INTRODUÇÃO AO APRENDIZADO ESTATÍSTICO

Mariane Branco Alves

DME-IM-UFRJ

mariane@im.ufrj.br

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

INSTITUTO DE MATEMÁTICA
Universidade Federal do Rio de Janeiro

lse laboratório de sistemas estocásticos

INTRODUÇÃO AO APRENDIZADO ESTATÍSTICO

Mariane Branco Alves

DME-IM-UFRJ

mariane@im.ufrj.br

Universidade Federal do Rio de Janeiro

Instituto de Matemática
Universidade Federal do Rio de Janeiro

Bibliografia Recomendada

1. Morris DeGroot e Mark J. Schervish (2002). *Probability and Statistics.* Addison Wesley.
2. Bolfarine, H. e Sandoval, C. (2001). *Introdução à Inferência Estatística.* IMPA. Coleção Matemática Aplicada
3. Casella, G., Berger, R. L (2001). *Statistical Inference.* Duxbury, Thomson Learning.
4. Migon, H. S., Gamerman, D. e Louzada, F. (2015). *Statistical Inference - An Integrated Approach.* Taylor Francis Group, LLC.
5. Murphy, K. P. (2012). *Machine Learning - A Probabilistic Perspective.* Massachusetts Institute of Technology

Critério de Avaliação

A avaliação da disciplina será baseada na soma dos pontos obtidos em exercícios de caráter aplicado, atribuídos ao longo do curso, a cada duas aulas.

Os exercícios serão desenvolvidos em horário extraclasse.

Conteúdo

Afinal, do que trata o Aprendizado Estatístico?

Afinal, do que trata o Aprendizado Estatístico?

Dados
disponíveis?

Afinal, do que trata o Aprendizado Estatístico?

Dados disponíveis?

Precisa prever/ tomar decisão?

Afinal, do que trata o Aprendizado Estatístico?

Dados disponíveis?

Precisa prever/ tomar decisão?

Alvo da decisão é universo maior?

Afinal, do que trata o Aprendizado Estatístico?

Dados disponíveis?

Ambiente sujeito a incerteza?

Precisa prever/ tomar decisão?

Alvo da decisão é universo maior?

Afinal, do que trata o Aprendizado Estatístico?

Aprendizado Estatístico!

Dados disponíveis?

Ambiente sujeito a incerteza?

Precisa prever/ tomar decisão?

Alvo da decisão é universo maior?

Universidade Federal do Rio de Janeiro

Instituto de Matemática
Universidade Federal do Rio de Janeiro

- Na disciplina de Modelagem de Incertezas, diversos modelos probabilísticos foram apresentados, de forma a quantificar incertezas associadas a uma variável aleatória Y.

Aprendizado Estatístico / Inferência Estatística

- Na disciplina de Modelagem de Incertezas, diversos modelos probabilísticos foram apresentados, de forma a quantificar incertezas associadas a uma variável aleatória Y.
- Exemplos são Normal (μ, σ^2), Poisson(λ), Binomial(n, π) etc.

- Na disciplina de Modelagem de Incertezas, diversos modelos probabilísticos foram apresentados, de forma a quantificar incertezas associadas a uma variável aleatória Y.
- Exemplos são Normal (μ, σ^2), Poisson(λ), Binomial(n, π) etc.
- Todos esses modelos são governados por parâmetros θ: $\theta = (\mu, \sigma^2)$; $\theta = \lambda$; $\theta = \pi$.

- Na disciplina de Modelagem de Incertezas, diversos modelos probabilísticos foram apresentados, de forma a quantificar incertezas associadas a uma variável aleatória Y.
- Exemplos são Normal (μ, σ^2), Poisson(λ), Binomial(n, π) etc.
- Todos esses modelos são governados por parâmetros θ: $\theta = (\mu, \sigma^2)$; $\theta = \lambda$; $\theta = \pi$.
- Dados valores para θ, tais modelos podem ser usados para calcular probabilidades associadas a Y e medidas sumário como valor esperado, quantis e variância de Y, entre outras.

Aprendizado Estatístico / Inferência Estatística

- Na disciplina de Modelagem de Incertezas, diversos modelos probabilísticos foram apresentados, de forma a quantificar incertezas associadas a uma variável aleatória Y.
- Exemplos são Normal (μ, σ^2), Poisson(λ), Binomial(n, π) etc.
- Todos esses modelos são governados por parâmetros θ: $\theta = (\mu, \sigma^2)$; $\theta = \lambda$; $\theta = \pi$.
- Dados valores para θ, tais modelos podem ser usados para calcular probabilidades associadas a Y e medidas sumário como valor esperado, quantis e variância de Y, entre outras.
- Questão: Na prática, os parâmetros θ são desconhecidos e não observáveis. Desejamos aprender sobre θ a partir de: i) dados coletados e, eventualmente, ii) informação subjetiva disponível sobre θ e externa aos dados.

Desejamos extrair informação/aprendizado a partir de dados, levando em conta que:

- a própria coleta dos dados está sujeita a variabilidade/incerteza;

Desejamos extrair informação/aprendizado a partir de dados, levando em conta que:

- a própria coleta dos dados está sujeita a variabilidade/incerteza;
- a extrapolação da informação observada nos dados para universos mais amplos está sujeita a incertezas;

Desejamos extrair informação/aprendizado a partir de dados, levando em conta que:

- a própria coleta dos dados está sujeita a variabilidade/incerteza;
- a extrapolação da informação observada nos dados para universos mais amplos está sujeita a incertezas;
- pode haver informação relevante externa aos dados (subjetiva ou vinda de outros conjuntos de dados).

