

Capítulo 1

Introdução

1.1 Processo de Aprendizagem

Todo processo de aprendizagem consiste em formular uma hipótese (H_1), deduzir as conseqüências desta hipótese a partir de *dedução*, comparar estas conseqüências com dados obtidos (experimentação ou pesquisa), se houver discrepâncias, utilizar os dados para formular novas hipóteses num processo de *indução* e assim por diante.

Exemplo 1.1.1. *Uma professora de escola primária pergunta às crianças o que acontecerá se ela misturar tinta guache das cores amarela e azul. Uma das crianças disse que H_1 : a mistura ficaria manchada. Quando a professora mistura as duas cores, obtém-se verde (dados), com isso as crianças aprendem.*

Exemplo 1.1.2. *As crianças querem obter a cor laranja mas somente possuem as cores azul, amarelo e vermelho.*

- H_1 : misturando-se azul e amarelo obtém-se laranja.
- Experimento 1: misturam-se as tintas de cores azul e amarelo.
- Dados: Obtém-se a cor verde.
- H_2 : misturando-se azul e vermelho obtém-se laranja.
- Experimento 2: misturam-se as tintas de cores azul e vermelho.
- Dados: Obtém-se a cor roxa.
- H_3 : misturando-se vermelho e amarelo obtém-se laranja.
- Experimento 3: misturam-se as tintas de cores vermelho e amarelo.
- Dados: Obtém-se a cor laranja.

Exemplo 1.1.3. *Um biólogo decide verificar as conseqüências de choques térmicos em um dos vetores de doenças de Chagas conhecido por *Panstrongylus megistus*.*

- H_1 : os insetos teriam uma queda na sobrevivência quando submetidos a choques de 35^0 Celsius.
- Experimento 1: O biólogo toma 100 espécimes adultos e submete 50 deles a um choque de 35^0 Celsius (grupo tratamento) e mantém os outros 50 insetos em temperatura ambiente (grupo controle). Os insetos são observados durante 30 dias.
- Dados: No final de 30 dias, morreram 15 insetos do grupo tratamento e 13 insetos do grupo controle.

- H_2 : os insetos teriam uma queda na sobrevivência quando submetidos a choques de 40^0 Celsius.
- Experimento 2: O biólogo toma 100 espécimes adultos e submete 50 deles a um choque de 40^0 Celsius (grupo tratamento) e mantém os outros 50 insetos em temperatura ambiente (grupo controle). Os insetos são observados durante 30 dias.
- Dados: No final de 30 dias, morreram 22 insetos do grupo tratamento e 15 insetos do grupo controle.
- H_3 : os insetos teriam uma queda na sobrevivência quando submetidos a choques de 45^0 Celsius.
- Experimento 3: O biólogo toma 100 espécimes adultos e submete 50 deles a um choque de 45^0 Celsius (grupo tratamento) e mantém os outros 50 insetos em temperatura ambiente (grupo controle). Os insetos são observados durante 30 dias.
- Dados: No final de 30 dias, morreram todos insetos do grupo tratamento e 11 insetos do grupo controle.

Exemplo 1.1.4. *Um biólogo decide verificar as conseqüências de choques térmicos em um dos vetores de doenças de Chagas conhecido por *Panstrongylus megistus*.*

- H_1 : os insetos teriam uma queda na sobrevivência quando submetidos a choques de 35^0 , 40^0 e 45^0 Celsius.

- Experimento 1: O biólogo toma 200 espécimes adultos e submete 50 deles a um choque de 35° Celsius (grupo tratamento 1), 50 deles a um choque de 40° Celsius (grupo tratamento 2), 50 deles a um choque de 45° Celsius (grupo tratamento 3) e mantém os outros 50 insetos em temperatura ambiente (grupo controle). Os insetos são observados durante 30 dias.
- Dados: No final de 30 dias, morreram 14 insetos do grupo tratamento 1, 17 insetos no grupo tratamento 2, 50 insetos no grupo tratamento 3 e 13 insetos do grupo controle.

Exemplo 1.1.5. *A Data Folha quer determinar qual a popularidade do presidente Lula. Neste caso são entrevistadas 2000 pessoas em diversas cidades do Brasil.*

1.2 Dificuldades atenuadas através da utilização de métodos estatísticos

Há diversas fontes de confundimento quando trabalhamos com experimentação e coleta de dados.

Eventos aleatórios Existem fenômenos nos quais há uma aleatoriedade intrínseca. Antes de realizar o experimento não é possível dizer qual será o resultado obtido. O mais simples destes fenômenos é o lançamento de uma moeda. Antes de realizar o experimento é impossível (a menos que esta tenha duas caras ou duas coroas) prever qual será o resultado obtido.

1.2. DIFICULDADES ATENUADAS ATRAVÉS DA UTILIZAÇÃO DE MÉTODOS ESTATÍSTICOS

Quando queremos determinar se uma moeda é ou não honesta podemos lança-la 400 vezes. Será que mesmo sendo a moeda não viciada podemos obter 180 caras e 220 coroas?

Erro experimental O erro experimental é a variação produzida por fontes conhecidas ou desconhecidas. Efeitos importantes podem ser mascarados pelo erro experimental. Por outro lado, o erro experimental pode levar a indução de efeitos inexistentes.

Os efeitos do erro experimental podem ser grandemente minimizados através da utilização de amostragem adequada e técnicas de planejamento de experimentos.

Suponha que deseja-se determinar o conteúdo de um pacote de café coletado em um supermercado. Mesmo que pesemos o mesmo pacote 200 vezes na mesma balança, não obteremos 200 resultados idênticos. Dependendo da precisão da balança é provável obtermos 200 resultados distintos.

Confundimento entre relação e causa Os dados representados na Figura 2.3.1 consistem da população da cidade de Oldenburg ao final de cada ano do período 1930–1936 versus o número de cegonhas observadas naquele ano. Embora, ninguém diria com base nesse gráfico que as cegonhas são a causa do aumento da população, este tipo de erro é feito com muita freqüência em outros contextos. Considere outros exemplos, crianças com pés maiores soletram melhor, bairros com maiores taxas de divórcio tem taxas de mortalidade menores, países que acrescentam fluor na água potável têm

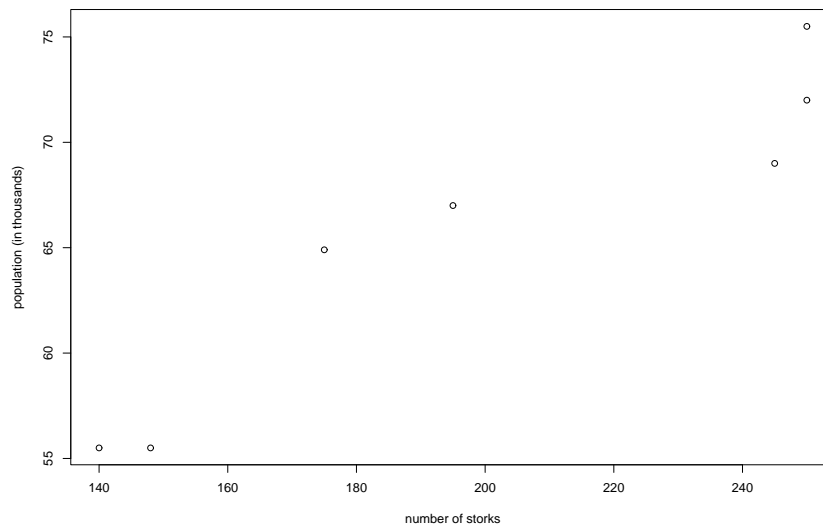
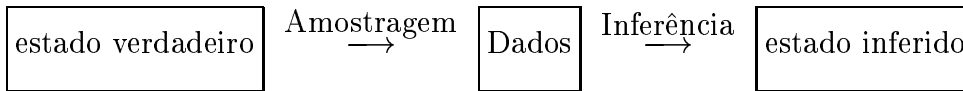


Figura 1.2.1: Gráfico da população de Oldenburg no final de cada ano no período de 1930–1936 versus o número de cegonhas avistadas durante o ano

maiores taxas de câncer. Embora, existam estudos que estabelecem todas estas relações, não podemos afirmar que exista causa e efeito. Nestes casos, provavelmente existe um terceiro fator não considerado na análise. Crianças com pés maiores tem maior habilidade em soletrar porque são mais velhas. Idade também é um fator a ser considerado no exemplo seguinte, como casais que são mais velhos são menos prováveis de se divorciarem e mais prováveis de morrerem do que aqueles em bairros onde o perfil demográfico é mais jovem. Nações que adicionam fluor na água são, em geral, mais ricas e mais conscienciosas em termos de saúde, e assim uma grande porcentagem de sua população vive o suficiente para adquirir câncer, a qual é uma doença predominantemente de idosos.

Complexidade dos efeitos estudados Suponha que uma nutricionista deseja estimar o efeito de 5 dietas para perder peso. Existe uma dieta padrão (controle) e as outras 4 dietas consistem em combinar 2 níveis de proteína e 2 níveis de carboidratos. Além disso, ela deseja estudar o efeito da dieta em homens e mulheres. Note que nesse caso, existem os efeitos de diversos fatores, proteínas, carboidratos, sexo e dieta/controle. Além disso, existe a interação entre os fatores. Por exemplo, as dietas fazem efeitos diferentes para homens e mulheres. Este tipo de problema pode ser “facilmente” contornado utilizando-se um bom planejamento de experimentos.

1.3 Objetivo da Inferência



Exemplo 1.3.1. Pesquisas de opinião

A fim de descobrir quanto é impopular a CPMF inicialmente concebida a fim de arrecadar recursos para a saúde, a DataFolha entrevistou 2000 pessoas e descobriu que 1740 destas pessoas são contrárias à volta do imposto. Isto nos diz que, nesta amostra, 87% das pessoas são contrárias ao imposto. O que isto diz a respeito da população do Brasil? Será que a proporção de pessoas contrárias à CPMF é próxima de 87%? Se sim, qual a margem de erro desta estimativa? Será que há evidência significativa que o valor verdadeiro é maior que 85%?

Exemplo 1.3.2. *Pesquisas de opinião.*

A fim de verificar a popularidade do governo Lula após um ano de mandato, a Datafolha realizou uma pesquisa entre os dias 8 a 12 e no dia 15 de dezembro em 396 cidades brasileiras. Foram entrevistados 12.180 brasileiros, e a margem de erro máxima para o levantamento é de dois pontos percentuais, para mais ou para menos. O desempenho do presidente Luiz Inácio Lula da Silva é aprovado por 42%, considerado regular por 41% e ruim ou péssimo por 15%.

