de distribuições aproximadamente simétricas e unimodais

Distribuições com formas similares a estas são encontradas com tanta freqüência que no início do século XIX os estatísticos começaram a pensar que esta deveria ser a forma "normal" de uma distribuição (daí o nome do modelo). Contudo, à medida em que aumentaram os recursos para a coleta e análise de dados esta idéia foi sendo abandonada (nos tempos de Gauss todas as contas e gráficos eram feitos à mão – para fazer os gráficos da Fig. **14**, por exemplo seria preciso organizar manualmente os dados em tabelas de distribuições de freqüência, depois desenhar os gráficos em papel – um processo que levaria horas, ou dias). Os estatísticos começaram a descobrir novas variáveis para as quais o modelo normal não servia.

Na verdade, é bom ter sempre em mente que nenhuma variável é *normal* – a distribuição normal é um modelo teórico, como os modelos da Geometria. Se formos a uma ambiente natural (uma floresta, por exemplo) não encontremos ali nenhum objeto que tenha formas iguais às dos modelos geométricos – nada que seja exatamente na forma de uma esfera, ou cone, ou mesmo de uma simples linha reta. Do mesmo modo, nenhuma variável real tem distribuição exatamente normal (ou exponencial, ou de Poisson, etc.). Modelos são construções teóricas que servem como descrições aproximadas da realidade; construções extremamente úteis, porque a partir delas as ciências podem tirar muitas conclusões importantes.

(iii) Porque serve de aproximação para outros modelos

Além de servir para descrever muitas variáveis encontradas nas ciências e na tecnologia, o modelo normal também é uma importante ferramenta para a aproximação de modelos de variáveis discretas, especialmente o *binomial* e o de *Poisson*, quando as amostras são grandes. A aproximação do modelo binomial pelo normal, aliás, é a base dos *testes de proporções* (seção 4.4). Veremos mais abaixo como estas aproximações são feitas (seção 3.4.4.5).

(iv) Porque está na base de várias técnicas estatísticas importantes

Várias técnicas estatísticas muito usadas são baseadas nas propriedades do modelo normal. Por exemplo, tanto os testes estatísticos de médias com amostras pequenas usando a *distribuição de Student*, quanto os testes comparativos de várias médias feitos pela *Análise de Variância*, partem do pressuposto de que a distribuição das variáveis seja normal. Além disso, há várias outras técnicas que se baseiam no pressuposto de que os *erros* de um modelo tenham distribuição normal; por exemplo, os modelos de regressão linear ou os modelos ARIMA para séries temporais. Por causa desta ampla gama de utilizações, o modelo normal ainda é a base de grande parte das aplicações da Estatística, mesmo que modelos mais específicos e mais complexos tenham sido desenvolvidos depois dele.

3.4.4.4. Seleção e ajuste de um modelo normal

Se estamos interessados em estudar as características de uma variável aleatória, ou calcular as probabilidades associadas a seus valores, precisamos de *ajustar* um modelo à distribuição desta variável. Este problema – a *modelagem* de uma variável – é muito importante na Estatística e será abordado na seção 5.6. Por enquanto, faremos aqui uma breve introdução ao assunto usando o modelo normal como exemplo, já que ele é o mais usado na prática.

Encontrar o modelo para uma variável implica em resolver dois problemas:

- escolher qual a família de modelos que possivelmente dará a melhor descrição da variável (esta etapa é em geral chamada de *identificação* do modelo)
- encontrar, dentro da família escolhida, qual o modelo particular que descreve a população com menor erro; isto é, encontrar os parâmetros que definem este modelo (esta etapa é chamada de *ajuste* do modelo).

Discutirmos abaixo estes dois problemas.

