CAPITULO 8: AMOSTRAGEM E ESTIMAÇÃO PONTUAL

Conceitos e resultados a serem apresentados neste capítulo:

8.1- Amostra aleatória

8.2- Estatística. Distribuição amostral

8.3- Média Amostral. Distribuição amostral de $\bar{X}$

8.4- Variância e Desvio Padrão amostrais

8.6- Proporção Amostral. Distribuição amostral de $\hat{p}$

8.7- Parâmetro, estimador e estimativa; Estimação Pontual de parâmetros

8.7.2- Estimador não Tendencioso. Viés de um estimador

8.7.3- Erro Quadrático Médio de um estimador

8.7.4- Erro Absoluto de estimacão

8.8- Dimensionamento da amostra

8.8.1 Dimensionando amostra para estimar a média popul., com σ conhecido

8.8.2 Dimensionando amostra para estimar a média popul., com σ desconhecido

8.8.4 Dimensionamento de Amostra para estimar a proporção

"A experiência sem teoria é cega, mas teoria sem experiência é mero jogo intelectual."

Immanuel Kant, filósofo

No Capítulo anterior foram apresentados os conceitos da **Análise Exploratória** para conjuntos de dados amostrais. A partir deste capítulo vamos mostrar como esse tipo de análise se relaciona com a **Teoria de Probabilidades** apresentada nos Capítulos 1 até 6. Visando estabelecer uma relação entre esses dois temas, apresentamos, a seguir, outra formulação para o conceito de **amostra aleatória**.

8.1 - Amostra aleatória

Seja X uma variável aleatória distribuída conforme um determinado modelo probabilístico.

Diremos então que **(X_1 , X_2 , ..., X_n) é uma amostra aleatória da variável aleatória X** se as n v.a.'s X_1, X_2, ..., X_n são **independentes e identicamente distribuídas (iid)**, com a mesma distribuição de X.

Para interpretarmos adequadamente a definição acima devemos considerar X_1 , X_2 , ..., X_n como n medições independentes da v.a. X. Para que cada X_i tenha a mesma distribuição de X é necessário fazer essas mensurações em **condições essencialmente iguais**, como por exemplo, usando um mesmo instrumento de medição, pessoal identicamente treinado, material extraído de um mesmo processo produtivo, etc. Além disso, através de um **adequado procedimento de aleatorização** devemos nos assegurar da **independência entre as diferentes mensurações**.

Para distinguir as variáveis aleatórias X_i's dos **valores que elas assumem**, denotaremos os valores com as **letras minúsculas** correspondentes. Desta maneira, os valores correspondentes à amostra aleatória (X_1 , X_2 , ..., X_n) serão representados por (x_1, x_2, ..., x_n).

Exemplo 8.1 **Carga de ruptura**

No Exemplo 7.4 foram mostradas as medições da carga de ruptura, em kg, de 30 espécimes de cabos naúticos.

Se X é a variável aleatória que representa a carga de ruptura, então a amostra aleatória de tamanho n=30 é representada por $(X_1, X_2, \ldots, X_{30})$ e os valores das medições desta amostra específica são representados por $(x_1, x_2, \ldots, x_{30})$, onde x_1= 83 kg , x_2 = 96 kg , ... , . x_{30}= 96 kg.

Exemplo 7.4: Carga de ruptura de cabos

A medição da carga de ruptura, em kg, para 30 espécimes de cabos

83	96	73	102	93	94	99	85	91	118	93	103	87	95	102
84	100	95	90	81	102	98	94	89	91	78	85	83	105	96

Se X for uma v.a. contínua, com função de densidade f, podemos obter a **função de densidade conjunta g do vetor aleatório** $(X_1, X_2, \ldots, X_n)$, fazendo uso da propriedade de **independência entre as variáveis e do fato de todas serem identicamente distribuídas**. Desta maneira teremos

$$g(x_1, x_2, \ldots, x_n) = f(x_1).f(x_2)\ldots f(x_n),$$

para todo vetor de dados $(x_1, x_2, \ldots, x_n)$.

Se X for uma **v.a. discreta**, com função de probabilidade p, a função de probabilidade conjunta, q, da amostra aleatória é

$$q(x_1,x_2, \ldots,x_n) = P(X_1=x_1).P(X_2=x_2)\ldots\ldots P(X_n=x_n) = p(x_1).p(x_2)\ldots p(x_n),$$

para todo vetor de dados $(x_1,x_2,\ldots,x_n)$.

8.2 - Estatísticas

No Capítulo 7 foram definidas medidas de centralidade e de dispersão para um conjunto de dados quantitativos. Essas medidas eram calculadas a partir do conjunto de dados e, como acabamos de ver, podem ser interpretadas como os valores assumidos por uma **amostra aleatória (X_1, X_2, ..., X_n)**. Desta maneira, elas podem ser definidas em termos das variáveis aleatórias que compõem a dita amostra.

Por exemplo, no Capítulo 7, dado um conjunto de dados quantitativos x_1, x_2, ...,x_n, a média aritmética amostral, $\bar{x}$, deste conjunto foi definida como $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. No enfoque do presente capítulo as observações **x_i's podem ser consideradas como sendo os valores particulares (x_1, x_2, ...,x_n) de uma amostra aleatória (X_1, X_2, ..., X_n)**. Assim, $\bar{x}$, pode ser considerado como um valor particular da média aritmética , $\bar{X}$, definida por $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

$\bar{X}$ é um exemplo do que chamaremos de **estatística**, conceito este cuja definição mais geral apresentamos a seguir.

Estatísticas

Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória de uma v.a. X, e sejam $x_1, x_2, \ldots, x_n$ os correspondentes valores obtidos por **amostragem**. Seja T uma função real, cujo argumento é um vetor n-dimensional de números reais.

A v.a. $Y = T(X_1, X_2, \ldots, X_n)$ é dita uma **estatística** que, para essa amostra, toma o valor particular $y = T(x_1, x_2, \ldots, x_n)$.

Como a estatística Y é uma v.a., podemos falar na distribuição de Y. Neste caso, ao invés de usar a expressão "distribuição de probabilidades" falaremos em "**distribuição amostral de Y**".

Assim como as medidas vistas no capítulo 7, as estatísticas também podem ser classificadas como **estatísticas de centralidade**, de **dispersão** e de **ordem**. Nas seções a seguir veremos os principais exemplos desses tipos de estatísticas.

8.3 A Média Amostral

A média aritmética , $\bar{x}$, definida por: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ é denominada **média amostral**.