Exemplo 1.3.3. *Confiabilidade*

Uma indústria produtora de fornos de microondas gostaria de saber quanto tempo deve-se dar de garantia a seus produtos de modo que somente 1% de seus produtos devam ser reparados no prazo de garantia. Através de testes acelerados, pode-se obter os tempos de vida de 50 fornos de microondas e com base nestes dados estimar o prazo de garantia.

Exemplo 1.3.4. *Controle de qualidade*

Uma indústria é fornecedora de rolamentos para indústria automobilística, os limites de especificação para rolamentos são 15mm e 18mm, peças fora da especificação são refugadas. Devemos então com base em uma amostra descobrir um intervalo de valores plausíveis para os rolamentos de modo que possamos saber se "grande" parte dos rolamentos produzidos estejam dentro dos limites de especificação. Se descobrimos que este intervalo é de (15.8; 17.2) podemos ficar tranquilos, mas se este intervalo for (12.0; 20.0) devemos

nos preocupar em diminuir a variabilidade do processo.

Exemplo 1.3.5. *Melhoramento industrial*

Uma indústria de sucos concentrados está desenvolvendo um novo procedimento para retirar a água do suco de laranja de modo que o suco reconstituído seja mais agradável ao paladar. O sabor do suco varia em uma escala de 1 à 10. Pelo procedimento atual este índice é uma variável aleatória normalmente distribuída com média 7 e desvio padrão 1. Como descobrir se o novo método é realmente melhor?

Exemplo 1.3.6. *Calibração*

A resistência à tração é uma característica internacionalmente utilizada para classificar a qualidade de ferros fundidos cinzentos, entretanto além do seu alto custo, pode ser impossível de ser determinada em alguns casos. Outro método de se determinar a resistência de ferros fundidos é a resistência à compressão entre cunhas. Como relacionar as duas medidas?

1.3.1 População e amostra:

O objetivo de uma investigação científica é descobrir (entender, estudar) alguma característica de certa população. Como, em geral, é impossível ou impraticável examinar toda a população, examinamos parte e com base nestes dados fazer inferências a respeito de toda população. Temos que distinguir entre população alvo e população amostrada.

População alvo: É aquela em que estamos interessados em estudar. No Exemplo 1.3.2, a população alvo são todos os eleitores do Brasil.

População amostrada: É aquela da qual retiramos a nossa amostra. No Exemplo 1.3.2, a população amostrada são adultos que moram nas 396 cidades selecionadas.

Se a população amostrada não é a mesma que a população alvo, as conclusões (inferências) obtidas através da amostra só são válidas para a população amostrada.

Pergunta: Como selecionar uma amostra?

- O tamanho da amostra é fixado através do nível de precisão desejado.
- Os requisitos de uma amostra é que seja finita, representativa da população, aleatória e não sistemática.

No Exemplo 1.3.1, o DataFolha decidiu amostrar 2000 pessoas e a partir de suas respostas inferir qual a proporção de pessoas contrárias à CPMF. Para simplificar e quantificar este exemplo defina:

$$X_i = \begin{cases} 1, & \text{se o } i\text{-ésimo entrevistado é contrário à CPMF} \\ 0, & \text{se o } i\text{-ésimo entrevistado é a favor à CPMF} \end{cases}$$

Após retirar a amostra temos disponíveis 2000 respostas 0 ou 1. Se $p =$ proporção de votos contrários à CPMF, como utilizar os 2000 dados para

estimar p ?

Note que $X_1, X_2, \dots, X_{2000}$ são variáveis aleatórias, não podemos prever seus valores antes de observar a amostra. Portanto $X_1, X_2, \dots, X_{2000}$ são variáveis aleatórias e têm uma distribuição conjunta.

Obs.:

Antes de retirar a amostra: $X_1, X_2, \dots, X_{2000}$ são variáveis aleatórias;

Depois de retirar a amostra: $x_1, x_2, \dots, x_{2000}$ são valores observados, realizações das v.a's.

Qual a distribuição conjunta de X_1, X_2, \dots, X_n ? Suponha que $n = 2$.

$$X_1 = \begin{cases} 1, & \text{se a primeira pessoa entrevistada é contrária à CPMF} \\ 0, & \text{caso contrário} \end{cases}$$

$$X_2 = \begin{cases} 1, & \text{se a segunda pessoa entrevistada é contrária à CPMF} \\ 0, & \text{caso contrário} \end{cases}$$

Neste caso, para $i = 1, 2$,

$$P(X_i = x) = p^x(1 - p)^{1-x}, x = 0 \text{ ou } 1.$$

Isto é, X_1 e X_2 são identicamente distribuídas.

Caso 1: Supondo que a amostra é feita com reposição:

$$P(X_1 = x, X_2 = y) = P(X_1 = x)P(X_2 = y).$$

Isto é, X_1 e X_2 são independentes.

Caso 2: Supondo que a amostra é feita sem reposição:

$$P(X_1 = x, X_2 = y) \neq P(X_1 = x)P(X_2 = y).$$

Isto é, X_1 e X_2 não são independentes.

Caso 2 é mais difícil de ser estudado. Vamos nos concentrar no Caso 1.

Definição 1.1. (1) X_1, X_2, \dots, X_n formam uma amostra aleatória de tamanho n de uma variável aleatória discreta X se, e somente se, a função de probabilidade conjunta de X_1, X_2, \dots, X_n é:

$$p_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n p_X(x_i).$$

Isto é, X_1, X_2, \dots, X_n são independentes e identicamente distribuídas (i.i.d.).

(2) X_1, X_2, \dots, X_n formam uma amostra aleatória de tamanho n de uma variável aleatória contínua X se, e somente se, a função de densidade conjunta de X_1, X_2, \dots, X_n é

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n f_X(x_i).$$

Isto é, X_1, X_2, \dots, X_n são independentes e identicamente distribuídas (i.i.d.).

Exemplo: Suponha que estamos interessados em estudar o tempo de vida X de certo componente eletrônico. Sabe-se que não há envelhecimento

neste componente, portanto há razões para se supor que X é exponencialmente distribuída com média $1/\theta$. Isto é,

$$f_X(x) = \theta e^{-\theta x}, x > 0.$$

A fim de determinar o valor de θ , 10 destes componentes foram selecionados ao acaso e seus tempos de vida anotados. Seja X_i = tempo de vida do i -ésimo componente eletrônico, $i = 1, \dots, 10$. Portanto, X_1, \dots, X_{10} formam uma amostra aleatória de tamanho 10 da distribuição exponencial com parâmetro θ .

Ache a densidade da amostra. Qual a probabilidade que todos os componentes vivam mais de 10 horas? Se todos os componentes viveram mais de 10 horas, você acharia razoável supor que $\theta = 1/5$?

1.4 Somas de Variáveis Aleatórias

Vamos agora tornar um pouco mais precisos alguns conceitos que utilizávamos intuitivamente a respeito de frequência relativa: “à medida que o número de repetições de um experimento cresce, a frequência relativa f_A de um evento A converge para a probabilidade teórica $P(A)$ ”. Por exemplo, se uma nova peça for ser produzida e não tivermos conhecimento à priori sobre quão provável a peça ser defeituosa (p = probabilidade de ser defeituosa é desconhecida), podemos proceder à inspeção de um grande número destas peças. Seja n = número de peças inspecionadas, X_1, \dots, X_n os indicadores das peças serem ou não defeituosas, isto é, X_1, \dots, X_n i.i.d. $b(1, p)$ e $\sum_{i=1}^n X_i$ = número de

peças defeituosas.

Portanto,

$$p = P(\text{peça defeituosa}) \approx \frac{\sum_{i=1}^n X_i}{n}.$$

Entretanto, $\sum_{i=1}^n X_i$ é variável aleatória, com $\sum_{i=1}^n X_i \sim b(n, p)$. Daí,

$$\frac{\sum_{i=1}^n X_i}{n} \rightarrow p \quad (\text{em algum sentido})$$

quando $n \rightarrow \infty$.

1.4.1 Lei dos Grandes Números

Lei Fraca dos Grandes Números (Bernoulli). Seja E um experimento e A um evento associado à E . Considere-se n repetições idênticas e independentes de E , seja N o número de vezes em que A ocorre nas n repetições. Seja $p = P(A)$ (a qual supõe-se seja a mesma para todas as repetições). Daí, $N \sim b(n, p)$ e

$$P(|N/n - p| \epsilon) \leq \frac{p(1-p)}{n\epsilon^2} \rightarrow 0$$

quando $n \rightarrow \infty$, para todo $\epsilon > 0$.

O resultado acima é muito fácil de ser provado utilizando-se a Desigualdade de Chebychev (**Exercício.**) Entretanto, pode-se facilmente verificar-se que este resultado não é restrito à variáveis aleatórias binomiais.

Lei Fraca dos Grandes Números: Seja X_1, X_2, \dots uma sequência de variáveis aleatórias i.i.d.. Sejam $\mu = E(X_i)$ e $\sigma^2 = \text{Var}(X_i)$, defina

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$$

então,

$$E(\bar{X}_n) = \mu, \quad \text{Var}(\bar{X}_n) = \sigma^2/n,$$

e também

$$P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \rightarrow 0$$

quando $n \rightarrow \infty$, para todo $\epsilon > 0$.

Exemplo: Um grande número de válvulas eletrônicas são testadas. Seja, T_i o tempo de vida da i -ésima válvula. suponha também que não há envelhecimento das peças e $T_i \sim \exp(\alpha)$. Portanto,

$$E(T_i) = 1/\alpha, \quad \text{Var}(T_i) = 1/\alpha^2.$$

Se definimos

$$\bar{T}_n = \frac{T_1 + \cdots + T_n}{n}$$

temos pela Lei Fraca dos Grandes Números

$$P(|\bar{T}_n - 1/\alpha| > \epsilon) \rightarrow 0$$

quando $n \rightarrow \infty$ para todo $\epsilon > 0$.

Ou seja, se o tamanho da amostra n é muito grande, será “muito provável” que o valor obtido para a média amostral esteja próximo de $1/\alpha$. **Exercício:** Quão provável?