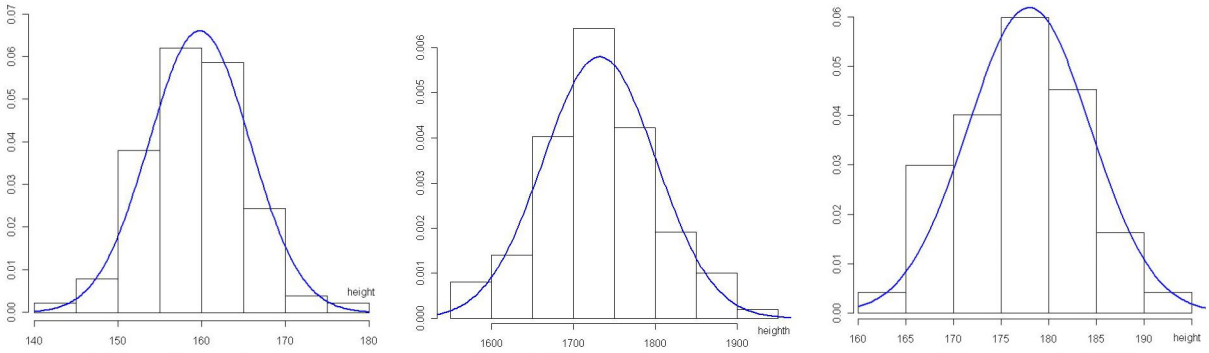
(i) Como identificar o modelo : histogramas e gráfico de quantis

A escolha de um modelo em qualquer ciência pode ser ditada por considerações teóricas deduzidas de conhecimento prévio, ou ser feita com base na observação empírica. Kepler, depois de experimentar com círculos e outras curvas, concluiu que o melhor modelo para a órbita dos planetas seria uma elipse, pois foi o modelo que se ajustou aos dados observados com menor erro. Newton, por outro lado, deduziu matematicamente que as trajetórias teriam que ser elipses, a partir de sua teoria da Gravitação.

Na Estatística, há situações em que o modelo é ditado por considerações teóricas. Quando fazemos Inferência com amostras pequenas (por exemplo, na estimação da média de uma população), sabemos que deve ser usado o modelo de Student, deduzido matematicamente a partir de alguns pressupostos. Na Análise de Variância, sabemos que devemos usar o modelo de Snedecor (F). Nestes dois casos, portanto, a escolha do modelo foi feita teoricamente. Contudo, tanto o modelo de Student quanto o de Snedecor partem de um mesmo pressuposto: o que a da distribuição da variável na população seja normal. Como podemos saber que o modelo normal é adequado para aquela variável?

Às vezes, o modelo pode ser *sugerido* por considerações teóricas. No Seção 3.4.4.1, por exemplo, dissemos que o modelo normal provavelmente seria uma escolha razoável para a variável "altura de homens adultos", porque a altura de uma pessoa é afetada por uma grande quantidade de fatores independentes, e o modelo normal em geral é bom para este tipo de problema. Esta é contudo apenas uma suposição (não sabemos realmente quais fatores estão influenciando a altura das pessoas em uma população), e teremos que confirmá-la empiricamente, por meio de análises de amostras tiradas da população.

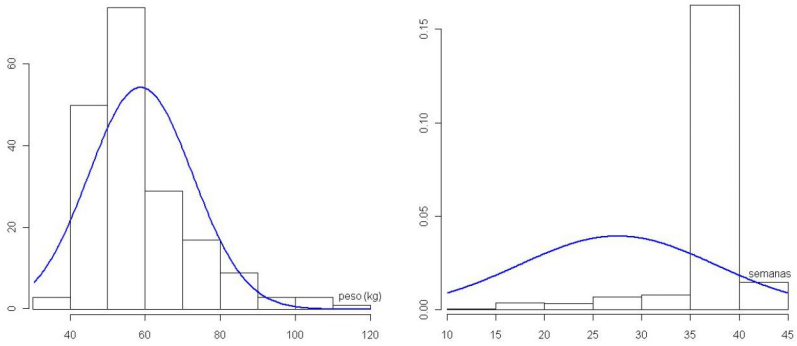
A primeira coisa a verificar é a forma do histograma da variável na amostra; este histograma deve dar uma indicação da forma do modelo adequado para a população. Se a distribuição da variável na população é coerente com o modelo normal, unimodal e simétrico, iremos esperar que o histograma também seja razoavelmente unimodal e simétrico. Usando o R, podemos sobrepor uma curva normal sobre o histograma, para facilitar a verificação visual, como na Fig. **15**.