Como procedemos?

Y é uma característica de interesse

Inviável observar Y
na população toda

$Y \sim f(y|\theta)$
Foco:θ

Como procedemos?

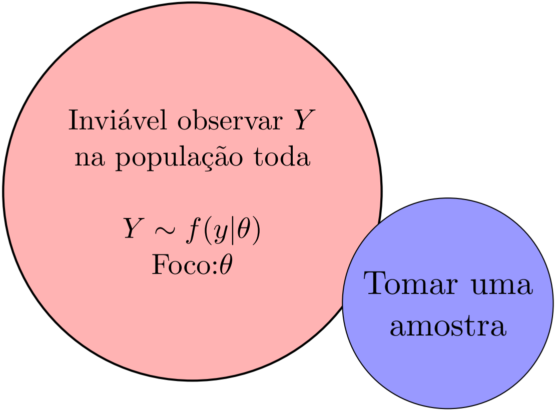
Y é uma característica de interesse

Inviável observar Y
na população toda

$Y \sim f(y|\theta)$
Foco: θ

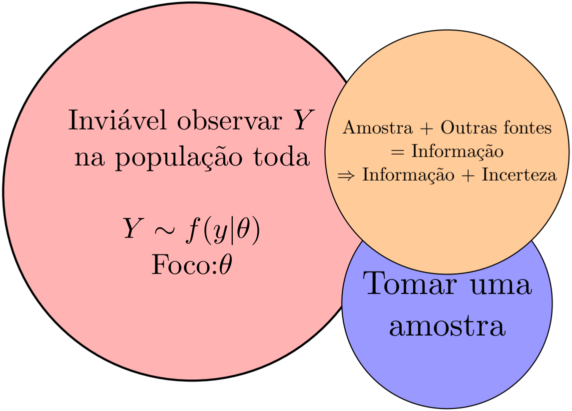
Como procedemos?

Y é uma característica de interesse



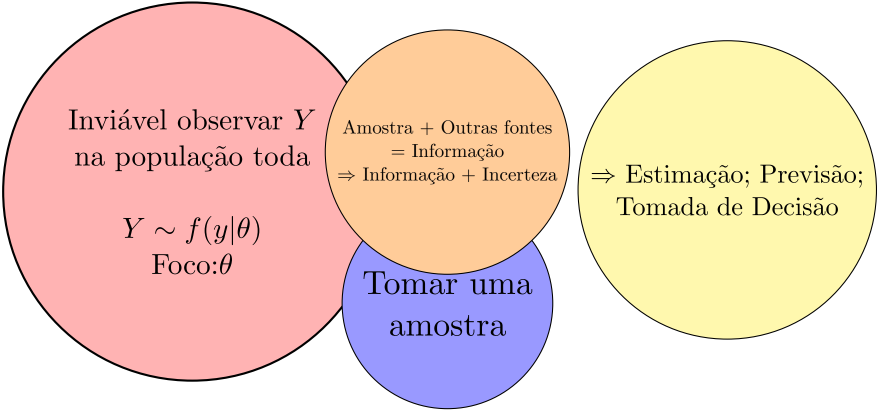
Como procedemos?

Y é uma característica de interesse



Como procedemos?

Y é uma característica de interesse



Universidade Federal do Rio de Janeiro

Instituto de Matemática
Universidade Federal do Rio de Janeiro

Elementos Fundamentais

- Inferência Estatística fornece métodos para obtenção de conclusões e tomadas de decisão em cenários sujeitos a incertezas.

- É processo indutivo: informação obtida com base na observação de amostra aleatória (e eventualmente outras fontes de informação) recebe tratamento probabilístico: modelagem de incertezas ao induzir conclusões para universo maior (população - finita ou não).

Conteúdo

Abordagens Inferenciais

- Clássica ou Frequentista: Fonte de informação lícita para estimar/ prever/ tomar decisões são os dados amostrais (ou empíricos).

- Bayesiana: Combina informação amostral a informação subjetiva para estimar/ prever/ tomar decisões. Instrumento usado para combinar as duas fontes de informação é Teorema de Bayes.

Conteúdo

Inferência Paramétrica

- Y é variável de interesse. Assume-se modelo probabilístico $f(y|\theta)$para descrever comportamento de Y na população - modelo observacional.

- θ é parâmetro (ou vetor de parâmetros), quantidade não observável que indexa o modelo observacional.

- Assume-se ser possível observar Y em amostra aleatória:

$$Y_1, \ldots, Y_n \overset{iid}{\sim} f(y|\theta).$$

$$\downarrow$$

(Informação amostral)

- Foco da Inferência no parâmetro $\theta \Rightarrow$ Inferência paramétrica.

Inferência Paramétrica - Modelo observacional

- No contexto de risco de crédito, Y é a variável aleatória que indica se cada indivíduo é um "bom" ou "mau" risco (por exemplo, para concessão de crédito).

- Assume-se que uma empresa dispõe de bases de dados, podendo classificar seus clientes passados, de acordo com critérios estabelecidos, como "bons" ou "maus".