1.4.2 Teorema Central do Limite

Não somente é importante saber que a média amostral “se aproxima” da média populacional, mas muito mais importante é sabermos quantificar a probabilidade de estarmos errando quando utilizamos a média amostral como uma estimativa para a média populacional. Para isto é necessário verificar-se a velocidade de convergência. Isto é, intuitivamente temos que

$$\frac{\sum_{i=1}^n X_i}{n} - \mu \rightarrow 0$$

em probabilidade. O que acontece se fizermos:

$$\sqrt{n} \left(\frac{\sum_{i=1}^n X_i}{n} - \mu \right)$$

este limite existe? Converge a uma constante? Converge a uma v.a.? Este é o teor do Teorema Central do Limite (TCL).

Teorema 1.4.1. *Seja X_1, X_2, \dots uma sequência de v.a.'s i.i.d. (i.e., uma amostra aleatória) com $E(X_i) = \mu$ e $Var(X_i) = \sigma^2$. Defina $S_n = X_1 + \dots + X_n$, então*

$$E(S_n) = n\mu, \quad Var(S_n) = n\sigma^2$$

e

$$Z_n = \frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} = \frac{S_n - n\mu}{\sqrt{n}\sigma} \rightarrow N(0, 1)$$

em distribuição. Isto é, se $G_n(z) = P(Z_n \leq z)$ então

$$\lim_{n \rightarrow \infty} G_n(z) = \Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

Exemplo: Seja X_1, X_2, \dots uma sequência de v.a.'s de Bernoulli independentes, ($P(X_i = 1) = p$). Então $S_n =$ número de sucessos em n ensaios de

Bernoulli independentes e

$$S_n \sim b(n, p).$$

Pelo Teorema Central do Limite (TCL),

$$\frac{S_n - np}{\sqrt{np(1-p)}} \rightarrow N(0, 1).$$

Suponha que somos produtores de arruelas, cerca de 5% das quais são defeituosas. Se num lote, 100 arruelas são inspecionadas, qual a probabilidade que pelo menos 4 sejam defeituosas?

S_{100} = número de arruelas defeituosas encontradas numa amostra de tamanho 100, temos

$$S_{100} \sim b(100, 0.05)$$

e

$$\begin{aligned} P(S_{100} \leq 4) &= \sum_{k=0}^4 \binom{100}{k} (0.05)^k (0.95)^{100-k} \\ &= P\left(\frac{S_{100} - 100 \times 0.05}{\sqrt{100 \times 0.05 \times 0.95}} \leq \frac{4 - 100 \times 0.05}{\sqrt{100 \times 0.05 \times 0.95}}\right) \\ &= P\left(Z_{100} \leq \frac{-1}{2.179}\right) = P(Z_{100} \leq -0.459) \\ &\approx \Phi(-0.459) = 0.3228. \end{aligned}$$

E se encontrarmos 8 defeituosos, ainda acreditamos que $p = 0.05$?

$$\begin{aligned} P(S_{100} \geq 8) &= 1 - P(S_{100} \leq 7) \\ &= 1 - P\left(\frac{S_{100} - 100 \times 0.05}{\sqrt{100 \times 0.05 \times 0.95}} \leq \frac{7 - 100 \times 0.05}{\sqrt{100 \times 0.05 \times 0.95}}\right) \\ &= 1 - P\left(Z_{100} \leq \frac{2}{2.179}\right) \\ &\approx 0.166. \end{aligned}$$

Note que,

$$Z_n = \frac{S_n - E(S_n)}{\sqrt{\text{Var}(S_n)}} = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{n(S_n/n - \mu)}{\sqrt{n}\sigma} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}.$$

1.4.3 V.a.'s t de Student, χ^2 e F de Snedecor

Definição 1.2. Se T é uma v.a. contínua com densidade

$$f_T(t) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{n\pi}} \left[1 + \frac{t^2}{n}\right]^{-(n+1)/2}, t \in \mathbf{R}$$

dizemos que T tem uma distribuição t de Student com n graus de liberdade.

Not.: $T \sim t(n)$.

Temos $E(T) = 0$, se $n > 1$ e $\text{Var}(T) = n/(n-2)$, se $n > 2$.

Obs.: $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx$.

Definição 1.3. Se Y é uma v.a. contínua, positiva ($P(Y > 0) = 1$) com densidade

$$f_Y(y) = \frac{1}{\Gamma(n/2)} \frac{1}{2^{n/2}} y^{(n/2)-1} e^{-y/2}, y > 0$$

dizemos que Y tem distribuição qui-quadrado com n graus de liberdade.

Not.: $Y \sim \chi^2(n)$.

Temos que $E(Y) = n$, e $\text{Var}(Y) = 2n$, função geradora de momentos $m_Y(t) = (1 - 2t)^{-n/2}$ para $t < 1/2$.

Definição 1.4. Se Z é uma v.a. contínua, positiva ($P(Z > 0) = 1$) com densidade

$$f_Z(z) = \frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)} \frac{(m/n)^{m/2} z^{(m-2)/2}}{(1+(mz/n))^{(m+n)/2}}, z > 0$$

dizemos que Z tem distribuição F de Snedecor com m graus de liberdade no numerador e com n graus de liberdade no denominador.

Not.: $Z \sim F(m, n)$.

Temos que $E(Z) = n/(n-2)$, para $n > 2$ e $\text{Var}(Z) = (2n^2(m+n-2))/(m(n-2)^2(n-4))$, para $n > 4$.

Todas as distribuições acima estão tabeladas.

Teorema 1.1. Seja X_1, \dots, X_n uma amostra aleatória de uma distribuição $N(\mu, \sigma^2)$, então

$$U = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2$$

tem distribuição $\chi^2(n)$.

Prova: Sejam as v.a.'s i.i.d.

$$Z_i = \frac{X_i - \mu}{\sigma} \sim N(0, 1)$$

e temos $U = \sum_{i=1}^n Z_i^2$. A qual tem função geradora de momentos,

$$\begin{aligned} m_U(t) &= E[e^{tU}] = E \left[e^{t \sum_{i=1}^n Z_i^2} \right] \\ &= E \left[\prod_{i=1}^n e^{t Z_i^2} \right] = \prod_{i=1}^n E[e^{t Z_i^2}] \end{aligned}$$

mas,

$$\begin{aligned}
 E[e^{tZ_i^2}] &= \int_{-\infty}^{+\infty} e^{tz^2} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\
 &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{(-1/2)(1-2t)z^2} dz \\
 &= \frac{1}{\sqrt{1-2t}} \underbrace{\int_{-\infty}^{+\infty} \frac{\sqrt{1-2t}}{\sqrt{2\pi}} e^{(-1/2)(1-2t)z^2} dz}_{=1}, \quad t < 1/2 \\
 &= \frac{1}{\sqrt{1-2t}}, \quad t < 1/2.
 \end{aligned}$$

Assim, $U \sim \chi^2(n)$ pois

$$m_U(t) = (1 - 2t)^{-n/2}.$$

Teorema 1.2. Se Z_1, Z_2, \dots são v.a.'s i.i.d. $N(0, 1)$. Temos

- (i) $\bar{Z}_n \sim N(0, 1/n)$;
- (ii) \bar{Z}_n e $\sum_{i=1}^n (Z_i - \bar{Z}_n)^2$ são v.a.'s independentes;
- (iii) $\sum_{i=1}^n (Z_i - \bar{Z}_n)^2 \sim \chi^2(n-1)$.

Prova: Ver Teorema 6, página 241 e Teorema 8, página 243 do livro Mood, Graybill and Boes.

Suponha que temos X_1, \dots, X_n uma amostra aleatória de uma distribuição $N(\mu, \sigma^2)$. Então

$$Z_i = \frac{X_i - \mu}{\sigma} \sim N(0, 1)$$

e por (i)

$$\begin{aligned}
 \frac{1}{n} \sum Z_i &= \frac{1}{n} \sum_{i=1}^n \frac{X_i - \mu}{\sigma} = \frac{1}{n} \frac{(\sum_{i=1}^n X_i - n\mu)}{\sigma} \\
 &= \frac{\bar{X} - \mu}{\sigma} \sim N(0, 1/n)
 \end{aligned}$$

Portanto,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Por (ii) temos que

$$\frac{\bar{X} - \mu}{\sigma} \text{ e } \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2}$$

são independentes e por (iii)

$$\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1).$$

Corolário 1.4.2. *Seja*

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

a variância amostral então

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Teorema 1.3. *Se $Z \sim N(0, 1)$ e $U \sim \chi^2(n)$ são v.a.'s independentes, então*

$$\frac{Z}{\sqrt{U/n}} \sim t(n).$$

Aplicação: Seja X_1, \dots, X_n uma amostra aleatória $N(\mu, \sigma^2)$, sabemos que

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

e

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

são v.a.'s independentes e portanto,

$$\begin{aligned} \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{(1/(n-1)) \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2}}} &= \frac{\bar{X} - \mu}{\sqrt{1/n} \sqrt{(1/(n-1)) \sum_{i=1}^n (X_i - \bar{X})^2}} \\ &= \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1). \end{aligned}$$

Então, se $n = 30$, pela tabela temos que

$$P(T > 1.699) = 0.05$$

e

$$P\left(\left|\frac{\bar{X} - \mu}{S/\sqrt{n}}\right| \leq 1.699\right) = 0.90$$

i.e.,

$$P\left(\bar{X} - 1.699\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.699\frac{S}{\sqrt{n}}\right) = 0.90.$$

Teorema 1.4. *Se $X \sim \chi^2(m)$ e $Y \sim \chi^2(n)$ são v.a.'s independentes então*

$$F = \frac{X/m}{Y/n} \sim F(m, n).$$

Aplicação: Seja X_1, \dots, X_m uma amostra aleatória $N(\mu_1, \sigma^2)$ e Y_1, \dots, Y_n uma amostra aleatória $N(\mu_2, \sigma^2)$ independentes. Se definimos

$$S_X^2 = \sum_{i=1}^m \frac{(X_i - \bar{X})^2}{m-1} \quad \text{e} \quad S_Y^2 = \sum_{i=1}^n \frac{(Y_i - \bar{Y})^2}{n-1}$$

então temos

$$\frac{(m-1)S_X^2}{\sigma^2} \sim \chi^2(m-1)$$

e

$$\frac{(n-1)S_Y^2}{\sigma^2} \sim \chi^2(n-1)$$

são independentes e

$$\frac{S_X^2}{S_Y^2} \sim F(m-1, n-1).$$

Capítulo 2

Estatísticas

Quando queremos estudar um fenômeno aleatório, devemos tirar uma amostra aleatória X_1, \dots, X_n da variável de interesse. Como estas características numéricas são aleatórias, o melhor que podemos fazer para descrevê-las, é descrever sua lei de probabilidade, se as v.a.'s são discretas isto é feito através de sua função de probabilidade. Se as v.a.'s são contínuas precisamos descrever a densidade de probabilidade. Primeiramente precisamos determinar a forma da distribuição. Isto é feito, através de considerações teóricas sobre o experimento em questão, por exemplo, se a distribuição é contínua, discreta, simétrica ou não, etc. Se isto não for possível é necessário utilizar inferência não paramétrica. Se podemos determinar a forma da distribuição, em geral, faltam alguns parâmetros numéricos que precisam ser determinados com base na amostra. Por exemplo, se estamos estudando tempo de vida de lâmpadas fluorescentes, podemos argumentar que o tempo de vida de lâmpadas é uma v.a. contínua, positiva e como não há envelhecimento pode ser considerada exponencial. Assim se tiramos uma a.a. X_1, \dots, X_n de uma distribuição

$\exp(\theta)$, falta ainda determinar θ , como isso não é possível de ser feito exatamente, utilizaremos a a.a. da “melhor forma possível” para estimar o parâmetro θ .

