UNIDADE

7

Modelos probabilísticos mais comuns

Objetivo

Nesta Unidade, você vai conhecer os modelos probabilísticos mais importantes para variáveis aleatórias discretas e contínuas. Você aprenderá a identificar as situações reais em que podem ser usados para o cálculo de probabilidades e a importância disso para o administrador.

Modelos probabilísticos para variáveis aleatórias discretas

Nas Unidades 5 e 6, vimos os conceitos gerais de probabilidade e variáveis aleatórias: podemos construir um modelo probabilístico do zero para um problema de administração, a partir de dados históricos ou experimentais.

Embora plenamente possível, o processo de construção de um modelo probabilístico do zero pode ser bastante longo: é preciso coletar os dados (ver Unidades 1 e 2), fazer a análise exploratória deles (ver Unidades 3 e 4), obter as probabilidades e validar o modelo. Mesmo tomando todos os cuidados, muitas vezes vamos reinventar a roda e correndo o risco de ela sair quadrada...

Por que não usar os conhecimentos prévios desenvolvidos ao longo de centenas de anos de pesquisa e experimentação? Vamos procurar, dentre os vários modelos probabilísticos existentes, aquele mais apropriado para o fenômeno que estamos estudando, que é materializado através de variáveis aleatórias.

Através da Análise Exploratória de Dados, podemos avaliar qual modelo é mais apropriado para os nossos dados. Contudo, para fazer isso precisamos conhecer tais modelos.

Nesta Unidade, vamos estudar os modelos mais usados para variáveis aleatórias discretas (binomial e Poisson) e para variáveis aleatórias contínuas (uniforme, normal, t e qui-quadrado).

Aqui é importante avaliar com cuidado a **variável aleatória*** discreta.

É preciso identificar se o **espaço amostral** é **finito*** ou **infinito numerável***: alguns modelos são apropriados para um caso, e não para o outro.

GLOSSÁRIO

*Variável aleatória – é uma função matemática que associa números reais aos resultados de um espaço amostral, por sua vez, vinculado a um experimento aleatório. Fonte: Barbetta, Reis e Bornia (2004).

*Espaço amostral finito – é aquele formado por um número limitado de resultados possíveis. Fonte: Barbetta, Reis e Bornia (2004).

*Espaço amostral infinito numerável – é aquele formado por um número infinito de resultados, mas que podem ser listados. Fonte: Barbetta, Reis e Bornia (2004).

Vamos ver os dois modelos mais importantes: binomial e Poisson.

Modelo binomial

Seja um experimento aleatório qualquer que apresenta as seguintes características:

- consiste na realização de um número finito e conhecido **n** de ensaios (ou repetições);
- cada um dos ensaios tem apenas dois resultados possíveis: "sucesso" ou "fracasso" (estão entre aspas, porque a definição de sucesso não quer necessariamente algo "positivo", e também porque poderá significar um grupo de resultados); e
- os ensaios são independentes entre si, apresentando probabilidades de "sucesso" (**p**) e de "fracasso" (**1–p**) constantes.

GLOSSÁRIO
*Variável aleatória discreta – o espaço amostral ao qual ela está associada é finito ou infinito numerável. Fonte: Barbetta, Reis e Bornia (2004).

Neste caso, estamos interessados no número de "sucessos" obtidos nos **n** ensaios: como o espaço amostral é finito (vai de 0 a **n**), uma variável aleatória associada seria discreta. Este tipo de experimento é chamado de binomial.

Então, a variável aleatória discreta* X, número de "sucessos" nos **n** ensaios, apresenta uma distribuição (modelo) binomial com os seguintes parâmetros:

n = número de ensaios **p** = probabilidade de "sucesso"

Com esses dois parâmetros, é possível calcular as probabilidades de um determinado número de sucessos, bem como obter o valor esperado e a variância da variável X:

$$E(X) = n \times p \qquad V(X) = n \times p \times (1-p)$$

Exemplo 1: experimentos binomiais:

a) observar o número de caras em três lançamentos imparciais de uma moeda honesta: n=3; p=0,5;

b) observar o número de meninos nascidos em três partos de uma família: n=3; p = x; e

c) observar o número de componentes defeituosos em uma amostra de dez componentes de um grande número de peças que apresentaram anteriormente 10% de defeituosos: n = 10; p= 0,1.

Vamos ver com maiores detalhes o caso do número de meninos (e meninas) nascidos em uma família. Chamando menino de evento H, será o "sucesso", e menina de evento M, e sabendo pela história da família que P(H) = 0,52 e P(M) = 0,48 (então **p** = 0,52 e **1– p** = 0,48), quais serão as probabilidades obtidas para a variável aleatória número de meninos em três nascimentos? Vamos obter a distribuição de probabilidades.

Usando os conceitos gerais de probabilidade, é preciso primeiramente determinar o espaço amostral, como poderão ser os sexos das três crianças:

$$\mathbf{\Omega = \{H \cap H \cap H,\ H \cap H \cap M,\ H \cap M \cap H,\ M \cap H \cap H,\ H \cap M \cap M,\ M \cap H \cap M,\ M \cap M \cap H,\ M \cap M \cap M\}}$$

Supondo que os nascimentos sejam independentes, podemos calcular as probabilidades de cada intersecção simplesmente multiplicando as probabilidades individuais de seus componentes:

$$P\{H \cap H \cap H\} = P(H) \text{ x } P(H) \text{ x } P(H) = p \text{ x } p \text{ x } p = p^3$$

$$P\{H \cap H \cap M\} = P(H) \text{ x } P(H) \text{ x } P(M) = p \text{ x } p \text{ x } (1-p) = p^2\,(1-p)$$

$$P\{H \cap M \cap H\} = P(H) \text{ x } P(M) \text{ x } P(H) = p \text{ x } (1-p) \text{ x } p = p^2 \text{ x } (1-p)$$

$$P\{M \cap H \cap H\} = P(M) \text{ x } P(H) \text{ x } P(H) = (1-p) \text{ x } p \text{ x } p = p^2 \text{ x } (1-p)$$

$$P\{H \cap M \cap M\} = P(H) \text{ x } P(M) \text{ x } P(M) = p \text{ x } (1-p) \text{ x } (1-p) = p \text{ x } (1-p)^2$$

$$P\{M \cap H \cap M\} = P(M) \text{ x } P(H) \text{ x } P(M) = (1-p) \text{ x } p \text{ x } (1-p) = p \text{ x } (1-p)^2$$

$$P\{M \cap M \cap H\} = P(M) \text{ x } P(M) \text{ x } P(H) = (1–p) \text{ x } (1–p) \text{ x } p = p \text{ x } (1–p)^2$$

$$P\{M \cap M \cap M\} = P(M) \text{ x } P(M) \text{ x } P(M) = (1–p) \text{ x } (1–p) \text{ x } (1– p) = (1–p)^3$$

Observe que:

$P\{H \cap H \cap M\} = P\{H \cap M \cap H\} = P\{M \cap H \cap H\} = p^2 \text{ x } (1 – p)$ = Probabilidade de dois "sucessos"

$P\{H \cap M \cap M\} = P\{M \cap H \cap M\} = P\{M \cap M \cap H\} = p \text{ x } (1 – p)^2$ = Probabilidade de um "sucesso"

Importa apenas a "natureza" dos sucessos, não a ordem em que ocorrem: com a utilização de **combinações**, é possível obter o número de resultados iguais para cada número de sucessos. Supondo que o número de ensaios **n** é o número de "objetos" disponíveis, e que o número de "sucessos" em que estamos interessados (doravante chamado **k**) é o número de "espaços" onde colocar os objetos (um objeto por espaço), o número de resultados iguais será:

$$C_{n,k} = \frac{n!}{k! \times (n-k)!}$$

Para o caso acima, em que há três ensaios (**n** =3):

- para dois sucessos (**k** =2) $C_{3,2} = \frac{3!}{2! \times (3-2)!} = 3$ (o mesmo resultado obtido por enumeração);
- para um sucesso (**k** =1) $C_{3,1} = \frac{3!}{1! \times (3-1)!} = 3$ (o mesmo resultado obtido por enumeração).

O procedimento acima poderia ser feito para quaisquer valores de **n** e **k** (desde que **n** $\geq$ **k**), permitindo obter uma expressão geral para calcular a probabilidade associada a um resultado qualquer.

A probabilidade de uma variável aleatória discreta X, número de sucessos em **n** ensaios, com distribuição binomial de parâmetros **n** e **p**, assumir um certo valor **k** ($0 \leq$ **k** $\leq$ **n**) será:

$$P(X = k) = C_{n,k} \times p^k \times (1-p)^{n-k}, \text{ onde } C_{n,k} = \frac{n!}{k! \times (n-k)!}$$

É importante lembrar que a probabilidade de ocorrer **k** sucessos é igual à probabilidade de ocorrer **n – k** fracassos, e que todos os axiomas e propriedades de probabilidade continuam válidos.

Neste segundo exemplo, admitamos que a probabilidade de que uma companhia não entregue seus produtos no prazo é igual a 18%. Quais são as probabilidades de que, em três entregas, uma, duas ou todas as três entregas sejam feitas no prazo? Calcular também valor esperado, variância e desvio-padrão do número de entregas no prazo.

Para cada entrega ("ensaio"), há apenas dois resultados: no prazo ou não. Há um número limitado de realizações, **n** = 3. Definindo "sucesso" como no prazo, e supondo as operações independentes, a variável aleatória X, número de entregas no prazo em três terá distribuição binomial com parâmetros

n = 3 e **p** = 0,82 (e **1 – p** = 0,18).

Então:

$$P(X=0)=C_{3,0}\times 0{,}82^{0}\times(0{,}18)^{3}=\frac{3!}{0!\times(3-0)!}\times 0{,}82^{0}\times(0{,}18)^{3}=0{,}006$$

$$P(X=1)=C_{3,1}\times 0{,}82^{1}\times(0{,}18)^{2}=\frac{3!}{1!\times(3-1)!}\times 0{,}82^{1}\times(0{,}18)^{2}=0{,}080$$

$$P(X=2)=C_{3,2}\times 0{,}82^{2}\times(0{,}18)^{1}=\frac{3!}{2!\times(3-2)!}\times 0{,}82^{2}\times(0{,}18)^{1}=0{,}363$$

$$P(X=3)=C_{3,3}\times 0{,}82^{3}\times(0{,}18)^{0}=\frac{3!}{3!\times(3-3)!}\times 0{,}82^{3}\times(0{,}18)^{0}=0{,}551$$

Somando todas as probabilidades, o resultado é igual a 1, como teria que ser. O valor esperado, variância e o desvio-padrão serão:

Lembre-se que a soma das probabilidades de todos os eventos que compõem o espaço amostral é igual a 1. E que 0! = 1, e um número diferente de 0 elevado a zero é igual a 1.