- Existe, entretanto, a necessidade de classificar clientes em potencial (incerteza envolvida).

- Poderíamos "traduzir essa incerteza"por meio de um modelo probabilístico: $Y \sim Bernoulli(\theta)$.

Inferência Paramétrica - Modelo observacional

- Uma característica importante para a gestão de negócios é a frequência (contagem) de cancelamentos de contratos, contabilizada em períodos de tempo de igual duração.

- Seja Y a variável aleatória que denota frequência de cancelamentos e interesse em predizer cancelamentos, calcular probabilidade de que a contagem em um período futuro exceda certos patamares.

- Poderíamos, por exemplo, especificar um modelo probabilístico: $Y \sim Poisson(\theta)$.

- Outros modelos probabilísticos, adequados ao tratamento de contagens, poderiam ser usados.

- Outra característica importante para a gestão de negócios é o custo associado a cancelamentos de contratos.

- Seja Y a variável aleatória que denota tais custos. É bastante importante a caracterização de Y, por exemplo para avaliação de probabilidade de que custos totais ultrapassem limiares de equilíbrio financeiro.

- Uma possível especificação: $Y \sim Gama(\alpha, \lambda)$. Aqui, $\theta = (\alpha, \lambda)$.

- Outra possível especificação: definir $Y^* = ln(Y)$ e modelar Y^* segundo um modelo Normal (o que equivale a assumir que os custos, em sua escala original, seguem distribuição lognormal): $Y^* \sim Normal(\mu, \sigma^2)$. Aqui, $\theta = (\mu, \sigma^2)$.

Em cada uma das três situações, deseja-se formular um modelo probabilístico $f(y|\theta)$ que descreva o comportamento populacional (presumido) de Y, especificando o vetor paramétrico que o caracteriza.

Este é o primeiro passo da análise inferencial paramétrica: estabelecer um modelo observacional para descrição do comportamento de Y e definir sobre quais parâmetros se deseja inferir.

Modelos Probabilísticos Usuais

Distribuição de X	$f(y\|\theta)$	θ	$E[X\|\theta]$	$V[X\|\theta]$
Bernoulli(p)	$\theta^y(1-\theta)^{1-y}, \quad y=0,1, \theta=p$	p	$p(1-p)$	
Binomial(n,p)	$\binom{n}{y}\theta^y(1-\theta)^{n-y}, \quad y=0,1,\ldots,n$	p	np	$np(1-p)$
Pascal(r,p)	$\binom{y-1}{r-1}p^r(1-p)^{n-r}, \quad y=r,r+1,\cdots$	p	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$
Poisson(λ)	$\frac{e^{-\lambda}\lambda^y}{y!}, \quad y=0,1,2,\ldots$	λ	λ	λ
Exponencial(λ)	$\lambda e^{-\lambda y}, \quad y>0,\ldots$	λ	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gama(α,λ)	$\frac{\lambda^\alpha}{\Gamma(\alpha)}y^{\alpha-1}e^{-\lambda y}, \quad y>0$	(α,λ)	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
Qui-quadrado(k)	$\frac{1}{2^{k/2}}y^{k/2-1}e^{-y/2}, \quad y>0$	k	k	$2k$
Normal(μ,σ^2)	$\frac{}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}, \quad -\infty<y<\infty,\ldots$	(μ,σ^2)	μ	σ^2
Lognormal(μ,σ^2)	$\frac{1}{\sqrt{2\pi}\sigma y}e^{-(\log(y)-\mu)^2/2\sigma^2}$	(μ,σ^2)	$e^{(\mu+1/2\sigma^2)}$	$e^{(2\mu+\sigma^2)}(e^{\sigma^2}-1)$
Beta(α,β)	$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}, \quad 0<y<1$	(α,β)	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
t de Student(ν)	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{y^2}{\nu}\right)^{-(\nu+1)/2}$	ν	$0, \quad \nu>1$	$\frac{\nu}{\nu-2}, \quad \nu>2$

A observação de uma amostra aleatória é elemento comum às abordagens clássica e bayesiana.

Diz-se que $Y_1, Y_2, \ldots, Y_n$ formam uma amostra aleatória (simples) de $f(y|\theta)$ se $Y_1, Y_2, \ldots, Y_n$ são variáveis aleatórias independentes e igualmente distribuídas, segundo $f(y|\theta)$.

- o valor a ser observado para cada elemento é realização de uma variável aleatória;

Inferência Paramétrica - Amostra Aleatória

A observação de uma amostra aleatória é elemento comum às abordagens clássica e bayesiana.

Diz-se que $Y_1, Y_2, \ldots, Y_n$ formam uma amostra aleatória (simples) de $f(y|\theta)$ se $Y_1, Y_2, \ldots, Y_n$ são variáveis aleatórias independentes e igualmente distribuídas, segundo $f(y|\theta)$.

- o valor a ser observado para cada elemento é realização de uma variável aleatória;
- o valor a ser observado para Y_i independe de Y_j, se $i \neq j$ e

A observação de uma amostra aleatória é elemento comum às abordagens clássica e bayesiana.

Diz-se que $Y_1, Y_2, \ldots, Y_n$ formam uma amostra aleatória (simples) de $f(y|\theta)$ se $Y_1, Y_2, \ldots, Y_n$ são variáveis aleatórias independentes e igualmente distribuídas, segundo $f(y|\theta)$.