Definição 2.1. *Qualquer função dos elementos de uma amostra aleatória, a qual não depende de parâmetros desconhecidos, é chamada de estatística.*

Se X_1, X_2, \dots, X_n é uma amostra aleatória de uma distribuição com densidade (ou função de probabilidade) $f(x, \theta)$, então

$$X_1 + X_2, \frac{X_3}{X_4}, \bar{X}_n = \sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2, \prod_{i=1}^n X_i$$

$$S_X^2 = \sum_{i=1}^n (X_i - \bar{X})^2, \sum_{i=1}^n \log(X_i), \max(X_1, \dots, X_n), \min(X_1, \dots, X_n)$$

são estatísticas. Por outro lado, Se temos X_1, \dots, X_n i.i.d. $N(\mu, \sigma^2)$ temos que

$$X_1 - \mu, \bar{X} - \mu, \frac{\bar{X} - \mu}{\sigma}, \sum_{i=1}^n (X_i - \mu)^2$$

não são estatísticas., pois dependem de parâmetros desconhecidos μ e σ .

Obs.: Estatísticas são v.a.'s e portanto têm distribuição de probabilidade.

Por exemplo, se temos X_1, \dots, X_n i.i.d. $N(\mu, \sigma^2)$ então

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

é uma estatística e a sua distribuição é dada por:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Note que

$$T = \frac{\bar{X} - \mu}{S_X/\sqrt{n}} \sim t(n-1)$$

mas **não** é estatística.

2.0.4 Momentos amostrais

Denote por

$$\mu_k = \mathbf{E}(X^k)$$

o k -ésimo momento da v.a. X .

Definição 2.2. *Se X_1, \dots, X_n é a.a. com a mesma distribuição de X , o k -ésimo momento amostral é:*

$$M_k = \frac{1}{n} \sum_{i=1}^n X_i^k$$

para $k = 1, 2, \dots$

Note que para cada k fixo M_k é uma v.a. e é estatística.

Notação: m_k é o k -ésimo momento amostral observado (isto é, após retirarmos a amostra).

Obs.: Alguns momentos amostrais têm especial importância:

$$M_1 = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

é a média amostral e

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

é a variância amostral.

Teorema 2.1.

$$S^2 = \frac{n}{n-1} [M_2 - M_1^2].$$

Notação: $S = \sqrt{S^2}$ = desvio padrão amostral.

Distribuição dos momentos amostrais

Teorema 2.2. *Sejam X_1, \dots, X_n uma amostra aleatória de uma população X . Temos,*

$$\mathbf{E}[M_k] = \mu_k, k = 1, 2, \dots$$

e

$$\text{Var}[M_k] = \frac{1}{n} [\mu_{2k} - \mu_k^2].$$

Prova:

$$\begin{aligned} \mathbf{E}[M_k] &= \mathbf{E}\left[\frac{1}{n} \sum_{i=1}^n X_i^k\right] = \frac{1}{n} \sum_{i=1}^n \mathbf{E}[X_i^k] \\ &= \frac{1}{n} \sum_{i=1}^n \mu_k = \mu_k \\ \text{Var}[M_k] &= \text{Var}\left[\frac{1}{n} \sum_{i=1}^n X_i^k\right] = \frac{1}{n} \sum_{i=1}^n \text{Var}[X_i^k] \\ &= \frac{1}{n} \text{Var}[X_k] = \frac{1}{n} [\mathbf{E}[X_k^{2k}] - \mathbf{E}^2[X_k]] \\ &= \frac{1}{n} [\mu_{2k} - \mu_k^2]. \end{aligned}$$

Corolário 2.0.1.

$$\mathbf{E}[\bar{X}] = \mu, \quad \text{Var}[\bar{X}] = \frac{\sigma^2}{n}.$$

Corolário 2.0.2.

$$\mathbf{E}[S^2] = \sigma^2.$$

Prova:

$$\begin{aligned} \mathbf{E}[S^2] &= \frac{n}{n-1} [\mathbf{E}[M_2] - \mathbf{E}[\bar{X}^2]] \\ &= \frac{n}{n-1} [m_2 - (\frac{\sigma^2}{n} + \mu^2)] \\ &= \frac{n}{n-1} [m_2 - \mu^2 - \frac{\sigma^2}{n}] \\ &= \frac{n}{n-1} [\sigma^2 - \frac{\sigma^2}{n}] \\ &= \frac{n}{n-1} [\frac{n-1}{n} \sigma^2] \end{aligned}$$

Teorema 2.3. Se X_1, \dots, X_n são *i.i.d.* $N(\mu, \sigma^2)$.

(a) $\bar{X} \sim N(\mu, \sigma^2/n)$, *i.e.*,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

(b)

$$\sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2(n).$$

(c)

$$\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1).$$

(d)

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

2.0.5 Estatísticas de ordem

Definição 2.3. *Dada uma amostra aleatória X_1, \dots, X_n com função de distribuição comum F . Coloque a amostra em ordem crescente*

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$$

temos

$$X_{(1)} = \min(X_1, \dots, X_n),$$

$$X_{(n)} = \max(X_1, \dots, X_n),$$

$X_{(i)}$ = *i-ésima estatística de ordem.*

Note que $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ são v.a.'s, mas não são independentes. Pois, por exemplo,

$$\mathbf{P}[X_{(1)} \leq y | X_{(n)} \leq y] = 1.$$

Definição 2.4. *Dada uma amostra aleatória X_1, \dots, X_n a mediana amostral é dada por:*

$$M_0 = \begin{cases} X_{(\frac{n+1}{2})}, & \text{se } n \text{ é ímpar} \\ \frac{1}{2}[X_{(\frac{n}{2})} + X_{(\frac{n+2}{2})}], & \text{se } n \text{ é par.} \end{cases}$$

Teorema 2.4.

$$F_{X_{(n)}}(t) = \mathbf{P}(X_{(n)} \leq t) = [F(t)]^n$$

e

$$F_{X_{(1)}}(t) = \mathbf{P}(X_{(1)} \leq t) = 1 - [1 - F(t)]^n.$$

Prova:

$$F_{X_{(n)}}(t) = \mathbf{P}(X_{(n)} \leq t) = \mathbf{P}(X_1 \leq t, \dots, X_n \leq t)$$

$$\text{indep} = \mathbf{P}(X_1 \leq t) \dots \mathbf{P}(X_n \leq t)$$

$$= [F(t)]^n$$

$$F_{X_{(1)}}(t) = \mathbf{P}(X_{(1)} \leq t) = 1 - \mathbf{P}(X_{(1)} > t)$$

$$= 1 - \mathbf{P}(X_1 > t, \dots, X_n > t)$$

$$\text{indep} = 1 - \mathbf{P}(X_1 > t) \dots \mathbf{P}(X_n > t)$$

$$= 1 - [1 - F(t)]^n$$

Se as v.a.'s X_1, \dots, X_n são contínuas e têm densidade f temos

$$f_{X_{(n)}}(t) = n[F(t)]^{n-1}f(t)$$

e

$$f_{X_{(1)}}(t) = n[1 - F(t)]^{n-1}f(t).$$

2.1 Distribuições assintóticas

Um resultado assintótico é o Teorema Central do Limite. Podemos dizer que se X_1, \dots, X_n é uma a.a. com $\mathbf{E}(X_i) = \mu$, e $\text{Var}(X_i) = \sigma^2$. Então

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \rightarrow N(0, 1).$$

Exercício: Utilizando o MINITAB, verifique o Teorema Central do Limite.

Pergunta: Existe distribuição assintótica da mediana?

Resposta: Sim!

$$M_0 \approx N\left(\rho_{0.5}, \frac{1}{4n[f(\rho_{0.5})]^2}\right)$$

onde $\rho_{0.5}$ é a mediana populacional ($F(\rho_{0.5}) = 0.5$), f é a densidade das v.a.'s X_i .

2.2 Estimação Pontual

O problema de estimação paramétrica pode ser definida como:

Assuma que alguma característica dos elementos de uma população pode ser representada por uma v.a. X cuja densidade (ou função de probabilidade) é $f(\cdot, \theta)$ onde a forma da densidade é assumida ser conhecida exceto pelo fato que ela contém um parâmetro desconhecido θ . Nesta situação decidimos tomar uma amostra de tamanho n de X (X_1, \dots, X_n i.i.d. com densidade f) e com base nos valores observados x_1, \dots, x_n deseja-se um bom “chute” do valor θ ou uma função $\tau(\theta)$.

Exemplo 1: É razoável se supor que o número de clientes que vão ao Banespa no horário das 12 às 14hs é uma v.a. Poisson com média (desconhecida) λ . A fim de dimensionar o número de pessoas (caixas) que devem trabalhar nesse horário, observamos o movimento do banco durante 10 dias e com base nessas observações desejamos estimar λ .