E(X) = n x p = 3 x 0,82 = 2,46 entregas

V(X) = n x p x (1 – p) = 3 x 0,82 x 0,18 = 0,4428 entrgas².

$$\sigma(X)=\sqrt{V(X)}=\sqrt{0{,}4428}=0{,}665 \text{ entregas}$$

A média é quase igual ao número de operações, devido à alta probabilidade de sucesso.

Estudos anteriores mostraram que há 73% de chance de consumidores do sexo feminino apresentarem uma reação positiva a anúncios publicitários com crianças. Uma agência está conduzindo um estudo, apresentando um novo anúncio para cinco consumidoras. Vamos ver nesse Exemplo 3 qual é a probabilidade de que pelo menos três das cinco consumidoras apresentem reação positiva? Calcular também o valor esperado, a variância e o desvio-padrão do número de consumidoras que apresentam reação positiva.

Para cada consumidora ("ensaio"), há apenas dois resultados: reação positiva ou não. Há um número limitado de realizações, **n** = 5. Definindo "sucesso" como reação positiva, e supondo as consumidoras "independentes", a variável aleatória X, número de consumidoras com reação positiva em cinco que assistiram ao novo anúncio terá distribuição binomial com parâmetros

n = 5 e **p** = 0,73 (e **1 – p** = 0,27).

O evento de interesse é a recuperação de pelo menos três ratos (3 ou mais): **P(X ≥ 3)**.

$$\mathbf{P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5)}$$

É preciso calcular as três probabilidades acima e somá-las, então:

$$P(X=3) = C_{5,3} \times 0{,}73^3 \times (0{,}27)^2 = \frac{5!}{3! \times (5-3)!} \times 0{,}73^3 \times (0{,}27)^2 = 0{,}284$$

$$P(X=4) = C_{5,4} \times 0{,}73^4 \times (0{,}27)^1 = \frac{5!}{4! \times (5-4)!} \times 0{,}73^4 \times (0{,}27)^1 = 0{,}383$$

$$P(X=5) = C_{5,5} \times 0{,}73^5 \times (0{,}27)^0 = \frac{5!}{5! \times (5-5)!} \times 0{,}73^5 \times (0{,}27)^0 = 0{,}207$$

$$\mathbf{P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5) = 0{,}284 + 0{,}383 + 0{,}207 = 0{,}874}$$

A probabilidade de que pelo menos três das cinco consumidoras apresentem reação positiva é igual a 0,874 (87,4%).

Há duas outras formas de chegar ao mesmo resultado:

- através do complementar:

$$\mathbf{P(X \geq 3) = 1 - P(X<3) = 1 - [\ P(X=0) + P(X=1) + P(X=2)];}$$

- mudando a definição de sucesso, de reação positiva para **reação negativa** (**p** = 0,27), se pelo menos três consumidoras apresentam reação positiva, então, no máximo duas apresentam reação negativa.

Mas se o espaço amostral fosse infinito numerável? Teríamos que usar o modelo de Poisson. Você conhece este modelo? Sabe como tirar proveito de suas facilidades? Vamos estudar juntos para aprender ou para relembrar!

Modelo de Poisson

Vamos supor um experimento binomial, com apenas dois resultados possíveis, mas com a seguinte característica: apesar de a probabilidade **p** ser constante, o valor de **n** teoricamente é infinito.

Na situação acima, o modelo binomial não poderá ser utilizado. Nestes casos, deve ser utilizado o modelo de Poisson.

Como seria a solução para o caso acima?

Como **n** é "infinito", deve-se fazer a análise das ocorrências em um período contínuo (de tempo, de espaço, entre outros) subdividido em um certo número de subintervalos (número tal que a probabilidade de existir mais de uma ocorrência em uma subdivisão é desprezível, e supondo ainda que as ocorrências em subdivisões diferentes são independentes); novamente, é preciso trabalhar com uma quantidade constante, que será chamada de **m** também:

$m = \lambda \times t$

onde λ é uma taxa de ocorrência do evento em um período contínuo (igual ou diferente do período sob análise), e **t** é justamente o período contínuo sob análise.

Apesar do símbolo t, o período contínuo não é necessariamente um intervalo de tempo.

Como obter a taxa λ? Há duas opções: realizar um número suficiente de testes de laboratório para obter a taxa de ocorrência do evento a partir dos resultados, ou observar dados históricos e calcular a taxa.

Se uma variável aleatória discreta X, número de ocorrências de um evento, segue a distribuição de Poisson, a probabilidade de X assumir um valor **k** será:

$$P(X = k) = \frac{e^{-m} \times m^k}{k!}$$

Onde **e** é uma constante: **e** $\cong$ 2,71. E m = n x p ou m = λ x t.

Uma particularidade interessante da distribuição de Poisson é que o valor esperado e a variância de uma variável aleatória que siga tal distribuição serão iguais:

$E(X) = m = \lambda \times t$

$V(X) = m = \lambda \times t$

O modelo de Poisson é muito utilizado para modelar fenômenos envolvendo filas: filas de banco, filas de mensagens em um servidor, filas de automóveis em um cruzamento.

Vejamos neste Exemplo 4 os experimentos e fenômenos que seguem a distribuição de Poisson.

a) Número mensal de acidentes de tráfego em um cruzamento. Observe que é uma variável aleatória discreta, pode assumir apenas valores inteiros (0, 1, 2, 3,...). Cada realização do "experimento" (acidente) pode ter apenas dois resultados: ocorre o acidente ou não ocorre o acidente. Mas o número máximo de realizações é desconhecido! Assim, a distribuição binomial não pode ser usada, e a análise do número de acidentes precisa ser feita em um período contínuo (no caso, período de tempo, um mês), exigindo o uso da distribuição de Poisson.

b) Número de itens defeituosos produzidos por hora em uma indústria.
Novamente, uma variável aleatória discreta (valores inteiros: 0,1, 2, 3, ...). Cada realização só pode ter dois resultados possíveis (peça sem defeito ou peça defeituosa). Se o número máximo de realizações for conhecido, provavelmente a probabilidade de uma peça ser defeituosa será reduzida, e apesar de ser possível a utilização da distribuição binomial, o uso da distribuição de Poisson obterá resultados muitos próximos. Se o número máximo de realizações for desconhecido, a distribuição binomial não pode ser usada, e a análise do número de acidentes precisa ser feita em um período contí-

nuo (no caso, período de tempo, uma hora), exigindo o uso da distribuição de Poisson.

c) Desintegração dos núcleos de substâncias radioativas: contagem do número de pulsações radioativas a intervalos de tempo fixos.
Situação semelhante à dos acidentes em um cruzamento, só que o "grau de aleatoriedade" deste experimento é muito maior. O número máximo de pulsações também é desconhecido, obrigando a realizar a análise em um período contínuo, utilizando a distribuição de Poisson.

Neste Exemplo 5, uma telefonista recebe cerca de 0,20 chamadas por minuto (valor obtido de medições anteriores).

a) Qual é a probabilidade de receber exatamente cinco chamadas nos primeiros dez minutos?

b) Qual é a probabilidade de receber até duas chamadas nos primeiros 12 minutos?

c) Qual é o desvio-padrão do número de chamadas em meia hora?

Há interesse no número de chamadas ocorridas em um período contínuo (de tempo no caso). Para cada "ensaio", há apenas dois resultados possíveis: a chamada ocorre ou não. Observe que não há um limite para o número de chamadas no período (sabe-se apenas que o número mínimo pode ser 0): por esse motivo, a utilização da binomial é inviável. Contudo, há uma taxa de ocorrência (λ = 0,20 chamadas/minuto), e isso permite utilizar a distribuição de Poisson.

a) Neste caso, o período **t** será igual a 10 minutos (**t** = 10 min.), P(X = 5)?

m = λ x t = 0,20 x 10 = 2 chamadas

$$P(X = k) = \frac{e^{-m} \times m^k}{k!} = P(X = 5) = \frac{e^{-2} \times 2^5}{5!} = 0{,}0361$$

Então, a probabilidade de que a telefonista receba exatamente cinco chamadas em dez minutos é igual a 0,0361 (3,61%).

b) Neste caso, o período **t** será igual a 12 minutos (**t** = 12 minutos). O evento de interesse é até duas chamadas em 12 minutos ($X \leq 2$).

m = λ x t = 0,20 x 12 = 2,4 chamadas

P(X ≤ 2) = P(X = 0) + P(X = 1) + P(X = 2)

$$P(X=0) = \frac{e^{-2,4} \times 2,4^{0}}{0!} = 0,0907$$

$$P(X=1) = \frac{e^{-2,4} \times 2,4^{1}}{1!} = 0,2177$$

$$P(X=2) = \frac{e^{-2,4} \times 2,4^{2}}{2!} = 0,2613$$

P(X ≤ 2) = P(X = 0) + P(X = 1) + P(X = 2) = 0,0907 + 0,2177 + 0,2613 = 0,5697

Então, a probabilidade de que a telefonista receba até duas chamadas em 12 minutos é igual a 0,5697 (56,97%).

c) Neste caso, o período **t** será igual a 30 minutos (**t** = 30 minutos). Primeiro, calcula-se a variância:

V(X) = m = λ x t = 0,2 x 30 = 6 chamadas^2

O desvio-padrão é a raiz quadrada positiva da variância:

$$\sigma(X) = \sqrt{V(X)} = \sqrt{6} \cong 2,45 \text{ chamadas}$$

Há vários outros modelos para variáveis aleatórias discretas: hipergeométrico, geométrico, binomial negativo.

Na próxima seção, vamos ver os principais modelos de variáveis aleatórias contínuas.

Modelos para variáveis aleatórias contínuas

Nesta seção, estudaremos os modelos uniforme, normal, t e qui-quadrado.

Modelo uniforme

Quando o espaço amostral associado a um experimento aleatório é infinito, torna-se necessário o uso de uma variável aleatória contínua para associar números reais aos resultados. Os modelos probabilísticos vistos anteriormente não podem ser empregados: a probabilidade de que uma variável aleatória contínua assuma exatamente um determinado valor é **zero**.