- o valor a ser observado para cada elemento é realização de uma variável aleatória;
- o valor a ser observado para Y_i independe de Y_j, se $i \neq j$ e
- o modelo que descreve a incerteza sobre cada elemento amostral Y_i é o mesmo que descreve o comportamento de Y na população de interesse, ou seja, o modelo observacional $f(y|\theta)$.

Conteúdo

Informação Amostral - Função de Verossimilhança

Seja $f(y_1, \ldots, y_n|\theta) = f_n(\mathbf{y}|\theta)$ a distribuição conjunta de $(Y_1, \ldots, Y_n)$. Dado um valor para θ, $f_n(\mathbf{y}|\theta)$ é um modelo probabilístico $n-$dimensional, que pode ser usado para se calcular probabilidades associadas ao vetor aleatório $(Y_1, \ldots, Y_n)$.

Em um problema de inferência, tem-se uma amostra aleatória observada, $y_1, \ldots, y_n$, e θ é desconhecido. Passamos a tratar $f_n(\mathbf{y}|\theta)$ como função de θ, para $\mathbf{y}$ dado.

$f_n(\mathbf{y}|\theta)$ passa a ser denotada, então, por $l(\theta; \mathbf{y})$ e é denominada função de verossimilhança.

Universidade Federal do Rio de Janeiro

Instituto de Matemática
Universidade Federal do Rio de Janeiro

Informação Amostral - Função de Verossimilhança

A função de verossimilhança é instrumento utilizado, tanto na abordagem clássica quanto na bayesiana, para resumir a informação sobre θ contida em uma amostra observada.

Se $Y_1, \ldots, Y_n$ formam uma amostra aleatória simples, isto é, se $Y_1, \ldots, Y_n \overset{iid}{\sim} f(y|\theta)$, então a função de verossimilhança é dada por:

$$l(\theta; \mathbf{y}) = f_n(\mathbf{y}|\theta) = \prod_{i=1}^{n} f(y_i|\theta). \tag{1}$$

Exemplo: Verossimilhança - modelo Bernoulli(θ)

$Y_1, \ldots, Y_n \overset{iid}{\sim}$Bernoulli($\theta$): n indivíduos entrevistados quando ao hábito ou não de praticar exercícios físicos.

$$ll(\theta; \mathbf{y}) = f_n(\mathbf{y}|\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^t(1-\theta)^{n-t}, t = \sum_{i=1}^{n} y_i.$$

Exemplo: Verossimilhança - modelo Bernoulli(θ)

Código R

```
n<-10
t1<-1
t2<-5
theta<-seq(0,1,0.01)
vero1<-theta^t1*(1-theta)^(n-t1)
vero2<-theta^t2*(1-theta)^(n-t2)
m<-100
t3<-10
t4<-50
vero3<-theta^t3*(1-theta)^(m-t3)
vero4<-theta^t4*(1-theta)^(m-t4)
par(mfrow=c(2,2))
plot(theta,vero1,type="l",ylab="n=10, t=1")
plot(theta,vero2, type="l",ylab="n=10, t=5")
plot(theta,vero3,type="l",ylab="n=100, t=10")
plot(theta,vero4, type="l",ylab="n=100, t=50")
```

Exemplo: Verossimilhança - modelo Bernoulli(θ)

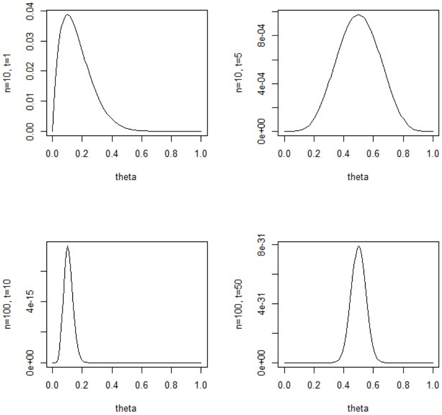


Figura 1: Comportamento da função de verossimilhança Bernoulli. Linha superior: $n = 10$, $t = 1$, $t = 5$. Linha inferior: $n = 100$, $t = 10$, $t = 50$.

Conteúdo

Objetivos neste curso:

- Estimação pontual de parâmetros: objetiva-se encontrar um único valor que aproxime o valor teórico de um parâmetro;

Objetivos neste curso:

- Estimação pontual de parâmetros: objetiva-se encontrar um único valor que aproxime o valor teórico de um parâmetro;

- Estimação intervalar de parâmetros: objetiva-se encontrar uma região de valores "razoáveis" para o parâmetro, admitindo-se formalmente a incerteza inerente ao processo de estimação;

Objetivos neste curso:

- Estimação pontual de parâmetros: objetiva-se encontrar um único valor que aproxime o valor teórico de um parâmetro;
- Estimação intervalar de parâmetros: objetiva-se encontrar uma região de valores “razoáveis”para o parâmetro, admitindo-se formalmente a incerteza inerente ao processo de estimação;
- Previsão Com base no que se aprendeu, a partir de observações passadas, deseja-se prever o comportamento de observações futura;

Objetivos neste curso:

- Estimação pontual de parâmetros: objetiva-se encontrar um único valor que aproxime o valor teórico de um parâmetro;
- Estimação intervalar de parâmetros: objetiva-se encontrar uma região de valores "razoáveis"para o parâmetro, admitindo-se formalmente a incerteza inerente ao processo de estimação;
- Previsão Com base no que se aprendeu, a partir de observações passadas, deseja-se prever o comportamento de observações futura;
- Testes de hipóteses: com base em amostras observadas, confrontam-se hipóteses sobre os parâmetros, objetivando-se escolher aquela que melhor retrate o real mecanismo gerador dos dados observados;

Objetivos neste curso:

- Dada uma amostra observada sobre uma variável aleatória Y, deseja-se avaliar se determinado modelo probabilístico teórico é adequado à descrição do comportamento de Y.