Exemplo 2: Na produção de esponjas Scotch-Brite, a fim de fazer o controle de qualidade, a cada 3 horas, 100 esponjas são selecionadas e o número de defeituosas são verificadas para se controlar o valor do parâmetro $p =$ proporção de defeituosas. Sabe-se que

$$X_i \sim b(100, p).$$

A estimação do parâmetro θ pode ser feita de dois modos:

(i) **Estimação Pontual:** Tomamos o valor de alguma estatística $t(X_1, \dots, X_n)$ para representar, ou estimar, $\tau(\theta)$, tal estimativa é chamada estimador pontual;

(ii) **Estimação por intervalo:** Definimos duas estatísticas $t_1(X_1, \dots, X_n)$ e $t_2(X_1, \dots, X_n)$ onde

$$t_1(X_1, \dots, X_n) < t_2(X_1, \dots, X_n)$$

de modo que $[t_1(X_1, \dots, X_n), t_2(X_1, \dots, X_n)]$ constitui um intervalo aleatório para o qual é possível se calcular a probabilidade que este intervalo $\tau(\theta)$. Este intervalo é chamado de intervalo de confiança.

Exemplo: Se queremos estudar o erro médio cometido por uma balança, utilizamos um objeto qualquer e fazemos n medições deste objeto X_1, \dots, X_n . É razoável se supor que $X_i \sim N(\mu, \sigma^2)$ e $\theta = (\mu, \sigma^2)$ é o parâmetro desconhecido, $\tau(\theta) = \mu$ e

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

é um estimador pontual para μ e

$$\left[\bar{X} - 2\sqrt{\frac{S^2}{n}}; \bar{X} + 2\sqrt{\frac{S^2}{n}} \right]$$

é um intervalo de confiança para μ .

Problemas:

- (i) Como encontrar um “bom” estimador?
- (ii) Como selecionar o “melhor” estimador?

2.2.1 Métodos para se encontrar estimadores

Assuma que X_1, \dots, X_n é amostra aleatória de uma distribuição $f(\cdot, \theta)$ (densidade ou função de probabilidade) e $\theta = (\theta_1, \dots, \theta_k)$ é um vetor de números reais (podemos ter $k = 1$). Seja Θ o espaço paramétrico, isto é, o conjunto de valores possíveis que θ pode assumir.

Objetivo: Queremos encontrar estatísticas T_1, \dots, T_k que “aproximem” $\theta_1, \dots, \theta_k$.

Definição 2.5. *Qualquer estatística cujos valores são usados para estimar $\tau(\theta)$ é dita ser um **estimador** de $\tau(\theta)$.*

Exemplo: Sejam X_1, \dots, X_n i.i.d. $N(\mu, \sigma^2)$. Temos como parâmetro $\theta = (\mu, \sigma^2)$, o espaço paramétrico $\Theta = \{(\mu, \sigma^2); \mu \in \mathbf{R}, \sigma^2 > 0\}$. Como estimadores podemos utilizar, \bar{X} para estimar μ e S^2 para estimar σ^2 .

2.2.2 Método dos momentos

Este método é o mais antigo, proposto por Karl Pearson em 1894. Este é um método simples com resultados “razoáveis”.

Seja X uma v.a. com distribuição $f(\cdot, \theta_1, \dots, \theta_k)$. Definimos

$$\mu_r = \mathbf{E}[X^r]$$

o r -ésimo momento de X . Em geral, μ_r é função de $\theta_1, \dots, \theta_k$. Seja X_1, \dots, X_n uma amostra aleatória de $f(\cdot, \theta)$ e denote

$$M_r = \frac{1}{n} \sum_{i=1}^n X_i^r$$

o r -ésimo momento amostral. Sabemos que

$$\mathbf{E}[M_r] = \mu_r,$$

daí é intuitivo se utilizar

$$M_r = \mu_r(\hat{\theta}_1, \dots, \hat{\theta}_k)$$

obtendo-se um sistema de k equações a k incógnitas e temos que a solução $(\hat{\theta}_1, \dots, \hat{\theta}_k)$ é o estimador de momentos de $(\theta_1, \dots, \theta_k)$.

Exemplo 1: Seja X_1, \dots, X_n uma a.a. de uma distribuição $N(\mu, \sigma^2)$. Neste caso,

$$\mu_1 = \mu, \quad \sigma^2 = \mu_2 - \mu_1^2.$$

Daí,

$$M_1 = \hat{\mu}, \quad M_2 = \hat{\sigma}^2 + \hat{\mu}^2$$

e

$$\hat{\mu} = M_1 = \bar{X}$$

e

$$\hat{\sigma} = \sqrt{M_2 - \bar{X}} = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n}}.$$

Exemplo 2: Seja X_1, \dots, X_n uma a.a. Poisson(λ). Queremos estimar λ pelo método de momentos. Como temos somente um parâmetro, temos somente uma equação

$$M_1 = \bar{X} = \hat{\lambda}.$$

Exemplo 3: Seja X_1, \dots, X_n uma a.a. exp(θ). Lembre-se que $\mu_1 = 1/\theta$. Queremos estimar θ pelo método de momentos. Como temos somente um parâmetro, temos somente uma equação

$$M_1 = \bar{X} = 1/\hat{\theta}$$

e conseqüentemente

$$\hat{\theta} = \frac{1}{\bar{X}}.$$

Exemplo 4: Sejam X_1, \dots, X_n i.i.d. $U[a, b]$, o parâmetro de interesse é $(\theta_1, \theta_2) = (a, b)$. Neste caso,

$$\mu_1 = \frac{a+b}{2}, \quad \mu_2 = \frac{a^2 + ab + b^2}{3}.$$

Daí,

$$M_1 = \hat{\mu}_1 = \frac{\hat{a} + \hat{b}}{2}, \quad M_2 = \frac{\hat{a}^2 + \hat{a}\hat{b} + \hat{b}^2}{3}.$$

Exemplo 5: Sejam X_1, \dots, X_n i.i.d. $U[0, \theta]$, o parâmetro de interesse é θ . Neste caso,

$$\mu_1 = \frac{\theta}{2}$$

e

$$\hat{\theta} = 2\bar{X}.$$

Entretanto, se obtemos uma amostra de tamanho 3 com os valores $x_1 = 6$, $x_2 = 50$ e $x_3 = 4$, obteremos $\hat{\theta} = 40$. Este valor não é admissível pois sabemos que $\theta > X_i$ para todo i .

2.2.3 Método de máxima verossimilhança

O método de máxima verossimilhança para gerar estimadores de um parâmetro desconhecido foi introduzido por Sir R.A. Fisher.

Este método produz muito “bons” estimadores. Veremos mais tarde as boas propriedades dos estimadores de máxima verossimilhança.

Considere o seguinte problema: temos duas moedas, uma é honesta e a outra é viciada (tem probabilidade de cara igual a 0.70). O problema é que misturamos as duas moedas e não sabemos diferenciá-las. Para decidir isto, tomamos uma das moedas e jogamos n vezes. Seja:

$X =$ número de caras nas n repetições;

Daí, $X \sim b(n, p)$, isto é:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} = f(k, n)$$

Aqui, $p = 0.5$ ou $p = 0.7$, isto é, $\Theta = \{.5; .7\}$. Se $n = 3$, temos

Valores Possíveis k	0	1	2	3
$f(k;0.5)$	0.125	0.375	0.375	0.125
$f(k;0.7)$	0.027	0.189	0.441	0.343

Note que se tiramos 3 caras em 3 lançamentos da moeda não acreditamos muito que $p = 0.5$, é mais “acreditável” (verossímil) que $p = 0.7$. Por outro lado, se tiramos 0 caras em 3 lançamentos é mais verossímil que $p = 0.5$.

Neste caso,

- Se tiramos 0 ou 1 cara dizemos que $\hat{p} = 0.5$;
- Se tiramos 2 ou 3 caras dizemos que $\hat{p} = 0.7$

Isto é,

$$f(0, 0.7) < f(0, 0.5) \Rightarrow \hat{p} = 0.5$$

$$f(1, 0.7) < f(1, 0.5) \Rightarrow \hat{p} = 0.5$$

$$f(2, 0.7) > f(2, 0.5) \Rightarrow \hat{p} = 0.7$$

$$f(3, 0.7) > f(3, 0.5) \Rightarrow \hat{p} = 0.7$$

Ou seja, escolhemos \hat{p} que faz com que $f(k, \hat{p})$ seja máximo:

$$\hat{p} = \arg \max_{p \in \Theta} f(k, p)$$

Da mesma forma, se $n = 10$ e $\Theta = [0, 1]$ temos

$$f(k, p) = P(X = k) = \binom{10}{p} p^k (1-p)^{10-k}$$

Queremos $\hat{p} = \arg \max_{p \in \Theta} f(k, p)$, para tanto derivamos $f(k, p)$, igualamos a derivada a zero e achamos o ponto crítico:

$$\begin{aligned} \frac{d}{dp} f(k, p) &= \binom{10}{p} k p^{k-1} (1-p)^{10-k} - \binom{10}{p} p^k (n-k) (1-p)^{10-k-1} \\ &= \binom{10}{p} p^{k-1} (1-p)^{10-k-1} [k(1-p) - (n-k)p] \\ &= \binom{10}{p} p^{k-1} (1-p)^{10-k-1} [k - np] \end{aligned}$$

Igualando a zero e resolvendo a equação temos como raízes os pontos 0, 1 e k/n . Se $0 < k < n$ temos que 0 e 1 são pontos de mínimo. Em todos os casos, $\hat{p} = k/n$ é ponto de máximo. Portanto, o estimador de máxima verossimilhança é:

$$\hat{p} = \frac{k}{n}$$

Definição 2.2.1. A função de verossimilhança de n variáveis aleatórias X_1, \dots, X_n é definida ser:

(1) a função de probabilidade conjunta das n variáveis aleatórias, se X_1, \dots, X_n são conjuntamente discretas. Se X_1, \dots, X_n formam uma amostra aleatória de uma variável aleatória discreta X com função de probabilidade $f(\cdot, \theta)$ dependendo de um parâmetro desconhecido θ então se x_1, \dots, x_n são os valores observados, a **função de verossimilhança** da amostra é:

$$L(\theta; x_1, \dots, x_n) = f(x_1, \theta) \dots f(x_n, \theta)$$

(2) a densidade conjunta das n variáveis aleatórias, se X_1, \dots, X_n são conjuntamente contínuas. Se X_1, \dots, X_n formam uma amostra aleatória de

uma variável aleatória contínua X com densidade $f(\cdot, \theta)$ dependendo de um parâmetro desconhecido θ então se x_1, \dots, x_n são os valores observados, a **função de verossimilhança da amostra** é:

$$L(\theta; x_1, \dots, x_n) = f(x_1, \theta) \dots f(x_n, \theta)$$

Definição 2.2.2. Seja $L(\theta) = L(\theta; x_1, \dots, x_n)$ a função de verossimilhança para as v.a.'s X_1, \dots, X_n . Se $\hat{\theta} [= \hat{\theta}(x_1, \dots, x_n)]$ é uma função das observações é o valor de θ no espaço paramétrico Θ que maximiza $L(\theta)$, então $\hat{\theta} = \arg \max_{\theta \in \Theta} L(\theta)$ é a **estimativa de máxima verossimilhança** de θ e $\hat{\Theta} = \hat{\theta}(X_1, \dots, X_n)$ é o **estimador de máxima verossimilhança** de θ .