Para entender melhor a declaração acima, vamos relembrar a definição clássica de probabilidade: a probabilidade de ocorrência de um evento será igual ao quociente entre o número de resultados associados ao evento pelo número total de resultados possíveis. Ora, se o número total de resultados é infinito ou tende ao infinito, para ser mais exato, a probabilidade de ocorrência de um valor específico é igual a zero. Por esse motivo, quando se lida com variáveis aleatórias contínuas, calcula-se a probabilidade de ocorrência de eventos formados por intervalos de valores, através de uma função densidade de probabilidades (ver Unidade 6). Uma outra conseqüência disso é que os símbolos $>$ e $\geq$ ($<$ e $\leq$ também) são equivalentes para variáveis aleatórias contínuas.

O modelo mais simples para variáveis aleatórias contínuas é o uniforme.

Seja uma variável aleatória contínua qualquer X que possa assumir valores entre A e B. Todos os valores entre A e B têm a mesma probabilidade de ocorrer, resultando no gráfico apresentado na Figura 50:

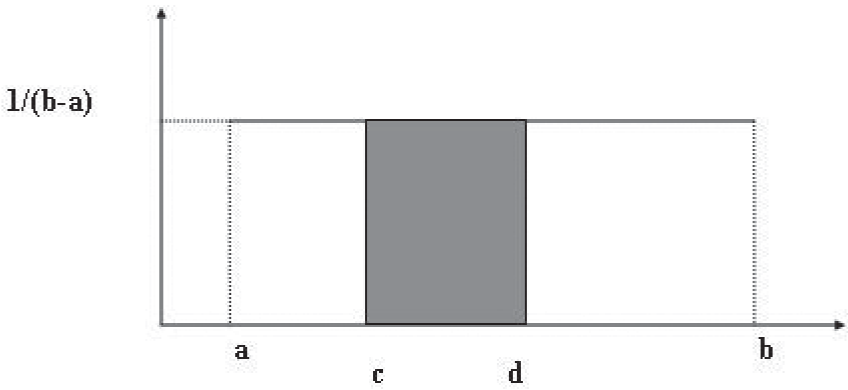

Figura 50: Modelo uniforme
Fonte: elaborada pelo autor

Para que a área entre a e b seja igual a 1, o valor da ordenada precisa ser igual a **1/(b – a)**, constante, portanto, para todo o intervalo. A área escura representa a probabilidade de a variável X assumir valores no intervalo **c – d**. Trata-se do modelo uniforme.

Dois intervalos de valores da variável aleatória contínua, que tenham o mesmo tamanho, têm a mesma probabilidade de ocorrer (desde que dentro da faixa de valores para os quais a função de densidade de probabilidades não é nula). Formalmente, uma variável aleatória contínua X tem distribuição uniforme, com parâmetros **a** e **b** reais (sendo **a** menor do que **b**), se sua função densidade de probabilidades for tal como a da Figuras 50.

A probabilidade de que a variável assuma valores entre c e d (sendo a < c < d < b) é a área compreendida entre c e d:

$$P(c < X < d) = (d - c) \times \frac{1}{(b - a)}$$

Seu valor esperado e a variância são:

$$E(X) = \frac{a + b}{2} \qquad V(X) = \frac{(b - a)^2}{12}$$

Intuitivamente, podemos supor que muitas variáveis aleatórias contínuas terão um comportamento diferente do caso acima: em algumas delas, haverá maior probabilidade de ocorrências de valores próximos ao limite inferior ou superior: para cada caso, deverá ser ajustado um modelo probabilístico contínuo adequado.

O modelo uniforme é bastante usado para gerar números pseudoaleatórios em processos de amostragem probabilística.

Neste Exemplo 6, a temperatura T de destilação do petróleo é crucial para determinar a qualidade final do produto. Suponha que T seja considerada uma variável aleatória contínua com distribuição uniforme de 150 a 300° C, e que o custo para produzir um galão de petróleo seja de 50 u.m. Se o óleo é destilado a menos de 200° C, o galão é vendido a 75 u.m. Se a temperatura for superior a 200° C, o produto é vendido a 100 u.m.

Adaptado de BUSSAB, W. O.; MORETTIN, P. A. *Estatística Básica*. 4. ed. São Paulo: Atual, 1987.

a) Fazer o gráfico da função densidade de probabilidade de T.

b) Qual é o lucro médio esperado por galão?

a) Os parâmetros a e b definem completamente uma distribuição uniforme; para fazer o gráfico, basta encontrá-los no enunciado acima. Identifica-se que o limite inferior, a, vale 150° C, e o superior, b, vale 300° C, resultando no gráfico a seguir:

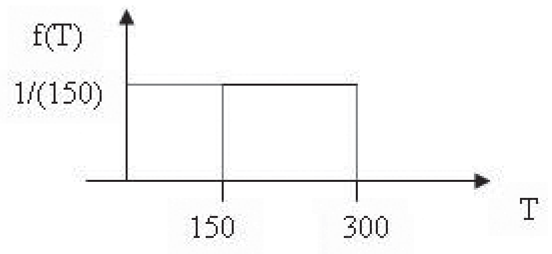

Figura 51: Temperatura de destilação do petróleo
Fonte: elaborada pelo autor

b) A variável aleatória de interesse, lucro, é discreta, somente pode assumir dois valores: 25 u.m. (caso o óleo seja destilado a menos de 200° C, posto que o galão custa 50 u.m. para ser produzido e será vendido a 75 u.m. nestas condições), ou 50 u.m. (caso o óleo seja destilado a mais de 200° C, posto que o galão custa 50 u.m. para ser produzido e será vendido a 100 u.m.). Sendo assim, seus valores possíveis serão: {25, 50}, sendo os resultados mutuamente exclusivos.

Lembrando das definições de distribuições de probabilidades, e de valor esperado e variância para variáveis aleatórias discretas (Unidade 6), para obter o lucro médio (valor esperado da variável lucro), é preciso obter as probabilidades de ocorrência dos seus dois valores (25 e 50). Relacionando com os valores de T:

$P(Lucro = 25) = P(T \leq 200)$ $\quad$ $P(Lucro = 25) = P(T > 200)$

Os valores das probabilidades acima correspondem às áreas abaixo da curva da função densidade de probabilidades para cada intervalo, calculando as áreas:

$$P(T \leq 200) = (200 - 150) \times \frac{1}{(300 - 150)} = \frac{50}{150}$$

$$P(T > 200) = (300 - 200) \times \frac{1}{(300 - 150)} = \frac{100}{150}$$

Então, a distribuição de probabilidades da variável lucro será (Quadro 22):

Lucro	Probabilidade
25	50/150
50	100/150
Total	1,0

Quadro 22: Distribuição de probabilidades da variável lucro
Fonte: elaborado pelo autor

Calculando o valor esperado:

$$E(Lucro) = \sum Lucro_i \times P(Lucro_i) \qquad E(Lucro) = 25 \times \frac{50}{150} + 50 \times \frac{100}{150} = 41{,}67 \text{ u.m.}$$

O lucro médio é de 41,67 u.m. Repare que a variável lucro não pode assumir este valor, o que significa que o valor esperado (a média) não é o valor mais provável. Neste problema, o valor mais provável, a moda (ver Unidade 4), vale 50 u.m., pois tem a maior probabilidade de ocorrência (66,67%).

Agora, vamos passar ao modelo mais importante para variáveis aleatórias contínuas.

Modelo normal

Há casos em que há maior probabilidade de ocorrência de valores situados em intervalos centrais da função densidade de probabilidades da variável aleatória contínua, e esta probabilidade diminui à medida que os valores se afastam deste centro (para valores menores ou maiores). O modelo probabilístico contínuo mais adequado talvez seja o modelo normal ou gaussiano.

O matemático alemão Gauss utilizou amplamente este modelo no tratamento de erros experimentais, embora não tenha sido o seu "descobridor".

Isso é especialmente encontrado em variáveis biométricas, resultantes de medidas corpóreas em seres vivos.

O modelo normal é extremamente adequado para medidas numéricas em geral, descrevendo vários fenômenos e permitindo fazer aproximações de modelos discretos. É extremamente importante também para a Estatística Indutiva (mais detalhes na próxima Unidade). O gráfico da distribuição de probabilidades de uma variável aleatória contínua que siga o modelo normal (distribuição normal) será como a Figura 52:

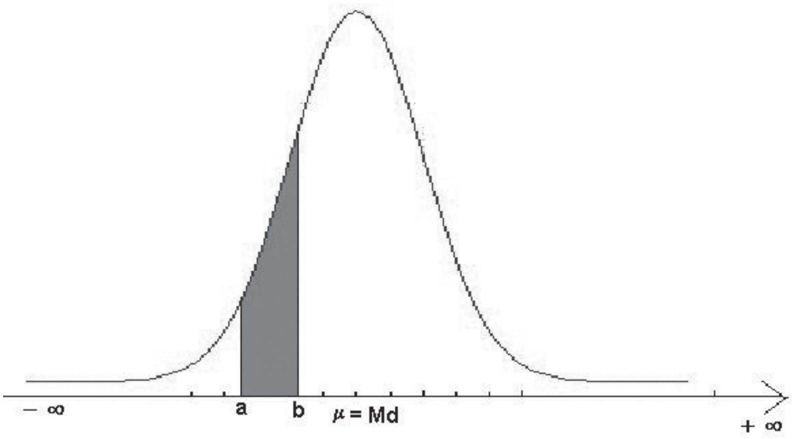

Figura 52: Distribuição normal
Fonte: elaborada pelo autor

Características do modelo normal0:

- a curva apresenta forma de sino, há maior probabilidade de a variável assumir valores próximos do centro;
- os valores de média (**μ**) e de mediana (**Md**) são iguais, significando que a curva é simétrica em relação à média;
- teoricamente, a curva prolonga-se de $-\propto$ a $+\propto$ (menos infinito a mais infinito), então a área total sob a curva é igual a 1 (100%);
- qualquer distribuição normal é perfeitamente especificada por seus parâmetros média (**μ**) e variância (σ^2) => **X: N (μ , σ²)** significa que a variável X tem distribuição normal com média **μ** e variância σ^2;
- a área escura na Figura 3 é a probabilidade de uma variável que siga a distribuição normal assumir valores entre **a** e **b**: esta área é calculada através da integral da função normal de **a** a **b**;
- cada combinação **(μ , σ²)** resulta em uma distribuição normal diferente; portanto, há uma família infinita de distribuições; e
- a função normal citada acima tem a seguinte (e aterradora...) fórmula para sua função densidade de probabilidade:

É comum a utilização de letras do alfabeto grego para representar algumas medidas. Não se esqueça que o desvio-padrão ( ) é a raiz quadrada positiva da variância.