A. 351 mulheres inglesas B. 199 homens ingleses C. 234 alunos da UFJF

Figura 15. Histogramas de três amostras, com as curvas normais sobrepostas

Nas três amostras, o modelo normal parece se ajustar bem aos dados do histograma. É importante enfatizar aqui. porém, que não estamos tentando modelar *a amostra*, e sim *a população* que esta amostra representa (esta é uma confusão que muitos alunos fazem). A amostra é conhecida, não precisa de modelos; a população é desconhecida, ou conhecida apenas parcialmente, e precisamos de modelos como ferramenta que nos permita fazer as contas e calcular as probabilidades. Um histograma representa a distribuição de freqüências (absolutas ou relativas) observadas numa amostra. Estas frequências relativas são estimativas das probabilidades; o histograma nos dá portanto estimativas da distribuição de probabilidades que existe na população; quanto maior a amostra, melhores serão estas estimativas.



A. Peso de mulheres adultas B. Duração de gestações, em semanas

Figura 16. Exemplos de variáveis para as quais o modelo normal não parece adequado

A Fig. **16** mostra exemplos de amostras de duas variáveis para as quais o modelo normal obviamente não parece adequado: (A) peso de 280 mulheres americanas entre 20 e 40 anos de idade, antes de engravidarem; (B) Tempo de duração de gestação destas mulheres. Estas amostras têm distribuições evidentemente assimétricas, o que não é coerente com o modelo normal.

Outra verificação que pode ser obtida por meio de gráficos é a fornecida pelos *gráficos de quantis* (*quantile plots*). Estes gráficos permitem que a distribuição observada numa amostra seja comparada visualmente com a de um modelo teórico qualquer; a função `qqnorm` do R compara a distribuição empírica na amostra com a distribuição teórica de um modelo normal.

Nestes gráficos, cada observação encontrada na amostra é representada pelos seus *quantis*: no eixo vertical, o seu percentil na distribuição empírica da amostra; no eixo hori-

zontal, o percentil que ela ocuparia na distribuição normal. Se estas duas distribuições coincidirem, ao pontos que representam as observações estarão distribuídos ao longo de uma linha reta com ângulo de 45º (y=x) no gráfico. A Fig. **17A** mostra o gráfico de quantis da distribuição de alturas cujo histograma está na Fig. **15A**, e indica que o modelo normal descreve bem o que foi encontrado na amostra, confirmando o que vimos no histograma. As Figs. **17B** e **17C** (peso de mulheres e duração de gestação) mostram os gráficos de quantis de variáveis para as quais o modelo normal não serve, confirmando a impressão obtida através dos histogramas da Fig. **16**.