Antes de olhar alguns exemplos, vamos lembrar um teorema de cálculo que é muito útil em encontrar máximos de funções. Geralmente, como $L(\theta)$ é um produto de funções de probabilidade ou densidades, é sempre positiva. Assim, $l(\theta) = \log(L(\theta))$ sempre pode ser definida e o valor de θ que maximiza $L(\theta)$ também maximiza $l(\theta)$.

Exemplo 1: Suponha que retiramos uma amostra aleatória de tamanho n de uma distribuição de Bernoulli

$$f(x, p) = p^x (1 - p)^{1-x} I_{\{0,1\}}(x), 0 \leq p \leq 1$$

Os valores amostrais x_1, \dots, x_n serão uma sequência de 0's e 1's e a função de verossimilhança é:

$$L(p) = \prod_{i=1}^n p^{x_i} (1 - p)^{1-x_i} I_{\{0,1\}}(x_i) = p^{\sum x_i} (1 - p)^{n - \sum x_i}$$

Podemos definir,

$$l(p) = \sum x_i \log(p) + (n - \sum x_i) \log(1 - p)$$

Como l é uma função contínua de p , se existir um valor \hat{p} tal que

$$\frac{d}{dp}l(\hat{p}) = 0, \frac{d^2}{dp^2}l(\hat{p}) < 0$$

então este valor maximiza a função l :

$$\frac{d}{dp}l(p) = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1 - p}$$

Assim,

$$\frac{\sum x_i}{\hat{p}} - \frac{n - \sum x_i}{1 - \hat{p}} = 0$$

Temos,

$$\hat{p} = \frac{\sum x_i}{n}$$

Como,

$$\frac{d^2}{dp^2} = -\frac{\sum x_i}{p^2} - \frac{n - \sum x_i}{(1 - p)^2} < 0$$

para todos os valores de p temos que \hat{p} corresponde a um ponto de máximo.

Portanto, o estimador de máxima verossimilhança de θ é:

$$\hat{P} = \frac{\sum X_i}{n}$$

Exemplo 2: Suponha que retiramos uma amostra aleatória de tamanho n de uma distribuição normal com média μ e variância 1. Se X_1, \dots, X_n é a amostra aleatória, a função de verossimilhança da amostra é:

$$\begin{aligned} L(\mu) &= \prod_{i=1}^n f(x_i, \mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-(x_i - \mu)^2/2} \\ &= (2\pi)^{-n/2} e^{-\sum (x_i - \mu)^2/2} \end{aligned}$$

cujo logaritmo é:

$$l(\mu) = -\frac{n}{2} \log(2\pi) - \sum \frac{(x_i - \mu)^2}{2}$$

e

$$\begin{aligned} \frac{d}{d\mu} l(\mu) &= \sum (x_i - \mu) = \sum x_i - n\mu \\ \frac{d^2}{d\mu^2} l(\mu) &= -n < 0 \end{aligned}$$

Assim,

$$\hat{\mu} = \frac{1}{n} \sum x_i = \bar{x}$$

é a estimativa de máxima verossimilhança de θ e o estimador de máxima verossimilhança é:

$$\hat{\mu} = \frac{\sum X_i}{n} = \bar{X}$$

Se a função de verossimilhança contém k parâmetros, isto é, se:

$$L(\theta_1, \dots, \theta_k) = \prod_{i=1}^n f(x_i; \theta_1, \dots, \theta_k)$$

então os estimadores de máxima verossimilhança são as estatísticas

$\hat{\theta}_1(X_1, \dots, X_n), \dots, \hat{\theta}_k(X_1, \dots, X_n)$ onde $\hat{\theta}_1, \dots, \hat{\theta}_k$ são os valores em Θ que maximizam $L(\theta_1, \dots, \theta_k)$.

Se certas condições de regularidade são satisfeitas, o ponto onde a função de verossimilhança é máxima é a solução das k equações:

$$\frac{\partial}{\partial \theta_1} L(\theta_1, \dots, \theta_k) = 0, \dots, \frac{\partial}{\partial \theta_k} L(\theta_1, \dots, \theta_k) = 0$$

ou equivalentemente,

$$\frac{\partial}{\partial \theta_1} l(\theta_1, \dots, \theta_k) = 0, \dots, \frac{\partial}{\partial \theta_k} l(\theta_1, \dots, \theta_k) = 0.$$

Exemplo 3: Uma amostra aleatória de tamanho n da distribuição normal de média μ e desvio padrão σ tem densidade:

$$f(x_1, \dots, x_n, \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i - \mu)^2}$$

and

$$L(\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{\frac{1}{2\sigma^2} \sum (x_i - \mu)^2\right\}$$

seu logaritmo sendo:

$$l(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 + \frac{1}{2\sigma^2} \sum (x_i - \mu)^2$$

onde $\Theta = \{(\mu, \sigma^2); -\infty < \mu < \infty, \sigma^2 > 0\}$. Portanto,

$$\frac{\partial}{\partial \mu} l(\mu, \sigma^2) = \frac{1}{\sigma^2} \sum (x_i - \mu)$$

$$\frac{\partial}{\partial \sigma^2} l(\mu, \sigma^2) = -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{\sigma^4} \sum (x_i - \mu)^2$$

Daí,

$$\begin{aligned} \frac{1}{\hat{\sigma}^2} \sum (x_i - \hat{\mu}) = 0 &\Rightarrow \sum (x_i - \hat{\mu}) = 0 \Rightarrow \hat{\mu} = \frac{\sum x_i}{n} \\ -\frac{n}{2} \frac{1}{\hat{\sigma}^2} + \frac{1}{\hat{\sigma}^4} \sum (x_i - \hat{\mu})^2 = 0 &\Rightarrow \hat{\sigma}^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n} \end{aligned}$$

e os estimadores de máxima verossimilhança são:

$$\hat{\mu} = \frac{\sum X_i}{n}$$

$$\hat{\sigma}^2 = \sum_{i=1}^n \frac{(X_i - \bar{x})^2}{n}$$

Exemplo 4: Seja a variável aleatória tendo densidade uniforme dada por:

$$f(x, \theta) = I_{[\theta-0.5; \theta+0.5]}(x)$$

onde $\Theta = (-\infty, \infty)$. A função de verossimilhança para uma amostra aleatória de tamanho n é dada por:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n f(x_i, \theta) \\ &= \prod_{i=1}^n I_{[\theta-0.5; \theta+0.5]}(x_i) \\ &= I_{[x_{(n)}-0.5; x_{(1)}+0.5]}(\theta) \end{aligned}$$

onde $x_{(1)} = \min\{x_1, \dots, x_n\}$ e $x_{(n)} = \max\{x_1, \dots, x_n\}$ e temos a última igualdade pois

$$\begin{aligned} \prod_{i=1}^n I_{[\theta-0.5; \theta+0.5]}(x_i) = 1 &\Leftrightarrow x_i \in [\theta - 0.5; \theta + 0.5], \text{ for all } i = 1, \dots, n \\ &\Leftrightarrow \theta - 0.5 \leq x_{(1)} \text{ e } \theta + 0.5 \geq x_{(n)} \\ &\Leftrightarrow \theta \leq x_{(1)} + 0.5 \text{ e } \theta \geq x_{(n)} - 0.5 \end{aligned}$$

Daí,

$$L(\theta) = \begin{cases} 1, & \text{se } x_{(n)} - 0.5 \leq \theta \leq x_{(1)} + 0.5 \\ 0, & \text{caso contrário} \end{cases}$$

Assim, qualquer estatística com valor $\hat{\theta}$ satisfazendo $X_{(n)} - 0.5 \leq \hat{\theta} \leq X_{(1)} + 0.5$ é estimador de máxima verossimilhança de θ . Por exemplo, $X_{(n)} - 0.5$, $X_{(1)} + 0.5$ ou $(X_{(1)} + X_{(n)})/2$, etc...

Exemplo 5: Seja X uma variável aleatória com densidade uniforme no intervalo $[0, \theta]$, achar o EMV de θ .

$$f(x, \theta) = \frac{1}{\theta} I_{[0; \theta]}(x)$$

onde $\Theta = (0, \infty)$. A função de verossimilhança para uma amostra aleatória de tamanho n é dada por:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n f(x_i, \theta) \\ &= \prod_{i=1}^n \frac{1}{\theta} I_{[0; \theta]}(x_i) \\ &= \theta^{-n} I_{[0; \theta]}(x_{(n)}) \\ &= \theta^{-n} I_{[x_{(n)}; \infty]}(\theta) \end{aligned}$$

onde $x_{(n)} = \max\{x_1, \dots, x_n\}$. Daí,

$$L(\theta) = \begin{cases} \theta^{-n}, & \text{se } x_{(n)} \leq \theta \\ 0, & \text{caso contrário} \end{cases}$$

Assim, o valor de θ que maximiza $L(\theta)$ é $\hat{\theta} = x_{(n)}$ e portanto o EMV de θ é $X_{(n)}$.