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

Saiba que não existe solução analítica para uma integral da expressão acima: qualquer integral precisa ser resolvida usando métodos numéricos de integração, que são extremamente trabalhosos quando implementados manualmente (somente viáveis se são usados meios computacionais). De Moivre, Laplace e Gauss desenvolveram seus trabalhos entre a metade do século XVIII e início do século XIX, e os computadores começaram a se popularizar a partir da década de 60 do século XX.

Gauss e todas as outras pessoas que usavam a distribuição normal para calcular probabilidades até recentemente resolviam as integrais usando métodos numéricos manualmente.

Porém, todas as distribuições normais apresentam algumas características em comum, independentemente de seus valores de média e de variância:

- 68% dos dados estão situados entre a média menos um desvio-padrão (**$\mu - \sigma$**) e a média mais um desvio-padrão (**$\mu + \sigma$**);
- 95,5% dos dados estão situados entre a média menos dois desvios-padrão (**$\mu - 2\sigma$**) e a média mais dois desvios-padrão (**$\mu + 2\sigma$**); e
- 99,7% dos dados estão situados entre a média menos três desvios-padrão (**$\mu - 3\sigma$**) e a média mais três desvios-padrão (**$\mu + 3\sigma$**).

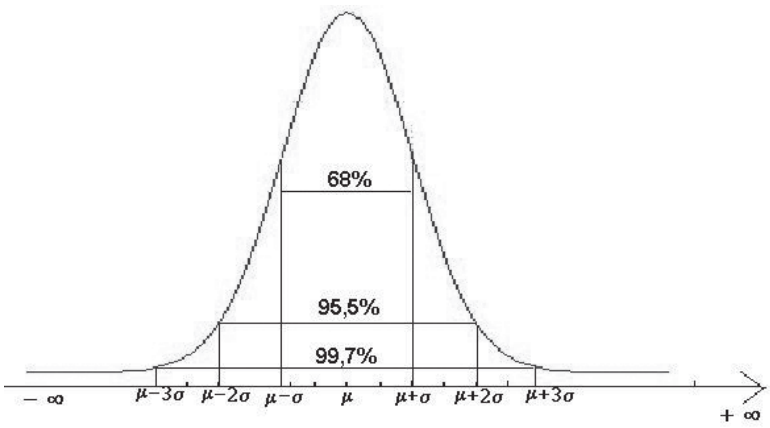

Figura 53: Percentuais de dados e número de desvios-padrão
Fonte: elaborada pelo autor

Por causa dessas características, alguém teve a idéia de criar um modelo normal-padrão: uma variável Z com distribuição normal de média igual a zero e desvio-padrão igual a 1 [Z: **N** ˜(, **1**)]. As probabilidades foram calculadas para esta distribuição-padrão e registradas em uma tabela. Através de uma transformação de variáveis chamada padronização, é possível converter os valores de qualquer distribuição normal em valores da distribuição normal-padrão e assim obter suas probabilidades – calcular o número de desvios-padrão, a contar da média, a que está um valor da variável, através da seguinte expressão:

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

Z – número de desvios-padrão a partir da média;
x – valor de interesse;
μ – média da distribuição normal de interesse;
σ – desvio-padrão da distribuição normal.

Z é um valor relativo: será negativo para valores de **x** menores do que a média e será positivo para valores de **x** maiores do que a média. Pela transformação, uma distribuição normal qualquer **X: N (μ , σ²)** passa a ser equivalente à distribuição normal-padrão Z: **N**˜ (, **1**), um valor de interesse x pode ser convertido em um valor z.

As probabilidades de uma variável com distribuição normal podem ser representadas por áreas sob a curva da distribuição normal-padrão. No AVEA, apresentamos uma tabela que relaciona valores positivos de z com áreas sob a cauda superior da curva. Os valores de z são apresentados com duas decimais. A primeira decimal fica na coluna da esquerda, e a segunda decimal, na linha do topo da tabela. A Figura 54 mostra como podemos usar essa Tabela para encontrar, por exemplo, a área sob a cauda superior da curva, além de z = 0,21.

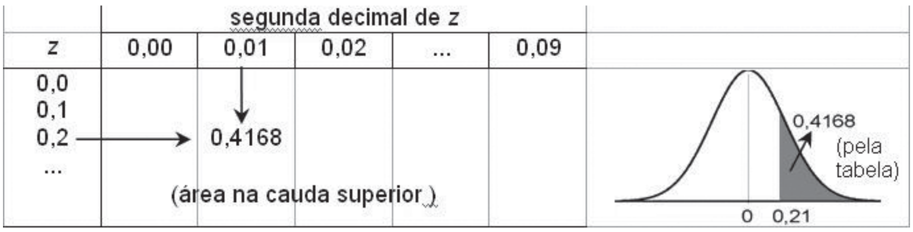

Figura 54: Ilustração do uso da tabela da distribuição normal-padrão (Tabela III do apêndice) para encontrar a área na cauda superior relativa ao valor de z = 0,21
Fonte: Barbetta, Reis, Bornia (2004)

No Exemplo 7, suponha uma variável aleatória X com média 50 e desvio-padrão 10. Há interesse em calcular a probabilidade do evento X > 55.

Primeiro, calculamos o valor de Z correspondente a 55. Z = (55 - 50)/ 10 = + 0,5.

Pelas Figuras 55 e 56, se pode ver a correspondência entre as duas distribuições:

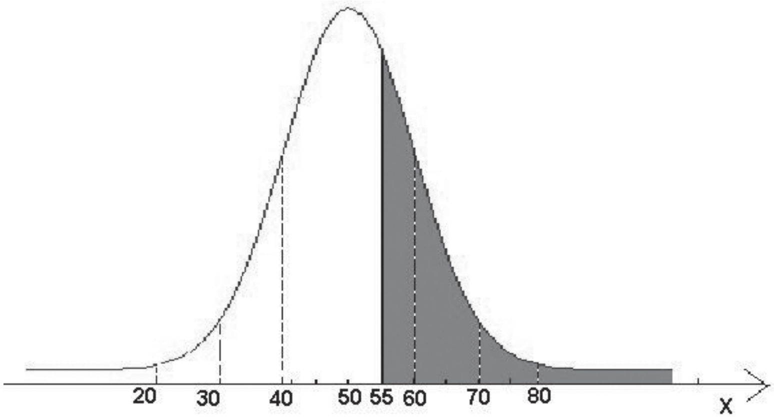

Figura 55: Distribuição normal N(50,102)
Fonte: elaborada pelo autor

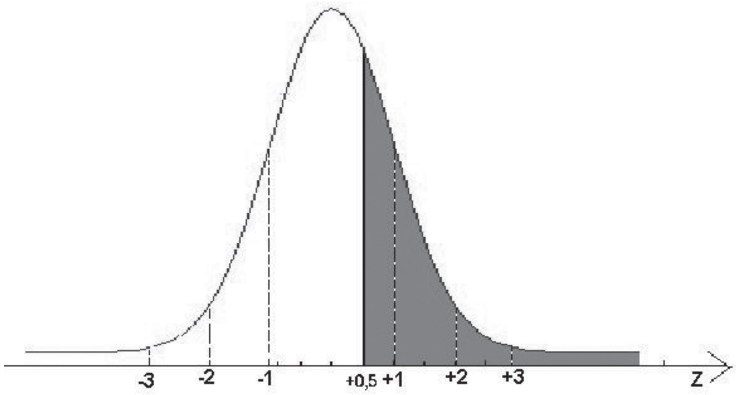

Figura 56: Distribuição normal-padrão
Fonte: elaborada pelo autor

O evento P (X>55) é equivalente ao evento P (Z> 0,5). Este valor pode ser obtido na tabela da distribuição normal-padrão (ver Ambiente Virtual). Os valores de Z são apresentados com duas decimais: o primeiro, na coluna da extrema esquerda, e o segundo, na linha do topo da tabela. Observe, pelas figuras que estão no alto da tabela, que as probabilidades são para eventos do tipo das figuras acima [P(Z> z1)]. Assim, poderíamos procurar a probabilidade do evento (Z > 0,5): fazendo o cruzamento do valor 0,5 (na coluna) com o valor 0,00 (na linha do topo), encontramos o valor 0,3085 (30,85%). Portanto, P(X>55) é igual a 0,3085. Observe a coerência entre o valor encontrado e as áreas nas figuras: a área é menor do que a metade das figuras (metade das figuras significaria 50%), e a probabilidade encontrada vale 30,85%.

Neste oitavo exemplo, supondo a mesma variável aleatória X com média 50 e desvio-padrão 10. Agora, há interesse em calcular a probabilidade de que X seja menor do que 40.

Primeiro, precisamos calcular o valor de Z correspondente a 40. Z = (40 - 50)/ 10 = -1,00.