Objetivos neste curso:

- Dada uma amostra observada sobre uma variável aleatória Y, deseja-se avaliar se determinado modelo probabilístico teórico é adequado à descrição do comportamento de Y.

- Para todos os processos descritos, é possível que haja informação relevante, externa aos dados amostrais. O paradigma bayesiano permite a inclusão dessas informações na análise.

Objetivos neste curso:

- Dada uma amostra observada sobre uma variável aleatória Y, deseja-se avaliar se determinado modelo probabilístico teórico é adequado à descrição do comportamento de Y.
- Para todos os processos descritos, é possível que haja informação relevante, externa aos dados amostrais. O paradigma bayesiano permite a inclusão dessas informações na análise.
- Importante ressaltar que, como veremos, o uso do paradigma bayesiano não se restringe às situações em que efetivamente tenhamos informações externas aos dados disponíveis.

Conteúdo

Universidade Federal do Rio de Janeiro

Instituto de Matemática
Universidade Federal do Rio de Janeiro

Admita $Y \sim f(y|\theta)$.

O conjunto de todos os possíveis valores de Y é denominado espaço amostral (induzido por Y)

O conjunto de todos os possíveis valores de θ é denominado espaço paramétrico (denotaremos por Θ).

Para cada valor de θ no espaço paramétrico Θ, tem-se uma curva $f(y|\theta)$

Estimação Pontual - Família paramétrica

$\mathcal{F} = \{f(y|\theta), \theta \in \Theta\}$ forma uma família paramétrica.

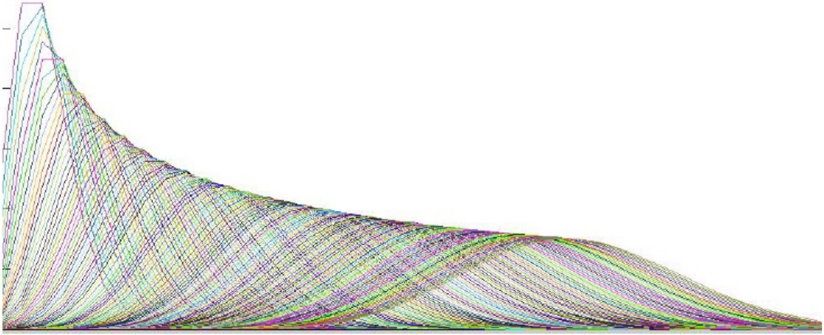


Figura 2: Exemplo de família paramétrica de distribuições: para cada valor de θ.

Estimação Pontual - Família paramétrica

O problema de estimação pontual consiste em aproximar o "verdadeiro" valor de θ.

Assuma que a escolha da família paramétrica tenha sido adequada. Objetiva-se, então, ao se estimar θ pontualmente, aproximar a curva $f(y|\theta)$ geradora da amostra $Y_1, Y_2, \ldots, Y_n$, dentro da família paramétrica $\mathcal{F} = \{f(y|\theta), \theta \in \Theta\}$.

Tal objetivo, claramente, estará sujeito a erro.

Estimação Pontual - Família paramétrica

"Verdadeiro" valor de θ é questionável.

A própria suposição de que o modelo observacional adotado é o "verdadeiro" macanismo gerador dos dados é questionável.

"*Essencialmente, todos os modelos estão errados, mas alguns são úteis.*" (George Box)

Desejamos obter modelos úteis e estimar seus parâmetros de forma a produzir resultados (estimações/predições) condizentes com observações práticas sobre os processos analisados; modelos que nos auxiliem a tomar decisões adequadas.

Seja $\mathbf{Y} = (Y_1, \ldots, Y_n)$ uma amostra aleatória. Uma estatística $T = h(\mathbf{Y})$ é uma função da amostra aleatória.

São exemplos de estatísticas:

- a média amostral, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$;

Estimação Pontual Clássica

Seja $\mathbf{Y} = (Y_1, \ldots, Y_n)$ uma amostra aleatória. Uma estatística $T = h(\mathbf{Y})$ é uma função da amostra aleatória.

São exemplos de estatísticas:

- a média amostral, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$;
- a variância amostral, $S^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$;

Seja $\mathbf{Y} = (Y_1, \ldots, Y_n)$ uma amostra aleatória. Uma estatística $T = h(\mathbf{Y})$ é uma função da amostra aleatória.

São exemplos de estatísticas:

- a média amostral, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$;
- a variância amostral, $S^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$;
- estatísticas de ordem, como mínimo e máximo amostrais: $Y_{(1)} = min_i Y_i$; $Y_{(n)} = max_i Y_i$.

Estimação Pontual Clássica

Um estimador pontual clássico para θ é uma estatística.

Ou seja, um estimador pontual clássico é uma função do vetor amostral $\mathbf{Y}$.