A média amostral, $\bar{x}$, é a estatística de centralidade mais utilizada. Como ocorre com toda estatística, faz sentido falar na sua **distribuição amostral**, na sua **esperança** e na sua **variância**.

Distribuição amostral de $\bar{X}$

Começaremos supondo que $(X_1, X_2, \ldots, X_n)$ é uma amostra aleatória de uma v.a. $X \sim N(\mu, \sigma^2)$. Assim **os X_i são iid como $N(\mu, \sigma^2)$**

Provamos, no Cap 6, que $Y_n = \sum_{i=1}^{n} x_i$ distribui-se conforme o modelo $N(n\mu, n\sigma^2)$.

No caso geral, em que não é feita a suposição de Normalidade, podemos usar o **Teorema Central do Limite** a fim de encontrar uma aproximação assintótica para a distribuição amostral de $\bar{X}$. Temos assim o seguinte resultado.

TCL: Sejam X uma v.a. com esperança μ e variância σ^2 e $(X_1, X_2, \ldots, X_n)$ uma **amostra aleatória de X**. Então $\frac{\bar{X}_n - E(\bar{X}_n)}{DP(\bar{X}_n)} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ tende à distribuição Normal padrão, quando **n**

Em outras palavras, se μ e σ^2 são, respectivamente, a média e a variância de uma variável aleatória X , então, para n suficientemente grande, a distribuição da média amostral de uma amostra aleatória de tamanho n, $\bar{X}_n$, **pode ser aproximada por uma Normal com média μ e variância σ^2/n** .

Nota: O desvio padrão de uma estatística é denominado **erro padrão** da estatística. Em particular, para a média amostral, diz-se que $\sigma/\sqrt{n}$ o **erro padrão de** $\bar{X}$.

Exemplo 8.2 **Especificação máxima de uma característica de qualidade**

As especificações de uma característica de qualidade estabelecem um limite máximo de 150,6 unidades. A medição desta característica comporta-se como uma v.a. X Normalmente distribuída com média 150 e desvio-padrão 2,1.

Determine a probabilidade de que a média amostral, $\bar{X}$, baseada em uma amostra aleatória de tamanho 49 ultrapasse a especificação limite de 150,6 ?

Solução:

Para X temos $E(X) = \mu = 150$, $DP(X) = \sigma = 2,1$.

Portanto, se n = 49, pelo TCL, $\bar{X}$ tem **distribuição Normal**

com média $E(\bar{X}) = \mu = 150$

e desvio padrão $DP(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{2,1}{\sqrt{49}} = 0,3.$

$$\text{Logo, } P(\bar{X} > 150,6) = P\left(\frac{\bar{X}-150}{0,3} > \frac{150,6-150}{0,3}\right) = 1 - \Phi\left(\frac{150,6-150}{0,3}\right) =$$

$$= 1 - \Phi(2,0) = 1 - 0,9772 = 0,0228 .$$

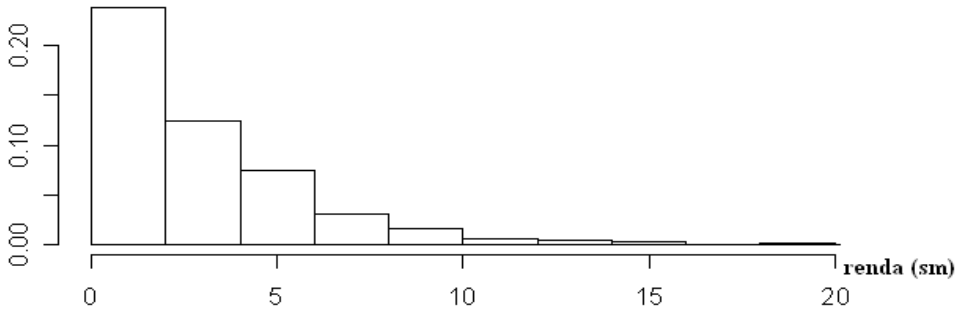
Ou seja, aproximadamente 2,3% das amostras de tamanho 49 apresentarão uma média aritmética da característica de qualidade acima da especificação máxima.

O Teorema Central do Limite (TCL)

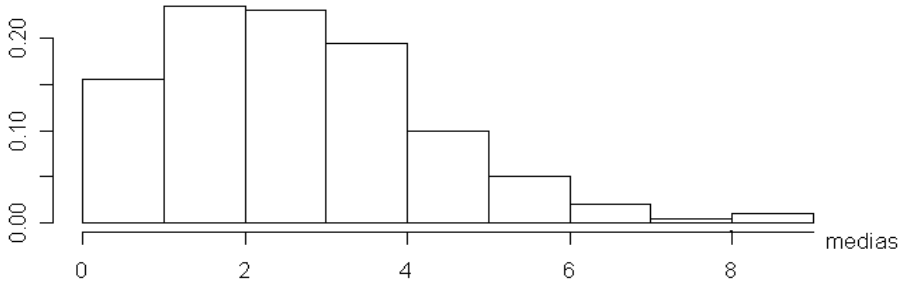
O Teorema Central do Limite (abreviadamente, TCL) diz respeito ao comportamento da **média amostral à medida que o tamanho n da amostra cresce indefinidamente**.

Exemplo: **A distribuição de renda e o TCL**

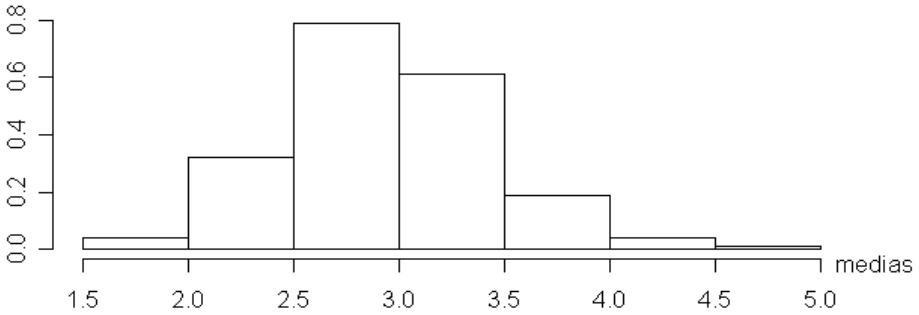
É um fato conhecido que a distribuição da renda pessoal dos habitantes de um país é usualmente muito desigual, ou seja, **muitos ganham pouco e poucos ganham muito**. Se forem sorteados 200 habitantes desse país e, com base nas suas rendas mensais construirmos um histograma, ele terá o aspecto.