Pelas Figuras 57 e 58, podemos ver a correspondência entre as duas distribuições:

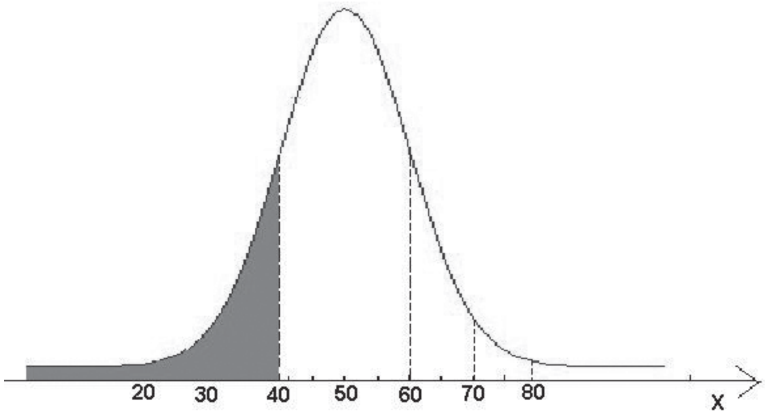

Figura 57: Distribuição normal N(50,102)
Fonte: elaborada pelo autor

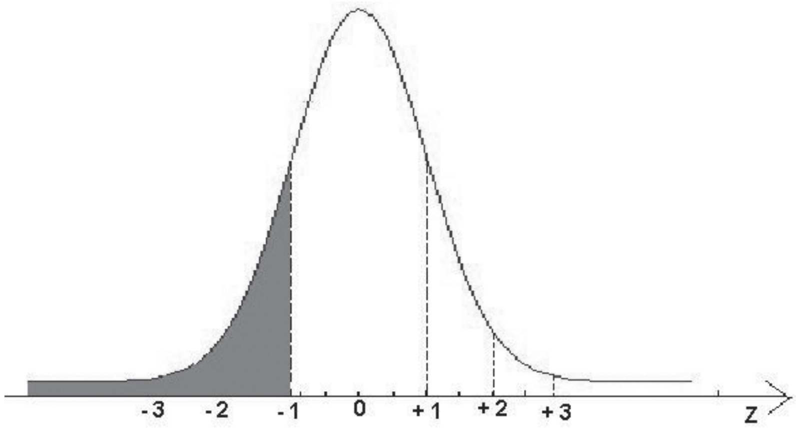

Figura 58: Distribuição normal-padrão
Fonte: elaborada pelo autor

O evento P (X<40) é equivalente ao evento P (Z < –1,00). Repare, porém, que queremos encontrar P (Z < –1,00), e a tabela nos apresenta valores apenas para P (Z > 1,00). Contudo, se rebatermos as figuras da distribuição normal para a direita, teremos o seguinte resultado (Figura 59):

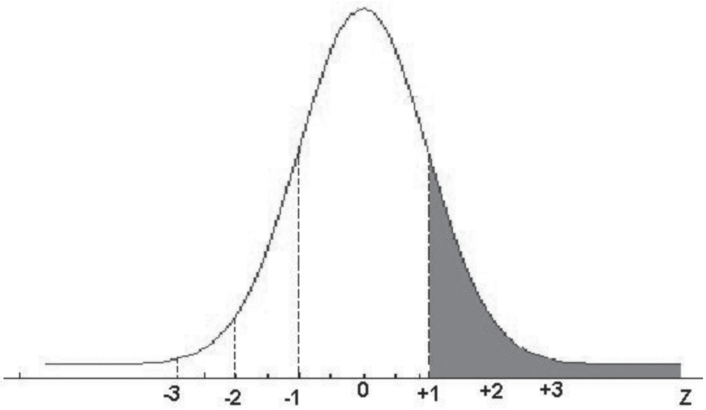

Figura 59: Distribuição normal-padrão
Fonte: elaborada pelo autor

Ou seja, a área P(Z < –1) = P(Z > 1). Esta probabilidade, nós podemos encontrar diretamente pela tabela, fazendo o cruzamento do valor 1,0 (na coluna) com o valor 0,00 (na linha do topo) encontramos o valor 0,1587 (15,87%). Portanto, P(X<40) = P(Z<-1) = P(Z>1), que é igual a 0,1587.

No nono exemplo, supondo a mesma variável aleatória X com média 50 e desvio-padrão 10. Agora, há interesse em calcular a probabilidade de que X seja maior do que 35.

Primeiro, precisamos calcular o valor de Z correspondente a 35. Z = (35 – 50)/ 10 = –1,50.

Pelas Figuras 60 e 61, se pode ver a correspondência entre as duas distribuições:

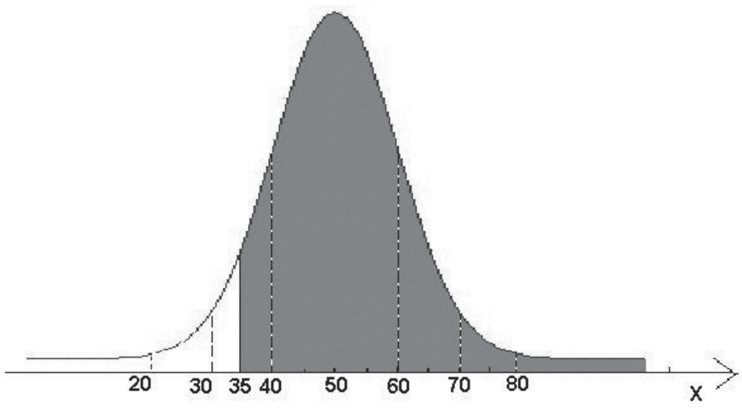

Figura 60: Distribuição Normal N(50,102)
Fonte: elaborada pelo autor

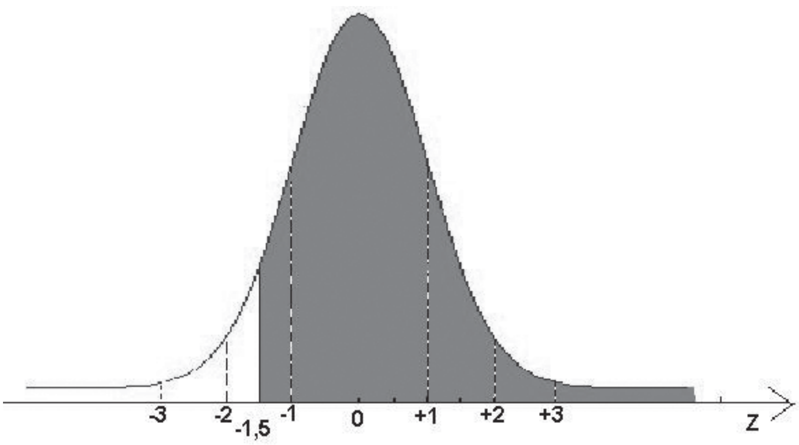

Figura 61: Distribuição normal-padrão
Fonte: elaborada pelo autor

Não podemos obter a probabilidade P(Z > –1,50) diretamente, pois a tabela do Ambiente Virtual apresenta apenas resultados para valores positivos de Z. Sabemos que a probabilidade total vale 1,0, podemos, então, considerar que P(Z > –1,50) = 1 - P(Z < –1,50). Usando o raciocínio descrito no Exemplo 8 (rebatendo as figuras para a direita), vamos obter: P(Z < –1,50) = P(Z > 1,50). Esta última probabilidade pode ser facilmente encontrada na tabela da distribuição normal-padrão: P(Z > 1,50) = P(Z < –1,50) = 0,0668. Basta substituir na expressão: P(Z > –1,50) = 1 – P(Z < –1,50) = 1 – 0,0668 = 0,9332 (93,32%). Observe novamente a coerência entre as áreas das figuras acima e o valor da probabilidade: a área nas figuras compreende mais do que 50% da probabilidade total, aproximando-se do extremo inferior da distribuição, perto de 100%, e a probabilidade encontrada realmente é próxima de 100%.

No Exemplo 10, supondo a mesma variável aleatória X com média 50 e desvio-padrão 10. Agora, há interesse em calcular a probabilidade de que X assuma valores entre 48 e 56.

Calcular P (48 < X < 56). Veja a Figura 62 abaixo:

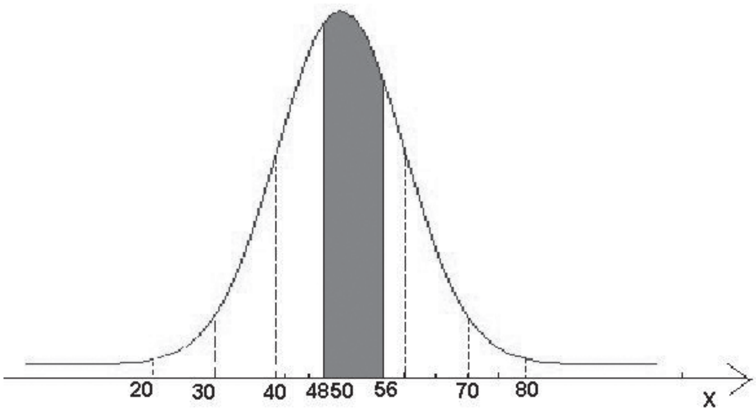

Figura 62: Distribuição normal $N(50, 10^2)$
Fonte: elaborada pelo autor

Novamente, precisamos calcular os valores de Z correspondentes a 48 e a 56.

$Z_1 = (48 – 50)/ 10 = – 0,20$ $\quad\quad$ $Z_2 = (56 – 50)/ 10 = 0,60$

Então: $P (48 < X < 56) = P (–0,20 < Z < 0,60)$

Repare que a área entre 48 e 56 é igual à área de 48 até $+\infty$ MENOS a área de 56 até $+\infty$:

$P(48 < X < 56) = P(X > 48) – P(X > 56) = P(– 0,20< Z <0,60) =$
$P(Z > – 0,20) – P(Z > 0,60)$

E os valores acima podem ser obtidos na tabela da distribuição normal-padrão:

$P(Z > 0,60) = 0,2743$

$P(Z > – 0,20) = 1– P(Z > 0,20) = 1– 0,4207 = 0,5793$

$P(48 < X < 56)= P(– 0,20< Z <0,60) = P(Z > – 0,20) – P(Z > 0,60)$
$= 0,5793 – 0,2743 = 0,3050$

Então, a probabilidade de a variável X assumir valores entre 48 e 56 é igual a 0,305 (30,5%).

A distribuição normal também pode ser utilizada para encontrar valores da variável de interesse correspondentes a uma probabilidade fixada.

No Exemplo 11, supondo a mesma variável aleatória X com média 50 e desvio-padrão 10. Encontre os valores de X, situados à mesma distância abaixo e acima da média, que contém 95% dos valores da variável.