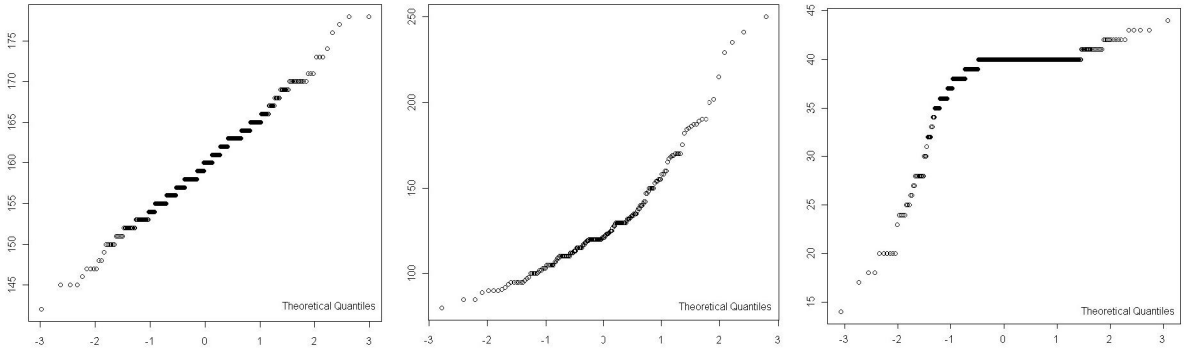


Figura 17. Três exemplos de gráfico de quantis

Outras verificações podem ser baseadas nas medidas obtidas nas amostras: o coeficiente de assimetria deve ser próximo de zero (indicando distribuição simétrica) e o de curtose próximo de 3 (ver seção 2.2.3). Por fim, há vários *testes de normalidade* – testes estatísticos que verificam qual a probabilidade de uma dada amostra ter vindo de uma população normal; este é um problema freqüente, se queremos usar testes de hipótese com amostras pequenas (testes t), modelos de Regressão Linear ou fazer Análise de Variância, etc. Na seção 5.6 veremos estes testes com mais detalhes.

(ii) Ajuste do modelo

Depois de escolhida a família de modelos a ser usada, é preciso encontrar, dentro desta família, aquele modelo particular que melhor pode representar a população. No caso do modelo normal, isto significa encontrar os valores dos parâmetros μ e σ que definem o modelo. Exemplos disto são os gráficos na Fig. **15**: foram usados modelos normais, mas em cada um deles os parâmetros (e mesmo as unidades de medida no eixo horizontal) são diferentes.

Este procedimento é chamado de *ajuste* do modelo. Há vários métodos para isto, e o mais comum é o que foi criado por Gauss, o método dos *mínimos quadrados ordinários, MQO* (*minimum square error, MSE*); aliás, foi para justificar matematicamente este método que Gauss criou o modelo normal.

O método MQO procura encontrar os valores de μ e σ que levem ao menor *erro quadrático*. "Erro" é a diferença entre o que o modelo prevê e o que foi encontrado numa amostra; no modelo para o peso dos alunos, por exemplo, a diferença entre o número de alunos que deveriam ser encontrados em cada faixa de peso, de acordo com o modelo, e o número que realmente foi encontrado na amostra. O método procura identificar os parâmetros *ótimos* do modelo, que são aqueles que levem ao menor total dos *quadrados* destes erros, somados para todas as faixas de peso (a razão de usarmos o *quadrado* dos erros é a mesma já explicada na seção sobre a *variância*, 2.2.2.4). No caso do modelo normal, o

procedimento é simples, pois pode ser demonstrado que os parâmetros ótimos são iguais à média aritmética e ao desvio-padrão da amostra.

(iii) Considerações finais

Estes tópicos – identificação e ajuste de modelos – é extremamente importante, e voltaremos a falar dele mais adiante. A maioria das técnicas de *Inferência Paramétrica* começam por identificar e ajustar um modelo para a variável que queremos estudar, e depois utilizar as propriedades deste modelo. Isto será visto a partir do Capítulo 4; por enquanto, há algumas observações importantes que devem ser feitas aqui.

Primeiro, o erro (entre modelo e amostra) sempre vai existir, e em cada amostra ele será diferente (se você lançar uma moeda 10 vezes não vai encontrar sempre 5 caras e 5 coroas). O melhor modelo é aquele que *minimiza* o erro (isto é, reduz o erro até o mínimo possível), mas nenhum modelo vai fazer o erro desaparecer.

Segundo, o erro sempre vai existir, mas até que ponto ele é aceitável? Se ele for grande demais, isto pode significar simplesmente que o modelo não serve, não descreve bem o que acontece na população. Veremos depois os *testes de normalidade*, testes estatísticos que procuram avaliar qual é a probabilidade de uma amostra ter sido produzida por um modelo normal; se esta probabilidade for pequena demais, isto quer dizer que o modelo normal não serve para esta variável.