Agora, se forem sorteadas 200 amostras, cada uma delas contendo 2 habitantes desse país, e se forem calculadas as 200 respectivas médias amostrais, a partir delas obteremos o histograma a seguir:



Agora, cada uma das 200 amostras sorteadas contendo 30 habitantes desse país, e se forem calculadas as 200 médias amostrais, o histograma seria :



Exemplo 8.3: Simulando o efeito do Teorema Central do Limite (TCL)

Como pode ser observado, no caso de n = 2 o histograma se aproxima mais de uma curva Normal do que no caso de n = 1. E no caso de n = 30, a semelhança do histograma com uma curva Normal é ainda maior.

O Teorema Central do Limite afirma que, independentemente de qual seja a distribuição original dos X_i's, a distribuição de probabilidade de $\overline{X}_n$ e a distribuição Normal com média μ e variância σ^2/n se aproximam cada vez mais uma da outra, à medida que n cresce.

Portanto, mesmo que a distribuição de probabilidade dos Xi's seja desconhecida, o Teorema Central do Limite garante a possibilidade de usarmos o modelo Normal para calcular, ainda que de forma aproximada, probabilidades relativas à média amostral , desde que n seja suficientemente grande.

Simulando o efeito do TCL

Para ilustrar o funcionamento do Teorema Central do Limite, vamos exibir agora um exemplo em que a distribuição original a partir da qual os dados são gerados é uma exponencial, modelo este que dá origem a uma função densidade bastante assimétrica (ao contrário do que ocorre com a curva Normal).

A densidade de uma exponencial com parâmetro λ é dada pela expressão:

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

$E(X) = DP(X) = 1/\lambda$

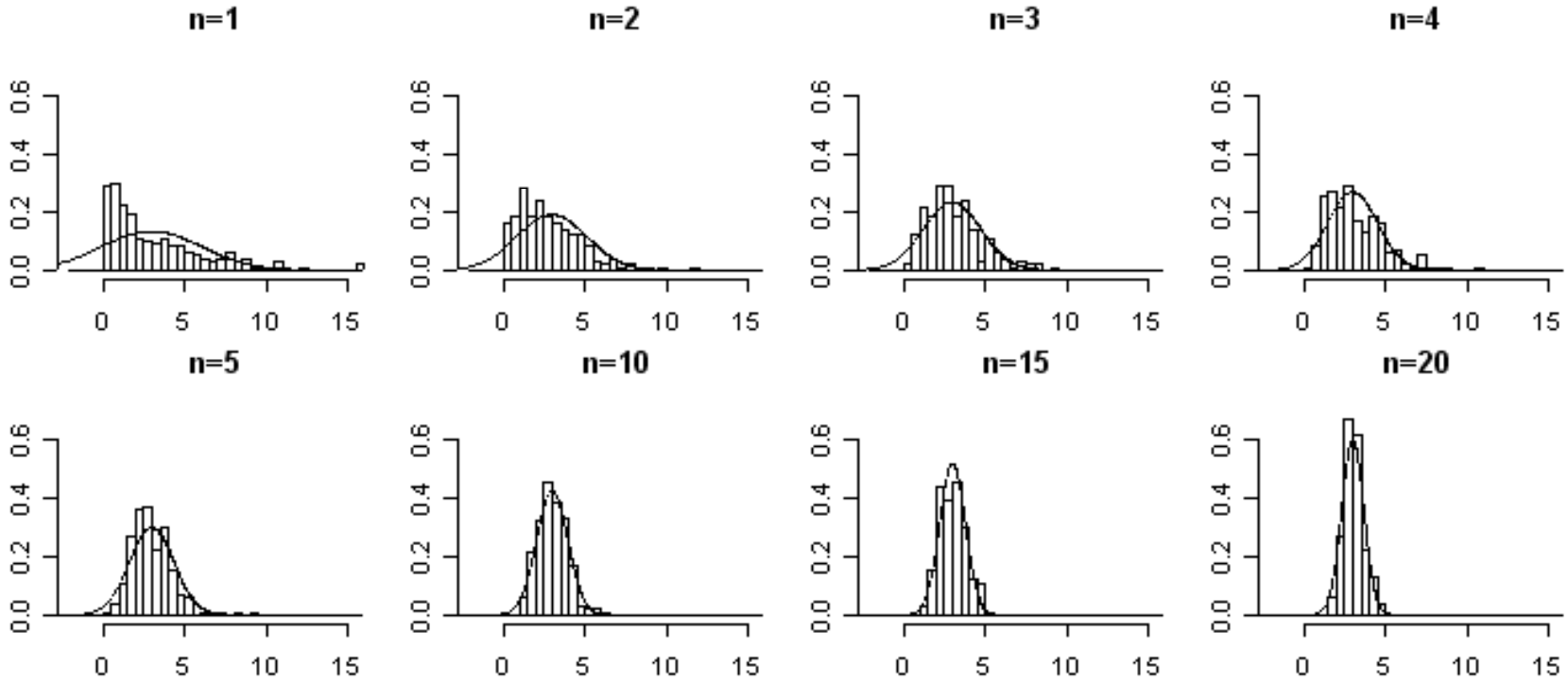
No R, **rexp(n,λ)** simula n valores

TCL: Exemplo

A densidade de uma exponencial com parâmetro λ é dada pela expressão: $f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$

Gerando dados por simulação a partir de uma exponencial com λ = 1/3, para cada um dos seguintes tamanhos n de amostra: 1, 2, 3, 4, 5, 10, 15 e 20,

1. Obtivemos 200 valores da média amostral ;
2. Utilizamos esses 200 valores para construir um histograma;
3. Traçamos no mesmo gráfico uma curva da densidade Normal com $E(\overline{X}_n)=3$ e $DP(\overline{X}_n)=3/\sqrt{n}$



Os 8 histogramas nos mostram que, à medida que o tamanho n da amostra cresce, a forma do histograma se aproxima cada vez mais de uma curva Normal.