Como a distribuição normal é simétrica em relação à média, e como neste problema os valores de interesse estão situados à mesma distância da média, “sobram” 5% dos valores, 2,5% na cauda inferior e 2,5% na superior, como na Figura 63:

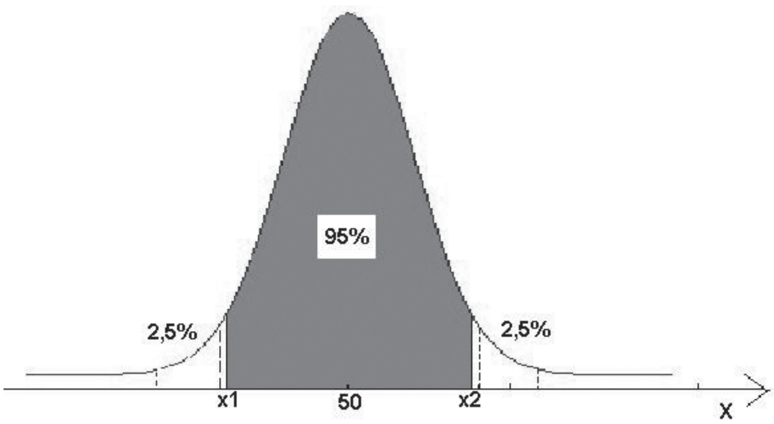

Figura 63: Distribuição normal N (50, 10^2)
Fonte: elaborada pelo autor

É preciso encontrar os valores de Z (na tabela da distribuição normal-padrão) correspondentes às probabilidades da figura acima, e a partir daí obter os valores de $\mathbf{x_1}$ e $\mathbf{x_2}$. Passando para a distribuição normal-padrão $\mathbf{x_1}$, corresponderá a um valor $\mathbf{z_1}$, e $\mathbf{x_2}$ a um valor $\mathbf{z_2}$, como na Figura 64:

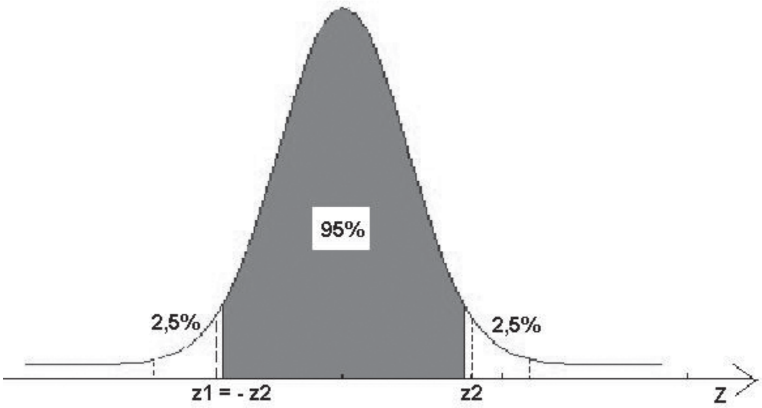

Figura 64: Distribuição normal-padrão
Fonte: elaborada pelo autor

Repare que a média da distribuição normal-padrão é igual a zero, fazendo com que $\mathbf{z_1}$ e $\mathbf{z_2}$ sejam iguais em módulo. Podemos encontrar $\mathbf{z_2}$, já que $P(Z > \mathbf{z_2}) = 0{,}025$

É necessário encontrar o valor da probabilidade na tabela da distribuição normal-padrão (ou o valor mais próximo) e obter o valor de Z associado.

Para o caso de $\mathbf{z_2}$, ao procurar pela probabilidade 0,025, encontramos o valor exato 0,025, e, por conseguinte, o valor de $\mathbf{z_2}$, que é igual a 1,96: $P(Z > 1{,}96) = 0{,}025$.

Como $\mathbf{z_1} = -\mathbf{z_2}$, encontramos facilmente o valor de z1: $\mathbf{z_2} = -1{,}96$. $P(Z < -1{,}96) = 0{,}025$.

Observe que os valores são iguais em módulo, mas corresponderão a valores diferentes da variável X. A expressão usada para obter o valor de Z, em função do valor da variável X, pode ser usada para o inverso:

$$Z = \frac{x - \mu}{\sigma} \quad => \quad x = \mu + Z \times \sigma$$

E assim obteremos os valores de $\mathbf{x_1}$ e $\mathbf{x_2}$, que correspondem a $\mathbf{z_1}$ e $\mathbf{z_2}$, respectivamente:

É muito importante que se preste atenção no sinal do valor de z ao obter o valor de x.

$$\mathbf{x_1 = \mu + (z_1 \text{ x } \sigma = 50 + [(-1{,}96) \text{ x } 10] = 30{,}4}$$

$$\mathbf{x_2 = \mu + (z_2 \text{ x } \sigma = 50 + (1{,}96 \text{ x } 10) = 69{,}6}$$

GLOSSÁRIO

*Modelo binomial – modelo probabilístico para variáveis aleatórias discretas que descreve o número de sucessos em n experimentos independentes (sendo n finito e conhecido). Os experimentos podem ter apenas dois resultados possíveis, e a probabilidade de sucesso permanece constante durante os n experimentos. Fonte: Barbetta, Reis e Bornia (2004) e Lopes (1999).

Observe que os resultados obtidos são coerentes: 30,4 está abaixo da média (1,96 desvios-padrão), e 69,6, acima (também 1,96 desvios-padrão). O intervalo definido por estes dois valores compreende 95% dos resultados da variável X.

Todo este trabalho poderia ter sido poupado, se houvesse um programa computacional que fizesse esses cálculos. Há vários softwares disponíveis no mercado, alguns deles de domínio público, que calculam as probabilidades associadas a determinados eventos, como também os valores associados a determinadas probabilidades.

Uma das características mais importantes do modelo normal é a sua capacidade de aproximar outros modelos, permitindo muitas vezes simplificar os cálculos de probabilidade. Na próxima seção, vamos ver como o modelo normal pode ser usado para aproximar o **binomial***.

Modelo normal como aproximação do binomial

O modelo binomial (discreto) pode ser aproximado pelo modelo normal (contínuo) se certas condições forem satisfeitas:

- quando o valor de **n** (número de ensaios) for tal que os cálculos binomiais sejam trabalhosos demais;

Para os que pensam que o advento dos computadores eliminou este problema, um alerta: em alguns casos, os números envolvidos são tão grandes que sobrepujam suas capacidades.

- quando o produto **n** × **p** (o valor esperado do modelo binomial) e o produto **n** × (**1 – p**) forem ambos maiores ou iguais a 5.

Se isso ocorrer, uma binomial de parâmetros **n** e **p** pode ser aproximada por uma normal com:

média = $\mu = n \times p$ (valor esperado do modelo binomial)

variância = $\sigma^2 = n \times p \times (1 - p)$ (variância do modelo binomial)

Usando o modelo normal (contínuo) para aproximar o binomial (discreto), é necessário fazer uma correção de continuidade: associar um intervalo ao valor discreto, para que o valor da probabilidade calculada pelo modelo contínuo seja mensurável. Este intervalo deve ser centrado no valor discreto e ter uma amplitude igual à diferença entre dois valores consecutivos da variável discreta: se, por exemplo, a diferença for igual a 1 (a variável somente pode assumir valores inteiros), o intervalo deve ter amplitude igual a 1, 0,5 abaixo do valor e 0,5 acima. **Esta correção de continuidade precisa ser feita para garantir a coerência da aproximação**.

Seja uma variável aleatória X com distribuição binomial.

1) Há interesse em calcular a probabilidade de X assumir um valor **k** genérico, P(X = **k**); ao fazer a aproximação pela normal, será: P(**k – 0,5** < X < **k + 0,5**).

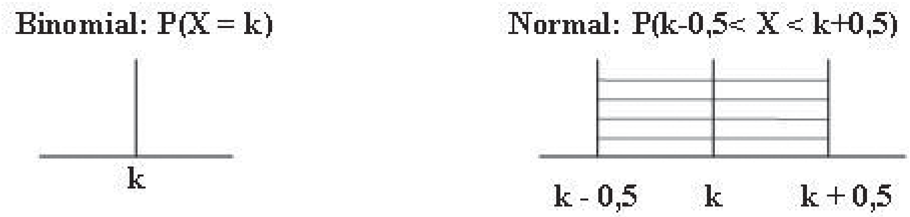

Figura 65: Correção de continuidade da aproximação do modelo binomial pelo normal – 1º caso
Fonte: elaborada pelo autor

2) Há interesse em calcular a probabilidade de X assumir valores menores ou iguais a um valor **k** genérico, P(X ≤ **k**); ao fazer a aproximação pela normal, será: P(X < **k** + **0,5**), todo o intervalo referente a k será incluído.

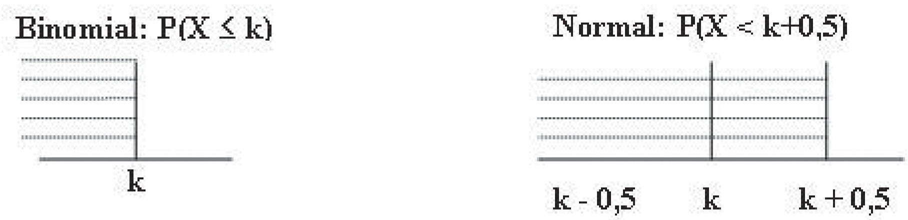

Figura 66: Correção de continuidade da aproximação do modelo binomial pelo normal – 2º caso
Fonte: elaborada pelo autor

3) Há interesse em calcular a probabilidade de X assumir valores maiores ou iguais a um valor **k** genérico, P(X ≥ **k**); ao fazer a aproximação pela normal, será: P(X > **k – 0,5**), todo o intervalo referente a k será incluído.

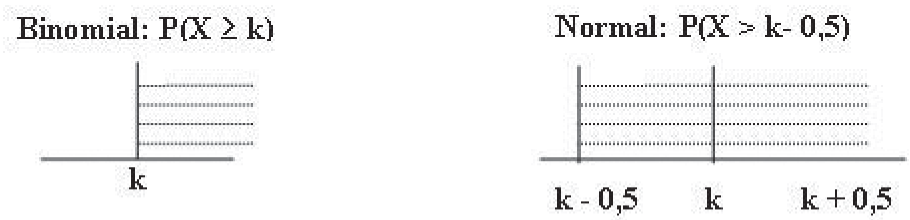

Figura 67: Correção de continuidade da aproximação do modelo binomial pelo normal – 3º caso
Fonte: elaborada pelo autor

4) Há interesse em calcular a probabilidade de X assumir valores menores do que um valor **k** genérico, P(X < **k**); ao fazer a aproximação pela normal, será: P(X < **k – 0,5**), todo o intervalo referente a k será excluído.

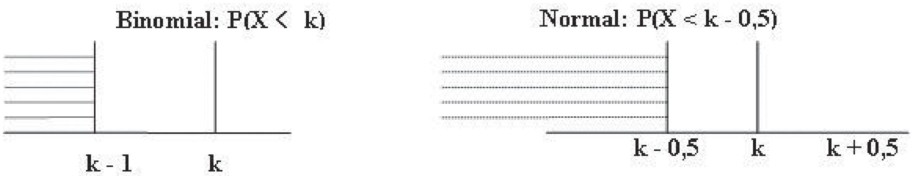

Figura 68: Correção de continuidade da aproximação do modelo binomial pelo normal – 4º caso
Fonte: elaborada pelo autor

5) Há interesse em calcular a probabilidade de X assumir valores maiores do que um valor **k** genérico, P(X > **k**); ao fazer a aproximação pela normal, será: P(X > **k + 0,5**), todo o intervalo referente a k será excluído.