Após observar-se uma particular amostra aleatória, tem-se uma estimativa pontual para θ: $\hat{\theta} = h(y_1, y_2, \ldots, y_n)$.

Exemplo: Estimador Pontual Clássico

Pode-se ter interesse em estimar o custo médio μ associado a cancelamentos de contratos.

Assuma que $Y \sim f(y|\theta)$ e que μ seja uma função de θ. Por exemplo, se assumimos que os custos Y de cancelamento seguem distribuição $Gama(\alpha, \lambda)$, temos $\mu = \alpha/\lambda$. Um estimador possível é a média amostral: $\hat{\mu} = h(Y_1, \ldots, Y_n) = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

Uma vez observada uma amostra aleatória de $n = 100$ custos, tem-se $\bar{y} = 1502,70$ u.m, que é uma estimativa pontual para μ.

Questão: Por que usamos a média amostral como estimador de μ? Haveria outros estimadores possíveis, ou seja, outras formas de resumir a informação amostral, para esse mesmo fim?

Métodos Clássicos de Estimação

Há diferentes estimadores para um parâmetro e diferentes métodos para obtenção de estimadores.

Os métodos clássicos de estimação propõem estimadores que são funções, unicamente, da amostra $(Y_1, \ldots, Y_n)$.

Estimação Pontual - Método dos Momentos

Usa a noção intuitiva de que quantidades teóricas populacionais devem ser estimadas por análogos amostrais.

Por exemplo, a média de uma população é estimada, de acordo com esse método, pela média amostral.

O momento teórico de ordem k de uma variável aleatória Y é dado por $\mu_k = E[Y^k]$, $k = 1, 2, \ldots$, se tal esperança é finita.

Obs: $\mu_1 = E[Y]$, $\mu_2 = V[Y] + (\mu_1)^2$.

O momento amostral de ordem k de uma variável aleatória Y com função (densidade) de probabilidade $f(y|\theta)$ é dado por $M_k = \frac{\sum_{i=1}^n Y_i^k}{n}$, $k = 1, 2, \ldots$.

Método dos Momentos - Operacionalização

- Se o vetor θ tem d componentes, utilizam-se d relações:

$$\mu_j = g_j(\theta), \quad j = 1, \ldots, d.$$

Método dos Momentos - Operacionalização

- Se o vetor θ tem d componentes, utilizam-se d relações:

$$\mu_j = g_j(\theta), \quad j = 1, \ldots, d.$$

- Transpõe-se as relações acima para o âmbito amostral, substituindo-se momentos teóricos por momentos amostrais e parâmetro(s) por estimador(es):

$$M_j = g_j(\hat{\theta}), \quad j = 1, \ldots, d.$$

Método dos Momentos - Operacionalização

- Se o vetor θ tem d componentes, utilizam-se d relações:

$$\mu_j = g_j(\theta), \quad j = 1, \ldots, d.$$

- Transpõe-se as relações acima para o âmbito amostral, substituindo-se momentos teóricos por momentos amostrais e parâmetro(s) por estimador(es):

$$M_j = g_j(\hat{\theta}), \quad j = 1, \ldots, d.$$

- o conjunto de d equações obtido forma um sistema que deve ser resolvido para $\hat{\theta}$, obtendo-se estimador(es) de θ.

Método dos Momentos - Exemplo: Exponencial

Seja Y a variável aleatória que denota o tempo até o resgate de um título e admita que $Y \sim Exponencial(\theta)$. Sabemos que

$$\mu_1 = E[Y|\theta] = g(\theta) = 1/\theta.$$

Transpondo essa relação para amostra, tem-se:

$$M_1 = \frac{\sum_{i=1}^{n} Y_i}{n} = g(\theta) = 1/\hat{\theta}$$

e, portanto, um estimador de θ, pelo método dos momentos, é $\hat{\theta} = 1/\bar{Y}$.

Método dos Momentos - Exemplo: Normal(μ, σ^2)

Seja Y a variável aleatória que denota o volume de preenchimento de latas em uma linha de produção.

Assuma que $Y \sim Normal(\mu, \sigma^2)$, ou seja, $\theta = (\mu, \sigma^2)$. Então:

$\mu_1 = E[Y|\theta] = g_1(\theta) = \mu; \quad \mu_2 = E[Y^2|\theta] = V[Y|\theta] + \mu_1^2 = g_2(\theta) = \sigma^2 + \mu^2.$

Transpondo essa relação para amostra, tem-se:

$$M_1 = \bar{Y} = g_1(\hat{\theta}) = \hat{\mu};$$

$$M_2 = \frac{\sum_{i=1}^n Y_i^2}{n} = g_2(\hat{\theta}) = \hat{\sigma^2} + \hat{\mu}^2$$

$$\Rightarrow \hat{\sigma^2} = \frac{\sum_{i=1}^n Y_i^2}{n} - \bar{Y}^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n}.$$

Estimação Pontual - Método de Máxima Verossimilhança

O estimador de máxima verossimilhança para θ é o argumento $\hat{\theta}$ para o qual a função de verossimilhança $l(\theta; \mathbf{y})$ assume valor máximo.

Obs: O argumento $\hat{\theta}$ que maximiza a função de verossimilhança é o mesmo que maximiza a fução de log-verossimilhança $L(\theta; \mathbf{y}) = \log[l(\theta; \mathbf{y})]$.