TCL: Códigos no R para elaboração da figura com simulações - Exponencial

```
tcl.exp=function(n, N=200, titulo=" ", yl=c(0, .4)) {      ## início da função – tcl.exp
        medias=numeric(N)
        for (i in 1:N)  medias[i]= mean(rexp(n,1/3))
                        hist(medias, xlim=c(-1,10), ylim=yl, freq=F, main=titulo)
        x=seq(-1,10, .02)
        points(x, dnorm(x, 3, 3*sqrt(1/n) ), type="l", lwd=3)
}                                                          ## fim da função

        graphics.off()
par(mfrow=c(2,4), mai=c(.3,.4,.1,.1))
tcl.exp(1,titulo="n=1")
tcl.exp(2,titulo="n=2")
tcl.exp(3,titulo="n=3")
tcl.exp(4,titulo="n=4")

tcl.exp(5,titulo="n=5")
tcl.exp(10,titulo="n=10",yl=c(0,.6))
tcl.exp(15,titulo="n=15",yl=c(0,.6))
tcl.exp(20,titulo="n=20",yl=c(0,.6))
```

Uma pergunta natural neste ponto seria: “Quão grande deve ser n para que possamos usar a aproximação fornecida pelo TCL com um nível de precisão aceitável?”

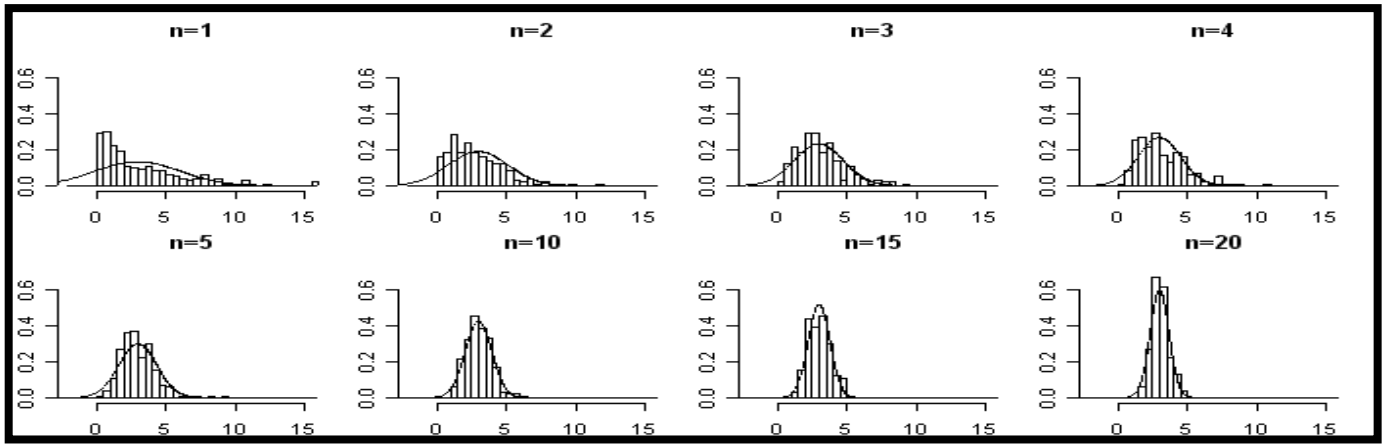
A rapidez com que essa convergência se dá depende de quão distante está a forma da distribuição original das Xi’s de uma curva Normal. Em outras palavras, se a distribuição das Xi’s já não for muito diferente de uma Normal, com um n não muito grande consegue-se uma boa aproximação. Caso contrário, somente para n bem grande (usualmente, n ≥ 30) a aproximação da distribuição de por uma Normal funcionaria adequadamente.

No exemplo a seguir vamos apresentar esse fenômeno, a saber, a convergência da distribuição de $\overline{X}_n$ para uma Normal à medida que n cresce, gerando por simulação os dados originais a partir de diferentes modelos probabilísticos. Em todos os casos, a distribuição original é bem diferente da Normal, E(X)=3 e DP(X)=3. No que se refere à Simulação, foi seguida a mesma seqüência de passos do exemplo anterior.

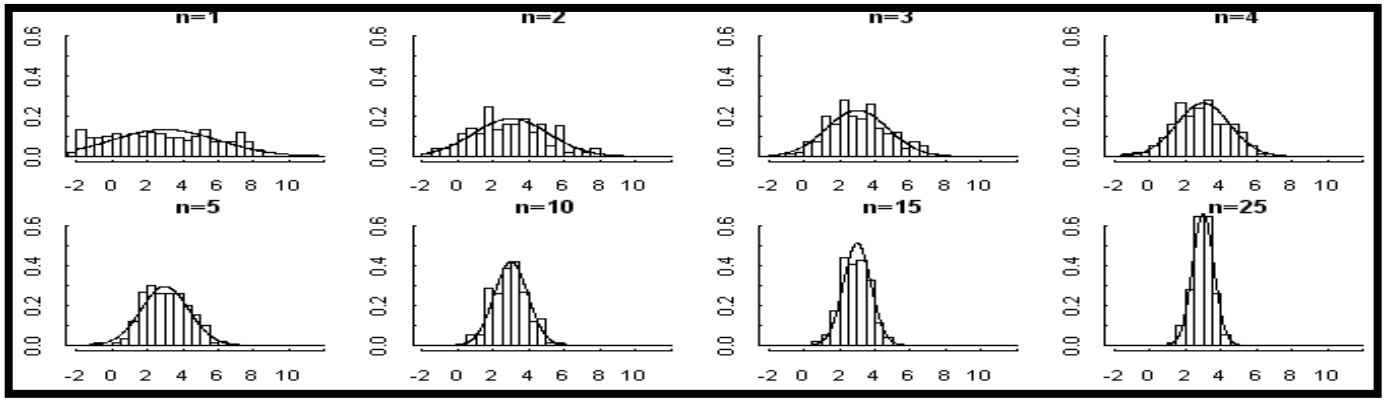
Gerando dados por simulação a partir de uma exponencial com $\lambda = 1/3$, para cada um dos seguintes tamanhos n de amostra: 1, 2, 3, 4, 5, 10, 15 e 20,

1. Obtivemos 200 valores da média amostral ;
2. Utilizamos esses 200 valores para construir um histograma;
3. Traçamos no mesmo gráfico uma curva da densidade Normal com $E(\overline{X}_n)=3$ e $DP(\overline{X}_n)=3/\sqrt{n}$