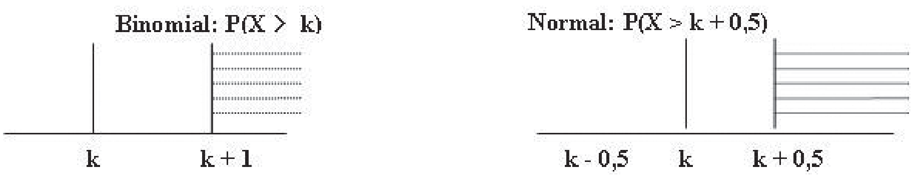

Figura 69: Correção de continuidade da aproximação do modelo binomial pelo normal – 5º caso
Fonte: elaborada pelo autor

Um município tem 40.000 eleitores. Para uma pesquisa de opinião eleitoral, uma amostra aleatória de 1.500 pessoas foi selecionada.

Vamos ver, nesse décimo segundo exemplo, qual é a probabilidade de que pelo menos 500 dos eleitores sejam menores de 25 anos se 35% dos 40.000 são menores do que 25 anos?

Este problema poderia ser resolvido usando o modelo binomial. Há apenas dois resultados possíveis para cada eleitor: menor de 25 anos ("sucesso") e maior ou igual a 25 anos ("fracasso"). Existe um limite superior de realizações, no caso, os 1.500 eleitores da amostra, e há independência entre as retiradas, pois a amostra foi retirada de forma aleatória (e a amostra representa menos de 5% dos 40.000 eleitores).

Então: "sucesso" = menor de 25 anos

p = 0,35 **1 - p** = 0,65 **n** = 1.500

A variável aleatória discreta X, número de eleitores menores de 25 anos em 1.500, terá distribuição binomial com parâmetros **n** = 1.500 e **p** = 0,35.

O evento "pelo menos 500 menores de 25 anos" seria definido como 500 ou mais eleitores:

$$P\,(X \geq 500) = P(X = 500) + P(X = 501) + ...+ P(X = 1.500)$$

Há cerca de 1.000 expressões binomiais.

Vamos ver se é possível aproximar pelo modelo normal.

O valor de **n** é grande:

$$\mathbf{n} \times \mathbf{p} = 1.500 \times 0{,}35 = 525 > 5 \text{ e } \mathbf{n} \times (\mathbf{1 - p}) = 1.500 \times 0{,}65 = 975 > 5.$$

Como as condições foram satisfeitas, é possível aproximar por um modelo normal:

média = □ = **n × p = 1.500 x 0,35 = 525**

desvio-padrão = □ $= \sqrt{n \times p \times (1-p)} = \sqrt{1500 \times 0{,}35 \times 0{,}65} = 18{,}47$

Pelo modelo binomial: P (X ≥ 500). Pelo modelo normal, será: P (X ≥ 499,5).

$$P(X \geq 499{,}5) = P(Z > z_1)$$

$$z_1 = (499{,}5 - 525)/18{,}47 = -1{,}38$$

$$P(Z > -1{,}38) = 1 - P(Z > 1{,}38)$$

Procurando na tabela da distribuição normal-padrão:

P (Z > 1,38) = 0,0838

Então:

P (X ≥ 500) ≅ P(X ≥ 499,5) = P (Z > –1,38) = 1 – P(Z >1,38) = 1 – 0,0838 = 0,9162.

A probabilidade de que pelo menos 500 dos eleitores da amostra sejam menores de 25 anos é igual a 0,9162 (91,62%).

Nas próximas duas seções, vamos ver modelos probabilísticos derivados do modelo normal, usados predominantemente em processos de inferência estatística. Vamos introduzi-los agora para facilitar nosso trabalho quando chegarmos às Unidades 9 e 10.

Modelo (distribuição) t de Student

Havia um matemático inglês, William Gosset, que trabalhava para a Cervejaria Guiness, em Dublin, Irlanda, no início do século XX. Ele atuava no controle da qualidade do cultivo de ingredientes para a fabricação de cerveja.

Nessa época, alguns estatísticos usavam a distribuição normal no estabelecimento de intervalos de confiança para a média a partir de pequenas amostras (veremos isso na Unidade 8). Calculavam média aritmética simples e variância da amostra, e generalizavam os resultados através do modelo normal, como fizemos no Exemplo 11.

Gosset descobriu que o modelo normal não funcionava direito para pequenas amostras e desenvolveu um novo modelo probabilístico, derivado do normal, introduzindo uma correção para levar em conta justamente o tamanho de amostra. Ele aplicou suas descobertas em seu trabalho e quis publicá-las, mas a Guiness apenas permitiu após ele adotar o pseudônimo “Student”. Por isso, o seu modelo é conhecido como t de Student para n – 1 graus de liberdade.

O valor **n –1** (tamanho da amostra menos 1) é chamado de número de **graus de liberdade** da estatística. Quando a variância amostral é calculada, supõe-se que a média já seja conhecida, assim apenas um determinado número de elementos da amostra poderá ter seus valores variando livremente; este número será igual a **n – 1**, porque um dos valores não poderá variar livremente, pois terá que ter um valor tal, que a média permaneça a mesma calculada anteriormente. Assim, a estatística terá **n – 1** graus de liberdade.

Trata-se de uma distribuição de probabilidades que apresenta média igual a zero (como a normal-padrão), é simétrica em relação à média, mas apresenta uma variância igual a n / (n – 2), ou seja, seus valores dependem do tamanho da amostra, apresentando maior variância para menores valores de amostra. Quanto maior o tamanho da amostra, mais a variância de **t** aproxima-se de 1,00 (variância da normal-padrão). A distribuição t de Student está na Figura 70 para vários graus de liberdade:

Esta é a correção propriamente dita, pois, ao usar pequenas amostras, o risco de que a variância amostral da variável seja diferente da variância populacional é maior, podendo levar a intervalos de confiança que não correspondem à realidade. A não-utilização desta correção foi a fonte de muitos erros no passado e, infelizmente, ainda de alguns erros no presente.

Para tamanhos de amostra maiores do que 30, supõe-se que a variância de **t** é igual a 1: por isso, a aproximação do item b.1.

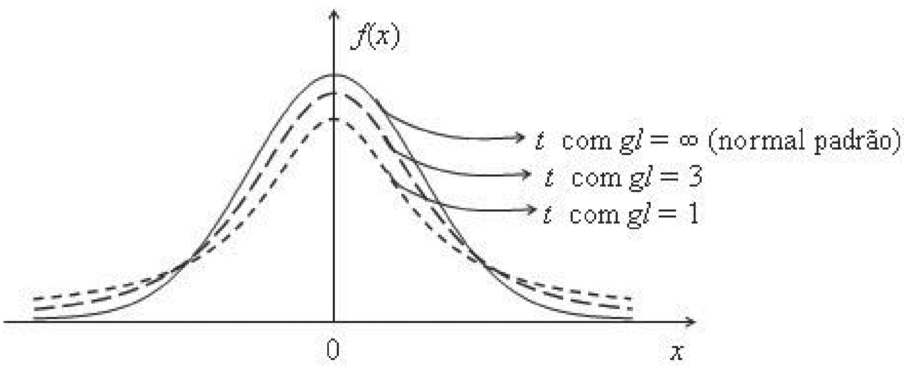

Figura 70: Distribuição t de Student para vários graus de liberdade
Fonte: Barbetta, Reis, Bornia (2004)

Observe que, tal como a distribuição normal-padrão, a distribuição **t** de Student é **simétrica** em relação à média (que é igual a zero).

A tabela da distribuição **t** de Student encontra-se no Ambiente Virtual, para vários graus de liberdade e valores de probabilidade. Vamos ver um exemplo.

Neste décimo terceiro exemplo, imagine a situação do Exemplo 12, obter os valores de t simétricos em relação à média que contêm 95% dos dados, supondo uma amostra de dez elementos.

Temos que encontrar os valores $\mathbf{t}_1$ e $\mathbf{t}_2$, simétricos em relação à média que definem o intervalo que contém 95% dos dados. Como supomos uma amostra de dez elementos, a distribuição t de Student terá 10 – 1 = 9 graus de liberdade. Repare que a média da distribuição t de Student é igual a zero, fazendo com que $\mathbf{t}_1$ e $\mathbf{t}_2$ sejam iguais em módulo. Podemos encontrar $\mathbf{t}_2$, já que $P(t > \mathbf{t}_2) = 0{,}025$. Veja a Figura 71:

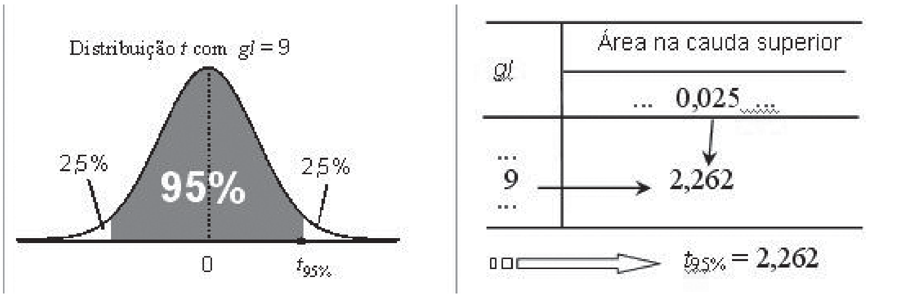

Figura 71: Uso da tabela da distribuição t de Student. Ilustração com gl = 9 e área na cauda superior de 2,5%
Fonte: Barbetta, Reis, Bornia (2004)

Vamos utilizar bastante a distribuição t de Student nas Unidades 9 e 10.

Modelo qui-quadrado

Trata-se de mais um modelo derivado da distribuição normal, embora não vamos discutir como se dá esta derivação aqui.

Na Unidade 3, estudamos como descrever os relacionamentos entre duas variáveis qualitativas, geralmente expressos através de uma tabela de contingências. No Exemplo 5 da Unidade 3, analisamos o relacionamento entre modelo e opinião geral sobre os veículos da Toyord. Havíamos concluído que havia relacionamento, pois os modelos mais baratos apresentavam maiores percentuais de insatisfeitos do que os mais caros.