Teorema 2.2.3. *Seja $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ o estimador de máxima verossimilhança de θ . Se $\tau(\theta) = (\tau_1(\theta), \dots, \tau_r(\theta))$, $1 \leq r \leq k$, é uma transformação no espaço paramétrico Θ , então o estimador de máxima verossimilhança de $\tau(\theta)$ é: $\tau(\hat{\theta}) = (\tau_1(\hat{\theta}), \dots, \tau_r(\hat{\theta}))$.*

Exemplo: Na densidade normal, seja $\theta = (\mu, \sigma^2)$. Suponha $\tau(\theta) = \mu + z_q \sigma$ onde z_q é tal que $\Phi(z_q) = q$, portanto $\tau(\theta)$ é o q -ésimo quartil. Portanto, o estimador de máxima verossimilhança de $\tau(\theta)$ é:

$$\bar{X} + z_q \sqrt{\frac{1}{n} \sum (X_i - \bar{X})^2}$$

Exercícios:

- (1) Suponha que X é uma variável normal com média 10 e variância σ^2 desconhecida. Qual o EMV para σ^2 , baseado em uma amostra aleatória de n observações de X ?
- (2) Suponha $X \sim P(\lambda)$. Dada uma amostra aleatória de tamanho n de X , qual o EMV de λ ?
- (3) Se $X \sim \text{geom}(p)$, qual o EMV de p baseado em uma amostra de tamanho n ?
- (4) Se $X \sim \text{exp}(\lambda)$, qual o EMV de λ baseado em uma amostra aleatória de n observações?

2.2.4 Outros métodos

Há vários outros métodos de se obter estimadores pontuais baseados em propriedades que desejamos obter: mínimos quadrados, método de Bayes, método quiquadrado, etc...

2.3 Propriedades de estimadores pontuais

Já vimos dois métodos de construção de estimadores pontuais para parâmetros desconhecidos. Em muitos casos os dois métodos obtém o mesmo estimador, mas em muitos casos importantes não. Também há outros métodos ainda não estudados para a obtenção de estimadores. As questões que nos vêm a mente agora são: “Qual estimador devo utilizar?”, “Como selecionar o mel-

hor estimador?”, “Quais as propriedades que um bom estimador deve ter?”. Se pudessemos encontrar uma escala de “bondade” de estimadores, sempre poderíamos escolher o melhor estimador para cada caso. Entretanto, não há uma escala universal de “bondade”.

O estimador ($\hat{\Gamma}$) de um parâmetro desconhecido (γ) é uma estatística e como tal uma v.a. que tem uma lei de probabilidade, portanto é sujeita a variabilidade, não é razoável de se esperar que a estimativa $\hat{\gamma}$ seja igual ao valor verdadeiro do parâmetro γ para todas as amostras retiradas. Se consideramos dois estimadores $\hat{\Gamma}$ e $\tilde{\Gamma}$ para o mesmo parâmetro γ , podemos derivar as leis de probabilidade dos estimadores e compará-las de algum modo. Por exemplo, se $\hat{\Gamma} \sim U(\gamma - 0.5; \gamma + 0.5)$ e $\tilde{\Gamma} \sim U(\gamma - 0.01; \gamma + 0.01)$, certamente preferiríamos $\tilde{\Gamma}$ como estimador de γ . Que pena que em geral esta comparação não seja tão direta.

Intuitivamente, queremos um estimador que seja “próximo” do verdadeiro valor do parâmetro. Há várias maneiras de se definir “próximo”. O estimador $\hat{\Gamma} = \hat{\Gamma}(X_1, \dots, X_n)$ é uma v.a. e portanto tem uma distribuição de probabilidade. A distribuição de $\hat{\Gamma}$ nos diz como os valores observados (estimativas) $\hat{\gamma}$ estão distribuídos e gostaríamos de ter valores de $\hat{\Gamma}$ distribuídos próximos de γ . Sabemos que a média e a variância de uma distribuição são medidas de locação e dispersão, daí o sentido de $\hat{\Gamma}$ ser “próximo” de γ poderia ser:

- $\mathbf{E}(\hat{\Gamma})$ “próxima” de γ ;

- $\text{Var}(\hat{\Gamma})$ “próxima” de 0.

Uma propriedade desejável para um estimador é que sua média seja o valor verdadeiro do parâmetro.

Definição 2.3.1. *Um estimador $\hat{\Gamma}$ de um parâmetro γ é não viciado se $\mathbf{E}(\hat{\Gamma}) = \gamma$.*

Exemplo: Se X_1, \dots, X_n forma uma amostra aleatória de uma distribuição tal que $\mathbf{E}(X_i) = \mu$ então sabemos que $\mathbf{E}(\bar{X}) = \mu$. Portanto, \bar{X} é um estimador não viciado de μ se $X_i \sim N(\mu, \sigma^2)$, de p se $X_i \sim b(1, p)$, de λ se $X_i \sim \text{Poisson}(\lambda)$.

A propriedade de ser não viciado, embora desejável para um estimador, não deve ser o único critério utilizado para se comparar estimadores, também devemos ter estimadores mais “concentrados” em torno do verdadeiro valor do parâmetro. Para isto eles devem ter variância pequena.

Definição 2.3.2. *Se $\hat{\Gamma}$ e $\tilde{\Gamma}$ são dois estimadores não viciados de γ , dizemos que $\hat{\Gamma}$ é mais eficiente que $\tilde{\Gamma}$ se*

$$\text{Var}(\hat{\Gamma}) < \text{Var}(\tilde{\Gamma}).$$

Exemplo: Suponha que X_1, \dots, X_n é uma amostra aleatória de uma distribuição $\text{Poisson}(\lambda)$. Portanto, $\hat{\Lambda} = \bar{X}$ e $\tilde{\Lambda} = (X_1 + X_2)/2$ são ambos estimadores não viciados de λ , entretanto,

$$\text{Var}(\hat{\Lambda}) = \frac{\lambda}{n}, \quad \text{Var}(\tilde{\Lambda}) = \frac{\lambda}{2}$$

Assim, se $n > 2$, $\hat{\Lambda}$ é mais eficiente que $\tilde{\Lambda}$.

2.3.1 Erro Quadrático Médio

Nem sempre um estimador viciado não é bom, às vezes o que perdemos por ter um viés pequeno pode ser compensado pela concentração em torno do valor verdadeiro. De alguma forma temos que combinar os dois fatores:

- $\mathbf{E}(\hat{\Gamma})$ “próxima” de γ ;
- $\text{Var}(\hat{\Gamma})$ “próxima” de 0.

Isto pode ser obtido através de uma medida muito útil de proximidade chamada *erro quadrático médio* (EQM).

Definição 2.3.3. *Seja $\hat{\Gamma} = \hat{\Gamma}(X_1, \dots, X_n)$ um estimador de γ baseado em uma amostra aleatória X_1, \dots, X_n . O erro quadrático médio (EQM) de $\hat{\Gamma}$ é:*

$$EQM(\hat{\Gamma}, \gamma) = \mathbf{E}_\gamma[(\hat{\Gamma} - \gamma)^2].$$

Obs.: Para v.a.'s contínuas com densidade $f(\cdot, \gamma)$,

$$\mathbf{E}_\gamma[(\hat{\Gamma} - \gamma)^2] = \int \dots \int [(\hat{\gamma}(x_1, \dots, x_n) - \gamma)^2] f(x_1, \gamma) \dots f(x_n, \gamma) dx_1 \dots dx_n.$$

Se $\hat{\Gamma}$ é não viciado, então $EQM(\hat{\Gamma}, \gamma) = \text{Var}_\gamma(\hat{\Gamma})$. Se $\hat{\Gamma}$ é viciado, então $EQM(\hat{\Gamma}, \gamma)$ pode ser pensado como uma medida de espalhamento de $\hat{\Gamma}$ em

torno de γ .

Se formos comparar estimadores baseados em seus EQM, naturalmente iremos preferir aquele com menor EQM. Geralmente, EQM depende de γ (parâmetro desconhecido) e não temos um estimador com EQM uniformemente menor. A situação abaixo é a mais comum:

- Se $\gamma \in [a, b]$ dizemos que Γ_1 é melhor que Γ_2 ;
- Se $\gamma \notin [a, b]$ dizemos que Γ_2 é melhor que Γ_1 .

Não temos base para escolher um estimador em detrimento do outro.

Exemplo: Sejam X_1, X_2, \dots, X_n i.i.d. $\exp(\beta)$ e tome

$$T_1 = \left(\sum_{i=1}^n X_i \right) / n$$

e

$$T_2 = \sum_{i=1}^n a_i X_i$$

onde $\sum_{i=1}^n a_i = 1$. Portanto, T_1 e T_2 são estimadores de $\tau(\beta) = 1/\beta$.

Calcule $\text{EQM}(T_1, \beta)$ e $\text{EQM}(T_2, \beta)$, veja se preferimos T_1 ou T_2 com base neste critério.

Como T_1 e T_2 são não viciados temos que

$$\begin{aligned} \text{EQM}(T_1, \beta) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n X_i\right) \\ &= \frac{1}{n} \text{Var}(X_1) = \frac{1}{n\beta^2} \end{aligned}$$

e

$$\begin{aligned} \text{EQM}(T_2, \beta) &= \text{Var}\left(\sum_{i=1}^n a_i X_i\right) \\ &= \sum_{i=1}^n a_i^2 \text{Var}(X_i) \\ &= \frac{1}{\beta^2} \sum_{i=1}^n a_i^2 \end{aligned}$$

Daí, $(EQM)(T_1, \beta) \leq (EQM)(T_2, \beta)$ se $(1/n) \leq \sum_{i=1}^n a_i^2$. Mas $\min \sum_{i=1}^n a_i^2$ sujeito a $\sum_{i=1}^n a_i = 1$ é quando $a_i = 1/n$ para todo $i = 1, \dots, n$. Portanto, T_1 é sempre melhor que T_2 .

Exemplo: Sejam X_1, X_2, \dots, X_n i.i.d. $\text{Poisson}(\lambda)$. Sejam $T_1 = 1$ e $T_2 = \bar{X}$ dois estimadores de λ . Daí,

$$\text{EQM}(T_1, \lambda) = \mathbf{E}_\lambda(1 - \lambda)^2 = (1 - \lambda)^2$$

$$\text{EQM}(T_2, \lambda) = \mathbf{E}_\lambda(\bar{X} - \lambda)^2 = \text{Var}(\bar{X}) = \lambda/n$$

Somente para exemplificar, vamos supor que $n = 2$. Pela Figura 2.3.1 podemos ver que

- Se $\lambda \in [1/2; 2]$ temos T_1 preferível a T_2 ;
- Se $\lambda \notin [1/2; 2]$ temos T_2 preferível a T_1 .