Na maioria dos casos, θ varia continuamente sobre um intervalo, ou mais geralmente, sobre uma região do $\Re^p$, para algum p.

Método de Máxima Verossimilhança - Operacionalização

$\Rightarrow$ O estimador de máxima verossimilhança, tipicamente, pode ser encontrado através da solução de

$$\frac{\partial l(\theta; \mathbf{y})}{\partial \theta_j} = 0, \quad j = 1, \ldots, p,$$

ou equivalentemente,

$$\frac{\partial L(\theta; \mathbf{y})}{\partial \theta_j} = 0, \quad j = 1, \ldots, p.$$

Deve-se, ainda, verificar se o ponto $\hat{\theta}$ ótimo obtido é, de fato, argumento máximo da função de verossimilhança.

Estimador de Máxima Verossimilhança - Exemplo: Poisson(λ)

Y: frequências de cancelamentos de contratos, por dia.
$Y_1, \ldots, Y_n \overset{iid}{\sim} Poisson(\lambda)$.

$$
\begin{aligned}
l(\lambda; \mathbf{y}) &= \prod_{i=1}^{n} f(y_i|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} = \frac{e^{-n\lambda}\lambda^{\sum_i y_i}}{\prod_i y_i!} \\
\Rightarrow L(\lambda; \mathbf{y}) &= -n\lambda + \sum_i y_i \log(\lambda) - \log(\prod_i y_i!) \\
\Rightarrow \frac{dL(\lambda; \mathbf{y})}{d\lambda} &= -n + \frac{\sum_i y_i}{\lambda} = 0 \\
& \hat{\lambda} = \frac{\sum_i y_i}{n} = \bar{Y}.
\end{aligned}
$$

Estimador de Máxima Verossimilhança - Exemplo: Poisson(λ)

Ou seja, o método de máxima verossimilhança sugere que o parâmetro λ, número médio de cancelamentos, por dia, deva ser estimado pelo número médio amostral de cancelamentos por dia.

Vamos, agora, obter a estimativa de máxima verossimilhança para λ, com base na seguinte amostra, de $n = 430$ observações:

y	Frequência
0	60
1	80
2	120
3	90
4	50
5	20
6	10
7+	0
Total (n)	430

$$\hat{\lambda} = \bar{Y} = \frac{0 \times 60 + \ldots + 6 \times 10}{430}$$

$$\approx 2,209302.$$

Estimativa de Máxima Verossimilhança - Obtenção Numérica

Vejamos como obter uma aproximação numérica para a estimativa de máxima verossimilhança, no R, sem necessidade de passar pelo desenvolvimento analítico para obtençaõ do estimador:

```
#Entrada de dados
y <- c(rep(0,60), rep(1,80), rep(2,120), rep(3,90),
rep(4, 50), rep(5,20), rep(6, 10))
#Visualização da distribuição da amostra observada
hist(y, main="Histograma de y",right=FALSE, prob=TRUE)
#tamanho amostral
n <- length(y)
theta<-NULL
# -logverossimilhança
neglogvero<-function(theta)
{ -sum(log(dpois(y,theta)))}
#resultado:
# A função nlm retorna o argumento mínimo. Por isso
#fornecemos a log verossimilhança negativa.
# p=c(0.5) é um chute inicial para a estimativa pontual.
saida<-nlm(neglogvero,p=c(0.5))
saida$estimate
```

Estimador de Máxima Verossimilhança - Exemplo: Gama(α, λ)

Y: Custos associados a cancelamentos de contratos. Assuma $Y_1, \dots, Y_n \overset{iid}{\sim} Gama(\alpha, \lambda)$.

Código R, para aproximação numérica de estimativas de máxima verossimilhança para os parâmetros α e λ retorna as estimativas $\hat{\alpha} \approx 1,35$ e $\hat{\lambda} \approx 0,36$:

```
#Entrada de dados
y <- c(2.07,0.38,11.77,3.66,4.02,8.20,1.88,2.92,2.83,1.76,
0.47,0.36,2.76,0.75,3.64,4.07,2.28,7.14,2.63,11.45)
#Visualização da distribuição da amostra observada
hist(y, main="Histograma de y",right=FALSE, prob=TRUE)
#tamanho amostral
n <- length(y)
a <- NULL
b <- NULL
# -logverossimilhança
theta<-c(a,b)
neglogvero<-function(theta){-sum(log(dgamma(y,theta[1],theta[2])))
#resultado:
saida¡-nlm(neglogvero,p=c(0.5,0.5))
saida$estimate
```

Universidade Federal do Rio de Janeiro

Instituto de Matemática
Universidade Federal do Rio de Janeiro

Propriedades de Estimadores

A distribuição de probabilidade de um estimador é denominada distribuição amostral.

Descreve a variação sofrida pelo estimador, caso fosse observado em diferentes amostras aleatórias de mesmo tamanho.

Como todo estimador clássico $\hat{\theta}$ é uma variável aleatória, pode-se obter seu valor esperado e sua variância.

Vício (ou tendência ou viés) de um estimador $\hat{\theta}$ é dado pela diferença entre seu valor esperado e o verdadeiro valor do parâmetro θ que se deseja estimar. Um estimador com vício nulo é dito não viciado(ou não tendencioso ou não viesado).