Terceiro: é comum dizermos que "a população (ou a variável) é normal" ou a "população (ou a variável) segue a distribuição normal"; por exemplo, em enunciados de exercícios que começam dizendo "Numa população, a altura dos homens segue um modelo normal, com média 173 cm e desvio-padrão 10 cm. Qual é a probabilidade, etc.". Esta maneira de dizer não é rigorosamente correta. A população não tem que "seguir" o modelo; o modelo é que tem que se seguir a população, isto é, deve dar resultados que se aproximem o mais possível do que existe na população. Isto é apenas uma maneira simplificada e um tanto inexata de dizer que a variável que nos interessa tem uma distribuição de probabilidades para a qual o modelo normal teórico fornece uma boa descrição.

3.4.4.5. Uso do modelo normal como aproximação de outras distribuições

(i) Aproximação da distribuição binomial

Vimos na Seção 3.3.5 que o modelo binomial pode ser usado para calcular as probabilidades associadas a cada valor de X, em problemas nos quais a variável de interesse é o número de sucessos obtidos em *n* repetições de um experimento de resultados binários. A probabilidade de obtermos 7 ou mais caras em 10 lançamentos de uma moeda, por exemplo, pode ser calculada pelo somatório:

$$P(X \geq 7) = P(X=7) + P(X=8) + P(X=9) + P(X=10)$$

onde cada um do termos no membro direito da equação pode ser calculado pelo modelo binomial com parâmetros n=10 e p=0,5, da forma :

$$P(X=7) = p(7) = C_{10}^{7} \times 0{,}5^{7} \times 0{,}5^{(10-7)} = 0.1172$$

$$P(X=8) = p(8) = C_{10}^{8} \times 0{,}5^{8} \times 0{,}5^{(10-8)} = 0.0439, \text{ etc.}$$

Este tipo de problema ocorre porém com freqüência em aplicações nas quais o número *n* de repetições é muito grande; por exemplo, problemas como:

a) Um fabricante afirma que dentre os parafusos que fabrica apenas 3% têm algum tipo de defeito. Se isto for verdade, qual é a probabilidade de que num lote de 100 destes parafusos, escolhidos aleatoriamente, sejam encontrados 7 ou mais defeituosos?
b) Um modelo genético criado por Mendel prediz que, entre os descendentes produzidos no cruzamento de ervilhas, ¼ dos ervilhas produzidas sejam de cor verde, e ¾ sejam de cor amarela. Se esta previsão for verdade, qual é a probabilidade de que, entre 500 cruzamentos, sejam encontrados mais de 135 descendentes de cor verde?
c) Um partido político afirma que seu candidato conta com os votos de 60% dos eleitores. Se isto for verdade, qual é a probabilidade de que, numa amostra de 1000 destes eleitores, selecionados aleatoriamente, menos da metade deles afirme que vai votar neste candidato?

Estes problemas são exatamente iguais ao do lançamento das moedas, em termos conceituais, mas usam amostras muito maiores (n=100, 500 e 1000, respectivamente), o que dificulta os cálculos (porque o cálculo dos fatoriais de números grandes é impossível, e é preciso usar aproximações). Pode ser demonstrado que a distribuição normal dá uma boa aproximação dos resultados, se *n* for grande e o valor de *p* não muito pequeno. O modelo binomial tem valor esperado e variância dados pelas expressões:

$$E(X) = np$$

$$V(X) = np(1-p)$$

Este modelo binomial poderá ser então aproximado por uma normal que tenha a mesma média e a mesma variância, isto é, que tenha:

$$E(X) = \mu = np$$

$$V(X) = \sigma^2 = np(1-p) \rightarrow \sigma = \sqrt{np(1-p)}$$

No problema (c), por exemplo, o que nos interessa calcular é

$$P(X \leq 500)$$

Usando uma distribuição normal de parâmetros:

$$\mu = np = 1000 \times 0{,}6 = 600$$

$$\sigma = \sqrt{np(1-p)} = 22{,}36$$

Padronizado o valor de X desejado:

$$Z = \frac{X - \mu}{\sigma} = \frac{500 - 600}{22{,}36} = -4{,}47$$

A probabilidade de um valor de Z menor ou igual a este é muito pequena:

$$P(Z \leq -4{,}47) = 0{,}0000$$

O gráfico da Fig. **18** mostra a posição do valor de *z* encontrado na curva normal padrão. o que reforça a conclusão de que este é um valor de ocorrência extremamente improvável.