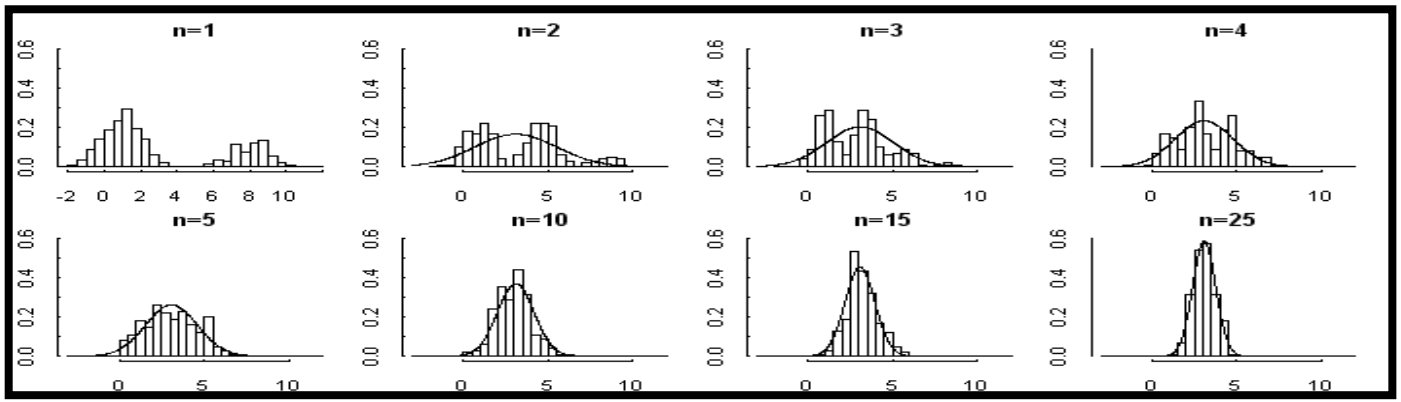
Cap. 3 – TCL: Exemplo



Exponencial



Uniforme



Mistura de Normais

TCL: Exemplo

Como se pode observar:

1. No caso da distribuição uniforme (A), o histograma de já se aproxima bastante de uma Normal quando n é da ordem de 4.

2. Já no caso da distribuição Exponencial (B) e da mistura de normais (C), modelos esses que se afastam muito mais de um "comportamento gaussiano", a aproximação pela Normal só se mostra mais adequada a partir de n em torno de 10.

3. No caso do modelo em (C), à medida que n cresce, tudo se passa como se houvesse a "erupção de um vulcão dentro do vale".

TCL: Códigos no R para elaboração da figura com as simulações - Uniforme

```
tcl.unif=function(n,N=100,titulo=" ", yl=c(0, .4)) {
medias=numeric(N)
  for (i in 1:N) medias[i]= mean(runif(n, 3-3*sqrt(3), 3+3*sqrt(3)))
  hist(medias, xlim=c(-6,10), ylim=yl, freq=F, main=titulo)
  x=seq(-6,10, .02)
  points(x, dnorm(x, 3, 3*sqrt(1/n) ), type="l", lwd=3)
####medias
}
graphics.off()
par(mfrow=c(2,4))                                    #####, mai=c(.3,.4,.1,.1))
tcl.unif(1,titulo="n=1",yl=c(0,.6))
tcl.unif(2,titulo="n=2",yl=c(0,.6))
tcl.unif(3,titulo="n=3",yl=c(0,.6))
tcl.unif(4,titulo="n=4",yl=c(0,.6))
tcl.unif(5,titulo="n=5",yl=c(0,.6))
tcl.unif(10,titulo="n=10",yl=c(0,.6))
tcl.unif(15,titulo="n=15",yl=c(0,.6))
tcl.unif(20,titulo="n=20",yl=c(0,.6))
```

8.4 A Variância e o Desvio Padrão amostrais

A **variância amostral** corresponde à variância definida no Capítulo 7 para um conjunto de dados quantitativos:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

A variância amostral, S^2, é uma **estatística de dispersão de uma amostra aleatória**. Um dos seus possíveis valores, que representaremos por s^2, corresponde à **variância amostral** definida no Capítulo 7 para um conjunto de dados quantitativos.

O **desvio-padrão amostral, S**, como a raiz quadrada não negativa de S^2.
Um valor que S venha a assumir é referido como **s**, ou seja, $\sqrt{s^2}$.

- Como acontece com todo estimador, **S^2 e S possuem distribuições amostrais**. Entretanto, não entraremos em detalhes sobre essas distribuições.
- Limitar-nos-emos a mencionar que **S^2 (a menos de uma constante) tem distribuição Qui-quadrado, desde que as X_i's sejam Normais** (ver Exercício P11.13).
- A distribuição Qui-quadrado será apresentada no capítulo 11, no contexto de Teste de hipóteses.

8.6 A Proporção Amostral

Considere uma amostra aleatória com **n** elementos extraída de uma determinada população e suponha que, entre eles, **Y** elementos possuam uma determinada característica de interesse. A proporção amostral a ela correspondente é dada por: $\hat{p} = \frac{Y}{n}$

Note que **Y é uma variável aleatória**. Portanto, $\hat{p}$ **também é uma v.a**.

A proporção nada mais é do que um caso particular da média, em que a variável considerada é do tipo 0 ou 1.

Muito do que é válido para a média amostral vale também para a proporção amostral, conforme veremos nas próximas seções.

A distribuição de probabilidade da proporção amostral

Admitamos que a população de interesse seja infinita ou muito grande. Então o processo de amostragem pode ser representado por n v.a. $X_1, X_2, X_3, ..., X_n$ iid tais que:

$X_i = 1$, se o elemento i possui a característica de interesse e

$X_i = 0$, caso contrário,

$P(X_i=1)=p$. Onde, p é a probabilidade de sucesso, "**possuir a característica**", o que significa que cada variável X_i tem distribuição de **Bernoulli com parâmetro p**.

Assim, $E(X_i) = p$ e $Var(X_i) = p(1-p)$, pelo **TCL:**

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} \sim N\left(E(\hat{p}) = p, Var(\hat{p}) = \frac{p(1-p)}{n}\right)$$

8.7.3 O Erro Quadrático Médio

Um bom estimador $\hat{\theta}$ é que ele seja não tendencioso, mas,

Ele deve gerar estimativas que estejam próximas do verdadeiro valor do parâmetro θ.

Mais importante do que ser não tendencioso é ele ser um estimador preciso.

Como podemos medir a precisão de um estimador?

Uma medida muito usada do **grau de precisão** de $\hat{\theta}$ como estimador de θ é o seu **erro quadrático médio**, definido por: $EQM(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$.

Quanto **menor** for $EQM(\hat{\theta})$, **mais preciso** será o estimador.

Pode-se provar que

$$EQM(\hat{\theta}) = Var(\hat{\theta}) + [B(\hat{\theta})]^2,$$

Logo, tanto uma variância grande quanto um viés grande (em módulo) podem prejudicar a precisão do estimador.