Na Unidade 10, vamos aprender a calcular uma estatística que relacionará as freqüências observadas de cada cruzamento entre os valores de duas variáveis qualitativas, expressas em uma tabela de contingências, com as freqüências esperadas desses mesmos cruzamentos, se as duas variáveis não tivessem qualquer relacionamento entre si. Esta estatística é chamada de qui-quadrado, χ^2, e caso a hipótese seja de que as variáveis não se relacionam, ela seguirá o modelo qui-quadrado com um certo número de graus de liberdade.

O número de graus de liberdade dependerá das condições da tabela: para o caso que será visto na Unidade 10, será o produto do número de linhas da tabela – 1 pelo número de colunas da tabela – 1. É uma distribuição assimétrica, sempre positiva, que tem valores diferentes, dependendo do seu número de graus de liberdade. Sua média é igual ao número de graus de liberdade, e a variância é igual a duas vezes o número de graus de liberdade.

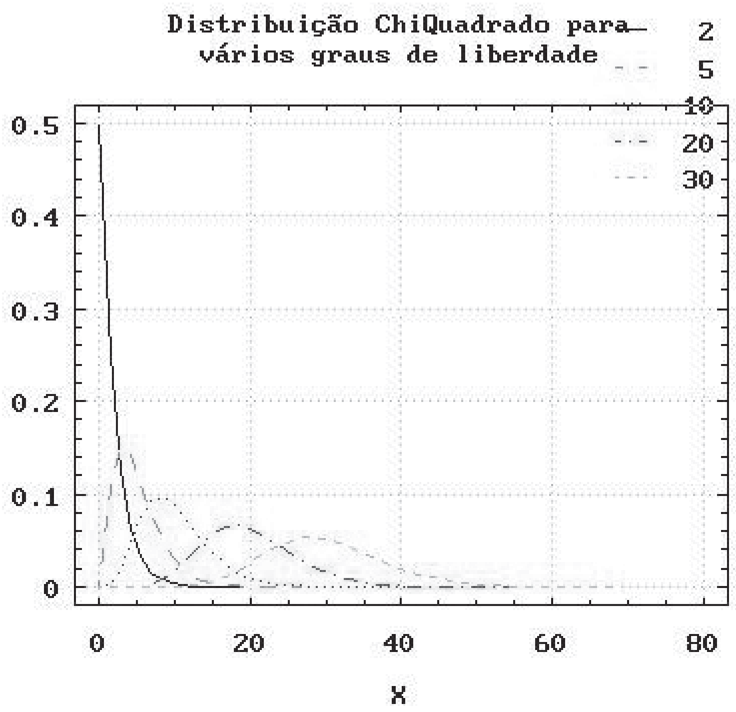

Figura 72: Modelo qui-quadrado com 2, 5, 10, 20 e 30 graus de liberdade
Fonte: adaptada pelo autor de Stagraphics®

A Figura 72 mostra as curvas do modelo (distribuição) qui-quadrado para 2, 5, 10, 20 e 30 graus de liberdade. Observe que a figura é assimétrica e como varia de forma, dependendo do número de graus de liberdade da estatística.

A tabela da distribuição qui-quadrado encontra-se no Ambiente Virtual de Ensino-Aprendizagem, para vários graus de liberdade e valores de probabilidade. Vamos ver um exemplo.

Neste décimo quarto exemplo, imagine que queremos encontrar o valor da estatística qui-quadrado, para três graus de liberdade, deixando uma área na cauda superior de 5%.

O valor da estatística qui-quadrado que define uma área na cauda superior de 5% pode ser encontrado através da Tabela, cruzando a linha de três graus de liberdade com a coluna de área na cauda superior igual a 0,05. Veja a Figura a seguir:

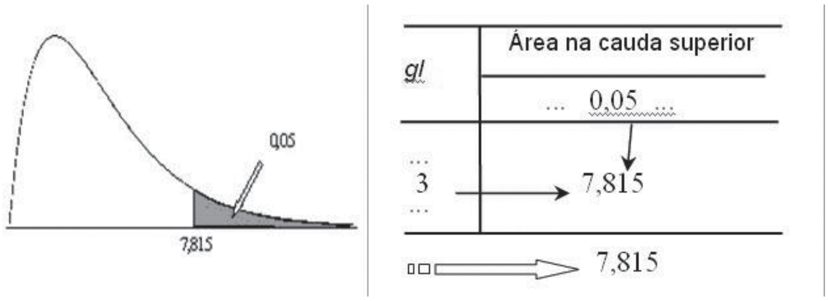

Figura 73: Uso da tabela da distribuição qui-quadrado. Ilustração com gl = 3 e área na cauda superior de 5%
Fonte: adaptado pelo autor de Barbetta, Reis, Bornia (2004)

Com este tópico, terminamos a Unidade 7. Na Unidade 8, você verá o importante conceito de distribuição amostral, que é indispensável para o processo de generalização (inferência) estatística que será estudado nas Unidades 9 e 10.

Saiba mais...

- Sobre modelos probabilísticos para variáveis aleatórias discretas:

BARBETTA, P. A. *Estatística Aplicada às Ciências Sociais*. 6. ed. Florianópolis: Ed. da UFSC, 2006, capítulo 7.

BARBETTA, P. A.; REIS, M. M.; BORNIA, A. C. *Estatística para Cursos de Engenharia e Informática*. São Paulo: Atlas, 2004, capítulo 5.

STEVENSON, Willian J. *Estatística Aplicada à Administração*. São Paulo: Harbra, 2001, capítulo 4.

- Sobre modelos probabilísticos para variáveis aleatórias contínuas:

BARBETTA, P. A. *Estatística Aplicada às Ciências Sociais*. 6. ed. Florianópolis: Ed. da UFSC, 2006. capítulo 8.

BARBETTA, P. A.; REIS, M. M.; BORNIA, A. C. *Estatística para Cursos de Engenharia e Informática*. São Paulo: Atlas, 2004, capítulo 6.

STEVENSON, Willian J. *Estatística Aplicada à Administração*. São Paulo: Harbra, 2001, capítulo 5.

- Sobre a utilização do Microsoft Excel para cálculo de probabilidades para os principais modelos probabilísticos, veja LEVINE, D. M. et al. *Estatística*: teoria e aplicações – usando Microsoft Excel em português. 5. ed. Rio de Janeiro: LTC, 2006, capítulos 4 e 5.

RESUMO

O resumo desta Unidade está mostrado na Figura 74:

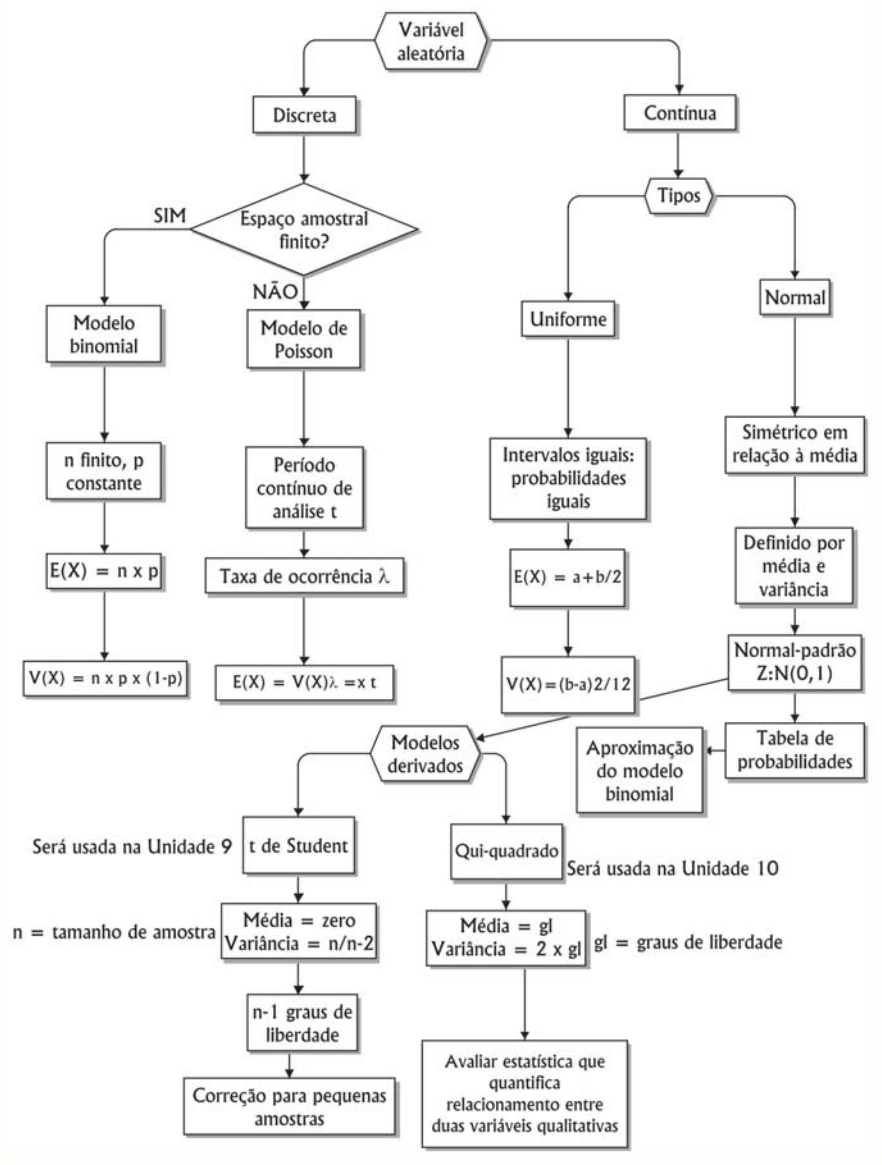

Figura 74: Resumo da Unidade 7
Fonte: elaborada pelo autor

Atividades de aprendizagem

As atividades de aprendizagem estão disponíveis no Ambiente Virtual de Ensino-Aprendizagem. Não deixe de respondê-las.

Caro estudante!

Chegamos ao final da Unidade 7 do nosso livro. Nela estudamos os modelos probabilísticos mais comuns. Esta Unidade foi repleta de figuras, quadros, representações, e exemplos de utilização das técnicas e das diferentes formas de utilização destes modelos. Releia, caso necessário, todos os exemplos, leia as indicações do Saiba mais e discuta com seus colegas. Responda as atividades de aprendizagem e visite o Ambiente Virtual de Ensino-Aprendizagem. Conte sempre com o acompanhamento da tutoria e das explicações do professor. Ótimos estudos!