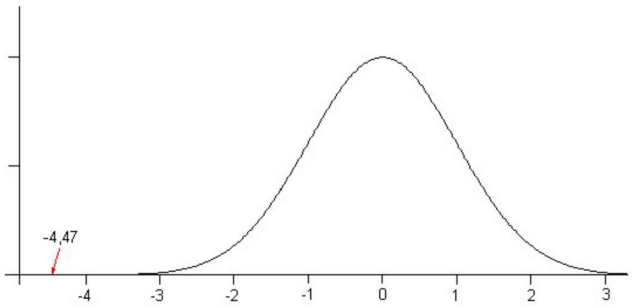


Figura 18. Posição do valor encontrado de z na curva normal padrão

Um raciocínio deste tipo de problema é a base dos *testes de hipótese*, que veremos a partir da seção 4.4. Neste exemplo, a afirmação do partido de que a porcentagem de eleitores favoráveis é de 60% (isto é, p=0,6) é considerada como uma *hipótese*, e o resultado encontrado na amostra é usado para *testar* esta hipótese. No exemplo, o modelo diz que se a hipótese for verdadeira, é praticamente impossível encontrar uma amostra onde menos da metade dos eleitores sejam favoráveis ao candidato. Se isto ocorreu na amostra que retiramos, chegaremos à conclusão de que a afirmação do partido provavelmente é falsa. As conclusões de um teste de hipótese nunca serão definitivas, porém, mas sempre baseadas em probabilidades; iremos mais tarde calcular as probabilidades de estas conclusões estarem certas, e de estarem erradas.

Esta aproximação da binomial à normal, à medida que o número *n* de repetições aumenta, pode ser demonstrada analiticamente. Esta demonstração não será feita aqui; em vez disto, mostramos nas Figs. **19** e **20** como o histograma das binomiais claramente tende para a forma de um modelo normal quando *n* aumenta, não apenas quando a binomial é simétrica (p=0,5), mas também quando ela é assimétrica ($p \neq 0{,}5$).

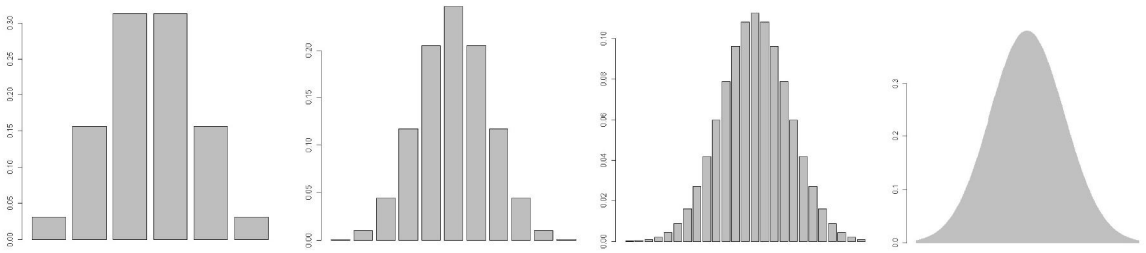


Figura 19. Distribuições binomiais de p=0,5 para diferentes números *n* de repetições

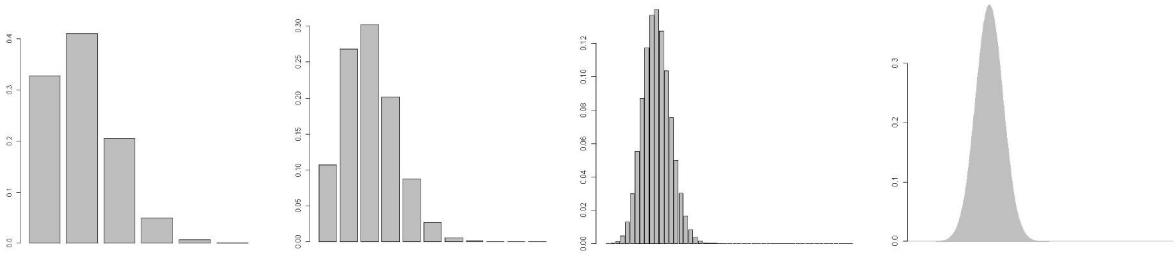


Figura 20. Distribuições binomiais de p=0,2 para diferentes números *n* de repetições