Observe que:

- $Var(\hat{\theta})$ é uma medida da variabilidade de $\hat{\theta}$ em torno da sua esperança $E(\hat{\theta})$;

enquanto que

- $[B(\hat{\theta})]^2$ é uma medida do afastamento entre a esperança $E(\hat{\theta})$ do estimador e o valor real θ do parâmetro.

se $\hat{\theta}$ é um estimador não tendencioso, $EQM(\hat{\theta}) = Var(\hat{\theta})$.

Exemplo 8.10: Tiro ao alvo, uma analogia com o processo de estimação de parâmetros

Consideremos um exercício de tiro ao alvo do qual participam 3 atiradores/estimadores A, B e C. Cada um deles teve direito a vários tiros. A figura abaixo mostra o desempenho de cada um dos atiradores.

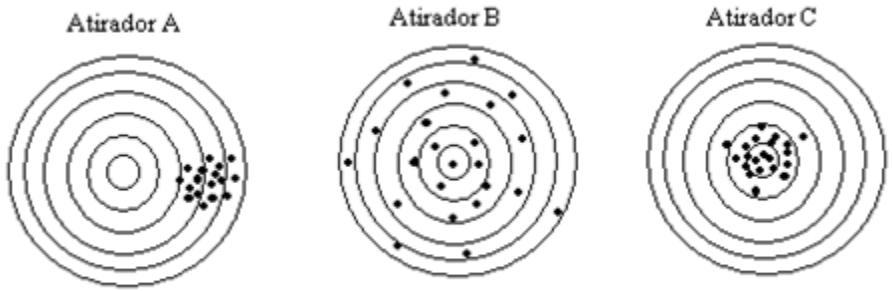


Figura 8.5 - Atiradores vistos como estimadores pontuais

Fazendo uma analogia entre "**atirar em um alvo**" e "**estimar um parâmetro**", pensemos em cada um dos 3 atiradores como um **estimador**. À luz dessa analogia:

- Quanto mais os tiros de um atirador se aproximarem da mosca, mais preciso ele estará sendo enquanto estimador;
- Quanto menos trêmulo for o seu braço na hora de atirar, menor será a sua variância enquanto estimador;
- Quanto mais livre de distorções for a sua visão do local onde está a mosca, menor será o seu viés enquanto estimador.

Exemplo 8.10: Tiro ao alvo, uma analogia com o processo de estimação de parâmetros

Consideremos um exercício de tiro ao alvo do qual participam 3 atiradores/estimadores A, B e C. Cada um deles teve direito a vários tiros. A figura abaixo mostra o desempenho de cada um dos atiradores.

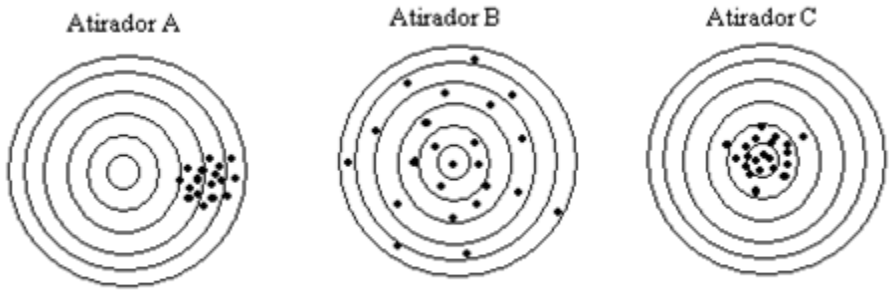


Figura 8.5 - Atiradores vistos como estimadores pontuais

O atirador A tem uma firmeza razoável nas mãos (baixa variância), mas tem problemas de visão (viés alto), o que compromete a sua precisão (EQM alto).

O atirador B (não tendencioso), mas tem as mãos muito trêmulas na hora de atirar (alta variância), o que compromete a sua precisão (EQM alto).

O atirador C tem uma firmeza razoável nas mãos (baixa variância) e visão perfeita (viés não tendencioso), o que lhe garante boa precisão (EQM baixo). Por isso ele é o melhor entre os 3 atiradores/estimadores.

Pergunta:
Como ficaria a performance de um atirador correspondente a um estimador tendencioso com variância nula?

8.8 - Dimensionamento da amostra

Qual deve ser o tamanho da minha amostra? Essa é certamente uma das dúvidas mais freqüentes de quem vai iniciar uma pesquisa envolvendo coleta de dados.

Quanto maior for a amostra, em princípio, mais precisas serão as estimativas, mas o custo tende a aumentar.

Uma das soluções mais conhecidas é arbitrar que o tamanho da amostra seja uma determinada fração do tamanho da população. Entretanto esta é uma alternativa demasiadamente simplista. A seguir mostramos que, em tais situações, a teoria também pode nos ajudar a tomar uma decisão quanto ao tamanho da amostra a ser utilizada, se formos capazes de especificar o nível de precisão desejado no processo de estimação do parâmetro μ em estudo.

É claro que não basta preocupar-se com o tamanho da amostra. É também de fundamental importância que o procedimento usado na seleção dos elementos que irão compor a amostra nos garanta uma boa representatividade.

Na discussão a seguir, estamos admitindo que seja usada a Amostragem Aleatória Simples, (sabidamente um bom procedimento amostral). Portanto, é suficiente que nos preocupemos apenas em dimensionar corretamente o tamanho da amostra.

Para dimensionarmos a amostra devemos fixar 2 constantes:

1) **d**, a distância máxima considerada tolerável entre a estimativa e o parâmetro
2) α, a probabilidade de essa distância ultrapassar d.

Dimensionamento de Amostra

Digamos que, em uma dada situação, se pretende usar a média amostral como estimativa da média populacional de uma certa variável.

Qual deveria ser o tamanho n da amostra a ser utilizada para que se possa garantir uma boa precisão na estimativa?

$$\left|\overline{X} - \mu\right| < d$$

Suponha que μ e σ são respectivamente a média e o desvio padrão populacionais.

Admita também que, nesse processo de estimação, o erro absoluto máximo considerado tolerável com uma probabilidade pré-fixada $1 - \alpha$, é igual a d, ou seja:

$$P\left[\left|\overline{X} - \mu\right| < d\right] = 1 - \alpha.$$

Dimensionamento de Amostra (Cont.)