Mas, em $\lambda = 1$, $\text{EQM}(T_1, 1) = 0 < \text{EQM}(T, 1)$ para qualquer estimador T de λ . assim, quando $\lambda = 1$ o estimador $T_1 = 1$ será preferível a qualquer

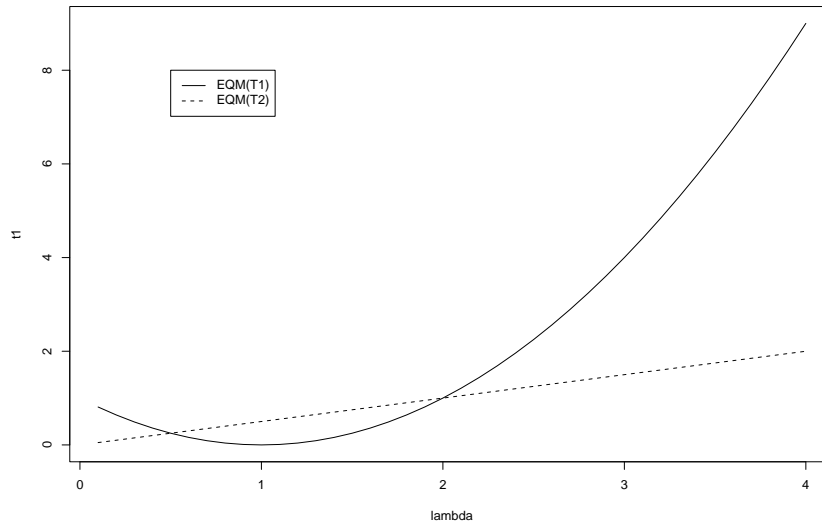


Figura 2.3.1: Erro médio quadrático para os estimadores T_1 e T_2 em termos de λ estimador.

Assim vemos que não existe um estimador $\hat{\Gamma}$ de γ que possa ser o melhor de todos considerando-se o critério de EQM.

Multiplicadores de Lagrange: Pode-se mostrar que o mínimo da função $g(\mathbf{x})$ sujeito a $h(\mathbf{x}) = K$ é encontrado achando-se o mínimo da função $g(\mathbf{x}) - \lambda h(\mathbf{x})$. Não vamos provar isto aqui, mas é possível se provar que se \mathbf{y} satisfaz $h\mathbf{y} = K$ e minimiza $g(\mathbf{x}) - \lambda h(\mathbf{x})$ para algum λ , então para qualquer outro \mathbf{x} tal que $h(\mathbf{x}) = K$,

$$g(\mathbf{x}) - \lambda h(\mathbf{x}) \geq g(\mathbf{y}) - \lambda h(\mathbf{y}),$$

ou, como $h(\mathbf{x}) = h(\mathbf{y})$,

$$g(\mathbf{x}) \geq g(\mathbf{y}).$$

Assim, \mathbf{y} é o ponto de mínimo. No caso que queremos minimizar $\sum a_i^2$ sujeito à $\sum a_i = 1$, minimizamos a função $f(\mathbf{a}) = \sum a_i^2 - \lambda \sum a_i = 1$. A derivada desta função com respeito a cada a_i deve ser zero:

$$2a_i - \lambda = 0, \quad j = 1, \dots, n$$

Portanto, os valores de a_i que minimizam $f(\mathbf{a})$ são todos iguais e como eles devem somar 1 devem ser todos iguais a $1/n$. Portanto, a média amostral é o estimador linear não viciado mais eficiente (de mínima variância).

O problema de encontrar um estimador que tenha uniformemente o menor EQM não tem solução. Já vimos que o “pior” estimador possível, tem um EQM de zero para um valor particular do parâmetro. Isto ocorre porque estamos procurando estimadores numa classe muito ampla. Algumas vezes pode-se encontrar estimadores com mínima variância na classe dos estimadores não viciados (veja ENVUMV); mas exceto pelo fato de que nesta classe o problema de minimalidade de EQM tem solução, a restrição a estimadores não viciados algumas vezes excluem estimadores que são bons.

Exemplo: Já vimos que se temos uma amostra aleatória de uma distribuição $N(\mu, \sigma^2)$, o estimador de máxima verossimilhança de σ é $\hat{\sigma}^2 = (1/n) \sum (X_i - \bar{X})^2$ e

$$\mathbf{E}[\hat{\sigma}^2] = \sigma^2 \left(1 - \frac{1}{n}\right)$$

portanto, $\hat{\sigma}^2$ tem um pequeno viés. Seu erro quadrático médio é:

$$\begin{aligned} EQM(\hat{\sigma}^2; \mu, \sigma^2) &= \text{Var}(\hat{\sigma}^2) + (\mathbf{E}(\hat{\sigma}^2) - \sigma^2)^2 \\ &= \frac{2\sigma^4(n-1)}{n^2} + \left(-\frac{\sigma^2}{n}\right)^2 \\ &= \frac{2n-1}{n^2}\sigma^4 \end{aligned}$$

Um estimador não viciado de σ é $S^2 = (1/(n-1)) \sum (X_i - \bar{X})^2$ (a variância amostral) e seu EQM é:

$$\begin{aligned} EQM(S^2; \mu, \sigma^2) &= \text{Var}(S^2) \\ &= \frac{n^2}{(n-1)^2} \frac{2(n-1)}{n^2} \sigma^4 \\ &= \frac{2\sigma^4}{n-1} < \frac{2n-1}{n^2} \sigma^4 \end{aligned}$$

Portanto, neste caso, o estimador não viciado tem um EQM maior que um estimador “um pouco” viciado.

Apesar de sua dependência nos parâmetros desconhecidos o EQM é útil quando estamos estudando a “performance” dos estimadores para grandes amostras. Neste caso, estamos procurando estimadores cujos EQM's sejam próximos a zero quando o tamanho cresce.

2.3.2 Consistência

Um estimador, em geral, depende do tamanho da amostra. Por exemplo, os momentos amostrais dependem de n e são definidos para todos os tamanhos amostrais, e.g., $\bar{X}_n = (1/n) \sum_{i=1}^n X_i$. Assim temos uma sequência de estimadores $\hat{\Gamma}_n$ que dependem do tamanho da amostra. É intuitivo se desejar

que quanto maior a amostra melhor seja o nosso estimador; assim um bom estimador $\hat{\Gamma}_n$ tem EQM que decresce a 0 quanto mais elementos contiver a amostra:

$$\lim_{n \rightarrow \infty} EQM(\hat{\Gamma}_n, \gamma) = \lim_{n \rightarrow \infty} \mathbf{E}(\hat{\Gamma}_n - \gamma)^2 = 0. \quad (2.3.4)$$

Definição 2.3.5. *Se condição (2.3.4) ocorre dizemos que a sequência de estimadores $\{\hat{\Gamma}_n\}$ é consistente em média quadrática.*

Note que condição (2.3.4) é verdadeira se, e somente se, o viés do estimador e a variância do estimador tende a 0 quando $n \rightarrow \infty$.

Definição 2.3.6. *Uma sequência de estimadores $\{\hat{\Gamma}_n\}$ é dita ser consistente se:*

$$\lim_{n \rightarrow \infty} \mathbf{P}(|\hat{\Gamma}_n - \gamma| \geq \epsilon) = 0, \text{ para todo } \epsilon > 0 \quad (2.3.7)$$

Obs.: Condição (2.3.4) implica em condição (2.3.7). Isto é, um estimador consistente em média quadrática é consistente.

Exemplo: Os momentos amostrais $M_{n,k} = (1/n) \sum_{i=1}^n X_i^k$ são consistentes em média quadrática dos correspondentes momentos populacionais μ_k pois satisfazem condição (2.3.4) :

$$\begin{aligned} \mathbf{E}(M_{n,k}) &= \mathbf{E}\left(\frac{1}{n} \sum_{i=1}^n X_i^k\right) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbf{E}(X_i^k) = \mu_k \end{aligned}$$

portanto, o vício é zero. Mais ainda,

$$\begin{aligned}
\text{Var}(M_{n,k}) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i^k\right) \\
&= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i^k) \\
&= \frac{1}{n} \text{Var}(X_1) \rightarrow 0
\end{aligned}$$

quando $n \rightarrow \infty$. Em particular, \bar{X} é um estimador consistente de μ e $\hat{\sigma}^2$ é estimador consistente de σ . A variância amostral também é um estimador consistente de σ^2 (por que?).

2.3.3 Normalidade Assintótica

Novamente vamos considerar uma sequência de estimadores $\hat{\Gamma}_n$ do parâmetro desconhecido γ .

Definição 2.3.8. *Uma sequência de estimadores $\hat{\Gamma}_n$ de γ é definida como sendo a **melhor sequência assintoticamente normal** (best asymptotically normal, BAN) se, e somente se, as 3 condições abaixo são satisfeitas:*

- (i) $\sqrt{n}(\hat{\Gamma}_n - \gamma) \approx N(0, \sigma^2(\gamma))$, quando $n \rightarrow \infty$;
- (ii) Para todo $\epsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbf{P}_\gamma[|\hat{\Gamma}_n - \gamma| > \epsilon] = 0$$

para todo γ . ($\hat{\Gamma}_n$ é fracamente consistente).

- (iii) Seja S_n uma outra sequência de estimadores fracamente consistentes de γ tal que

$$\sqrt{n}(S_n - \gamma) \approx N(0, \tilde{\sigma}^2(\gamma))$$

quando $n \rightarrow \infty$, Então $\sigma^2(\gamma) < \tilde{\sigma}^2(\gamma)$, para todo γ .

A utilidade desta definição se deriva parcialmente dos teoremas que garantem a existência de estimadores BAN e do fato que estimadores razoáveis e comuns são assintoticamente normalmente distribuídos.

Exemplos:

(1) $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ é BAN para μ . De fato,

$$\mathbf{P}[|\bar{X}_n - \mu| > \epsilon] \leq \frac{\text{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \rightarrow 0, \text{ quando } n \rightarrow \infty$$

e

$$\sqrt{n}(\bar{X}_n - \mu) \approx N(0, \sigma^2), \text{ quando } n \rightarrow \infty$$

e nenhum outro estimador com essas propriedades possui variância assintótica menor que σ^2 . Mas há muitos outros estimadores S_n que também são BAN, e.g.

$$S_n = \frac{1}{n+1} \sum_{i=1}^n X_i$$

também é BAN para μ .