Exemplo

Suponha que um aluno dê palpites aleatórios em todas as 100 questões de uma prova de concurso (questões de múltipla escolha, com cinco alternativas cada). Qual é a probabilidade de que este aluno acerte mais de 30 questões ?

O que nos interessa calcular é P(X≥30). Se usássemos um modelo binomial, ele teria como parâmetros:

n=100	(número de questões da prova)
p=1/5=0,2	(probabilidade de acerto em cada questão)

donde o valor esperado e a variância do número de acertos seriam dados por:

$$E(X) = np = 100 \times 0{,}2 = 20 \qquad V(X) = np(1-p) = 16$$

Usando uma distribuição normal cujos parâmetros sejam o valor esperado e o desvio-padrão da binomial:

$$\mu = E(X) = 20 \qquad \sigma = \sqrt{V(X)} = 4$$

Padronizando o valor de X desejado:

$$Z = \frac{X - \mu}{\sigma} = \frac{30 - 20}{4} = 2{,}50$$

A probabilidade de um valor de Z maior ou igual a este é (consulte a tabela na seção 3.6.3):

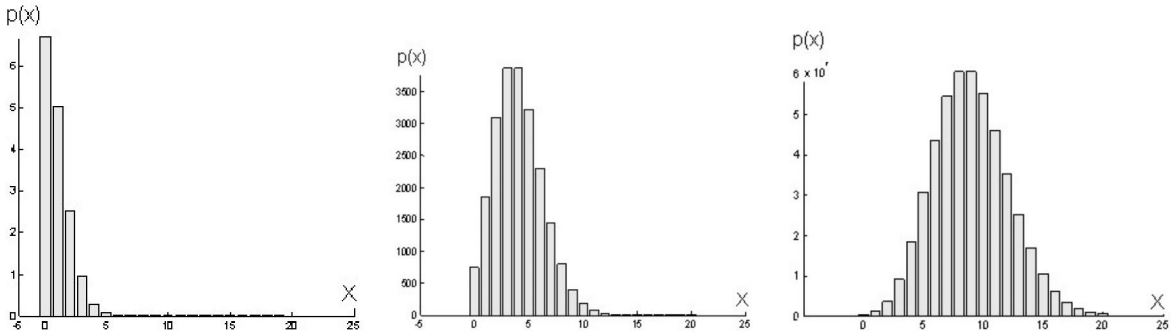
$$P(Z > 2{,}50) = 0{,}5 - 0{,}4938 = 0{,}0062$$

Note que, se a prova tivesse 10 questões, a probabilidade de um aluno ter a mesma proporção de acertos (≥30%) seria igual a P(X≥3) = 0,3222. Quanto mais questões tem a prova, portanto, mais difícil fica conseguir uma grande proporção de acertos dando apenas palpites aleatórios. (Por isso, provas de múltipla escolha geralmente têm grande número de questões, para garantir que a nota não tenha sido conseguida apenas por meio de palpites aleatórios...)

(ii) Aproximação da distribuição de Poisson

O modelo normal também pode ser usado como aproximação da distribuição de Poisson, quando a média desta é razoavelmente grande. Os gráficos da Fig. **21** mostram as funções de probabilidade de três modelos de Poisson. É possível ver que a distribuição já se torna praticamente simétrica quando λ=10, e que sua forma parece se aproximar de uma normal.

A aproximação do modelo de Poisson pela normal contudo é bem menos usada na prática do que a aproximação da binomial, porque o modelo de Poisson tipicamente é usado para modelar distribuições de eventos raros, com números médios de ocorrências muito pequenos.



λ=1,5 λ=5 λ=10

21. Distribuições de Poisson para diferentes λ