Como $Var(\overline{X}) = \frac{\sigma^2}{n}$, admitindo que n é suficientemente grande para que o Teorema Central do Limite seja aplicável, temos

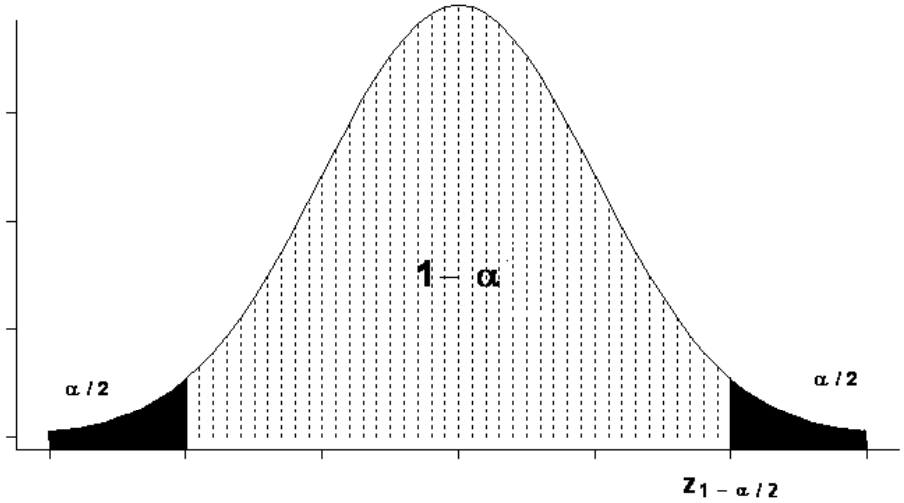
$$P\left[\frac{|\overline{X} - \mu|}{\sigma/\sqrt{n}} < \frac{d}{\sigma/\sqrt{n}}\right] = 1 - \alpha$$

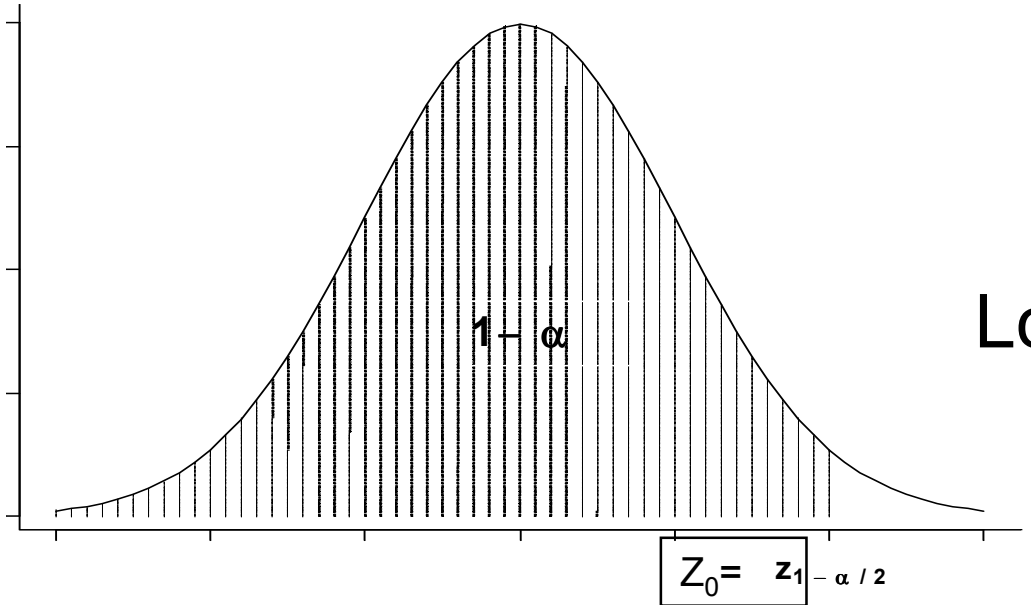
Então, se $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$, esta v.a. tem distribuição aproximadamente Normal(0;1), e a igualdade acima implica que

$$z_{1-\frac{\alpha}{2}} = \frac{d}{\sigma/\sqrt{n}}$$

onde $z_{1-\frac{\alpha}{2}}$ é o quantil $1 - \alpha/2$ da Normal(0,1)(*)

Figura - O quantil $z_{1-\alpha/2}$ da Normal(0,1)





Logo n = $\left(\frac{z_0 \cdot \sigma}{d} \right)^2$

8.8.1 Dimensionando a amostra para estimar a média populacional, com σ conhecido

Conforme afirmamos acima, para calcular o tamanho n da amostra devemos fixar d e α, ambos pequenos e tais que:

$$\mathbf{P}\left[\ |\bar{X}-\mu|\ > d\ \right]=\alpha\text{, ou equivalentemente, }\ \mathbf{P}\left[\ |\bar{X}-\mu|\le d\ \right]=1-\alpha.$$

$$\mathbf{P}\left[\ |\bar{X}-\mu|\le d\ \right]=1-\alpha\ \Rightarrow\ P\left[|Z|\le\frac{d}{\sigma/\sqrt{n}}\right]=1-\alpha\ \Rightarrow\ z_{1-\frac{\alpha}{2}}=\frac{d}{\sigma/\sqrt{n}}\ \Rightarrow\ n=\left(\frac{z_{1-\frac{\alpha}{2}}\cdot\sigma}{d}\right)^2$$

Exemplo 8.14: Pesquisa de clima interno (cont.)

Voltando ao caso da pesquisa de opinião do exemplo 8.12, qual deveria ser o tamanho n da amostra de empregados a serem entrevistados para que o erro absoluto na estimação do índice de satisfação médio estivesse limitado em 1,5 unidades com a mesma probabilidade de antes, isto é, 92,81%?

Solução:

$1-\alpha=0{,}9281$ implica em $z_{1-\frac{\alpha}{2}}=1{,}8$.

Então, como aqui temos d = 1,5, o tamanho da nova amostra teria que ser

$$n=\left(\frac{1{,}8\times 30}{1.5}\right)^2\cong\ 1296\text{ empregados.}$$

Isso significa que para a tolerância máxima do erro absoluto cair à metade do que era antes, o tamanho da amostra terá de quadruplicar.

8.8.2 Dimensionando a amostra para estimar a média populacional, com σ desconhecido

E se o desvio padrão σ for desconhecido, como é o caso em várias situações concretas? Afinal, se não sabemos nem o valor da média, não seria de se estranhar que também não soubéssemos o valor do desvio padrão! Como contornar essa dificuldade?

Em tais situações um caminho possível é extrair inicialmente uma pequena amostra piloto com, digamos, n_1 elementos, com base na qual se pode obter uma estimativa preliminar do desvio padrão σ, usando o seguinte estimador: $s_1 = \sqrt{\frac{\sum_{i=1}^{n_1}(x_i - \bar{x})^2}{n_1 - 1}}$

Substituindo σ por sua estimativa s_1 na fórmula $n = \left(\frac{z_{1-\frac{\alpha}{2}} \cdot \sigma}{d}\right)^2$, calcula-se então o tamanho da amostra a ser utilizada para que sejam atendidas as especificações de precisão: $n = \left(\frac{z_{1-\frac{\alpha}{2}} \cdot s_1}{d}\right)^2$

Depois de calcular o tamanho definitivo da amostra, pela fórmula acima, se for o caso, a amostra poderá ser complementada com a seleção de mais $(n - n_1)$ elementos.

8.8.4 Dimensionamento de Amostra para estimar a proporção populacional

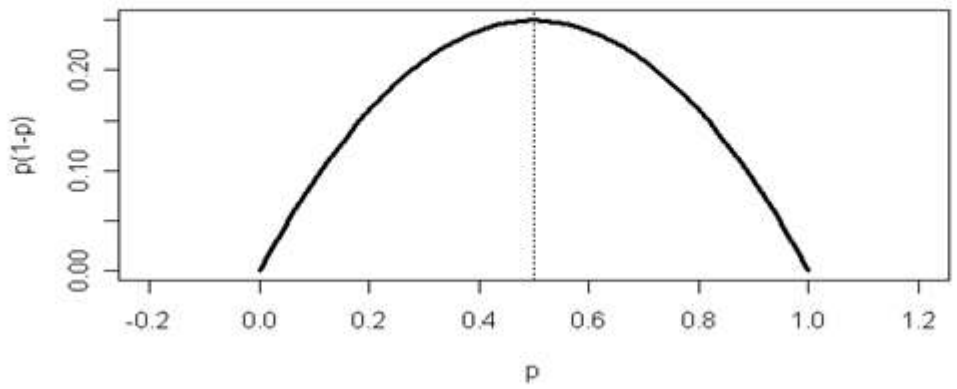
Quando o objetivo é estimar a proporção populacional, qual deve ser o tamanho n da amostra para que a probabilidade do erro absoluto exceder a tolerância d, esteja limitada por um certo α, onde d e α são ambos especificados pelo usuário?

Equivalentemente, para garantir que $P\left[\left|\hat{p} - p\right| \leq d\right] = 1 - \alpha$,

O tamanho da amostra deve ser $n = \left(\frac{z_{1-\frac{\alpha}{2}}}{d}\right)^2 p(1-p).$

n depende do próprio p, valor que desconhecemos e desejamos estimar.

A figura nos mostra que a função **p(1 – p)** atinge o seu ponto de máximo quando **p = 0,5**.



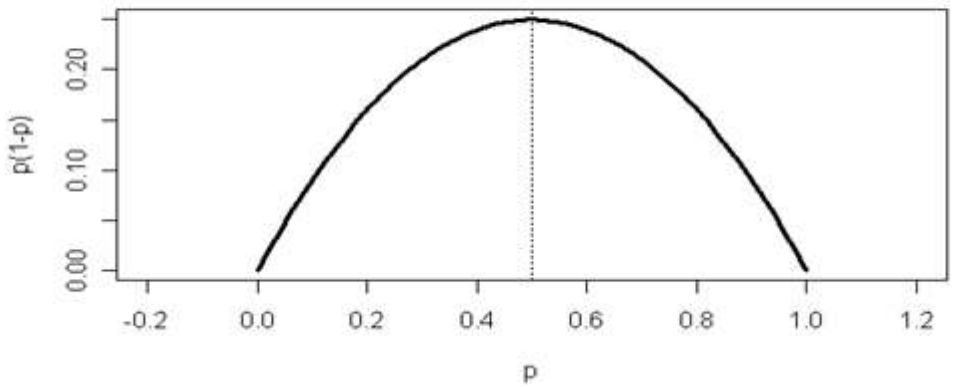
Concluímos que quanto mais próximo p estiver de 0,50, maior terá de ser o tamanho n da amostra para que seja atendida a especificação de precisão. Portanto podemos adotar o seguinte critério:

- Se nada sabemos acerca do valor verdadeiro de p, ou se sabemos que o intervalo de valores possíveis de p inclui o valor 0,5, usaremos como estimativa prévia p = 0,5.

- Se, com base em conhecimentos anteriores, a informação que temos sobre p é de que 0,5 não está entre os valores possíveis desse parâmetro, devemos usar o valor mais próximo de 0,5. Por exemplo, se a informação é de que p está entre 0,2 e 0,3, usamos o valor p = 0,3. Se, por outro lado, a informação é de que o verdadeiro valor de p é um número entre 0,6 e 0,8, o valor a ser usado é p = 0,6.

n depende do próprio p, valor que desconhecemos e desejamos estimar.

A figura nos mostra que a função **p(1 – p)** atinge o seu ponto de máximo quando **p = 0,5**.



Exemplo 8.16: Revisitando os diodos fora das especificações

Suponha que, no caso dos diodos, o objetivo é dimensionar a amostra a fim de termos uma probabilidade grande, digamos 0,90, de que o erro absoluto da estimativa da fração de diodos fora das especificações não ultrapasse 0,05. Qual deveria ser o tamanho amostral:

a) se não tivermos qualquer informação sobre o verdadeiro valor de p?

b) se temos a informação de que o verdadeiro valor da fração fora das especificações é inferior a 20%?

Solução

Temos d= 0,05 e para 1-α = 0,90 temos $z_{1-\frac{\alpha}{2}}$ = 1,64

a) Como nada sabemos sobre o possível valor da verdadeira fração de diodos fora das especificações usamos, como estimativa prévia, p = 0,5.

Temos, então $n = \left(\frac{1,64}{0,05}\right)^2 0,5(1-0,5) = 268,96 \Rightarrow n = 269 \text{ diodos.}$

b) Neste caso existe a informação de que o verdadeiro valor de p não ultrapassa 20%. Por isso, como estimativa prévia para calcular n usamos p = 0,2 . Daí:

$n = \left(\frac{1,64}{0,05}\right)^2 0,2(1-0,2) = 172,13 \Rightarrow n = 173 \text{ diodos.}$

Notas:

1. No caso do item (b), onde há uma informação prévia sobre o possível valor de p, o tamanho da amostra, necessário para se obter o mesmo nível de precisão na estimativa, é menor do que o calculado no item (a).
2. Na determinação do tamanho da amostra, por motivo de segurança, recomenda-se que o arredondamento seja feito sempre para cima.