Propriedades de Estimadores

A precisão de um estimador $\hat{\theta}$ é o inverso de sua variância.

Um estimador é dito preciso se as estimativas produzidas por este oscilam pouco, quando observado em diferentes amostras, ou seja, se sua variância é pequena.

Uma sequência de estimadores $\hat{\theta}_n$, formada pelo estimador $\hat{\theta}$ tomado em amostras de tamanho $n = 1, 2, \ldots$ é dita consistente se tal sequência converge (em algum sentido além dos propósitos deste curso) para o verdadeiro valor θ.

Propriedades de Estimadores

Problema de estimação pontual é como um “jogo de tiro ao alvo”. Como θ é desconhecido, precisamos explorar as propriedades dos estimadores para termos confiança de que estamos “atirando” bem, isto é, produzindo estimativas próximas ao alvo (acuradas).

Figura 3: Ilustração do problema de estimação: precisão e vício

Propriedades dos Estimadores de Máxima Verossimilhança

- Normalidade Assintótica: para amostras suficientemente grandes e sob certas condições de regularidade, o estimador de máxima verossimilhança de um parâmetro θ tem distribuição aproximadamente normal, com média θ e variância que depende inversamente do tamanho amostral.

Propriedades dos Estimadores de Máxima Verossimilhança

- Normalidade Assintótica: para amostras suficientemente grandes e sob certas condições de regularidade, o estimador de máxima verossimilhança de um parâmetro θ tem distribuição aproximadamente normal, com média θ e variância que depende inversamente do tamanho amostral.

- Ou seja, sob condições de regularidade, o vício de um estimador de máxima verossimilhança é desprezível e sua precisão, tão alta quanto se queira (ao menos em tese, pressupondo-se ser possível aumentar o tamanho amostral tanto quanto necessário).

Propriedades dos Estimadores de Máxima Verossimilhança

- Normalidade Assintótica: para amostras suficientemente grandes e sob certas condições de regularidade, o estimador de máxima verossimilhança de um parâmetro θ tem distribuição aproximadamente normal, com média θ e variância que depende inversamente do tamanho amostral.

- Ou seja, sob condições de regularidade, o vício de um estimador de máxima verossimilhança é desprezível e sua precisão, tão alta quanto se queira (ao menos em tese, pressupondo-se ser possível aumentar o tamanho amostral tanto quanto necessário).

- Invariância: Se $\hat{\theta}$ é estimador de máxima verossimilhança de θ e $\psi = g(\theta)$, então $\hat{\psi} = g(\hat{\theta})$.

Propriedades de Máxima Verossimilhança - Exemplo

No exemplo anterior, obtivemos, com base em uma amostra observada, estimativas de máxima verossimilhança para os parâemtros α e λ de um modelo Gama para custos de cancelamento Y.

O custo esperado é

$$\mu = E[Y|\alpha, \lambda] = \alpha/\lambda.$$

Então, pela propriedade de invariância, o estimador de máxima verossimilhança para μ é dado por

$$\hat{\mu} = \hat{\alpha}/\hat{\lambda}$$

e, com base na amostra observada no exemplo anterior, a estimativa pontual de μ é

$$\hat{\mu} = 1,35/0,36 = 3,75$$

Conteúdo

Quais são as informações disponíveis?

Elementos de Inferência Bayesiana

Quais são as informações disponíveis?

Informação contida na amostra coletada

Elementos de Inferência Bayesiana

Quais são as informações disponíveis?

Informação contida na amostra coletada

Informação externa à amostra?

Elementos de Inferência Bayesiana

Quais são as
informações
disponíveis?

Combinar
as duas fontes
de informação

Informação contida
na amostra coletada

Informação
externa
à amostra?

Elementos de Inferência Bayesiana

Inferência Bayesiana!

Quais são as informações disponíveis?

Combinar as duas fontes de informação

Informação contida na amostra coletada

Informação externa à amostra?

Como procedemos?

$$Y \sim f(y|\theta)$$

(modelo observacional)

Como procedemos?

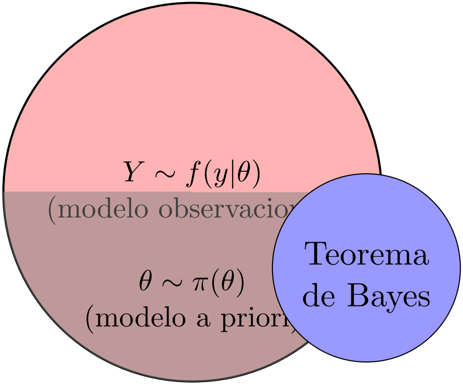
$$Y \sim f(y|\theta)$$

(modelo observacional)

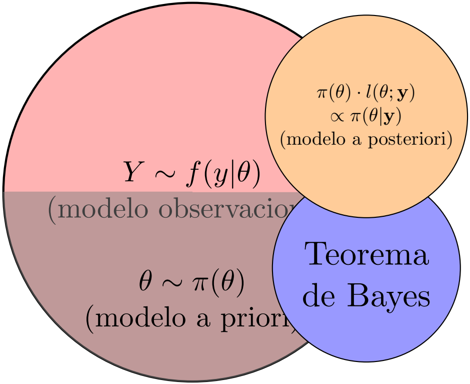
$$\theta \sim \pi(\theta)$$

(modelo a priori)

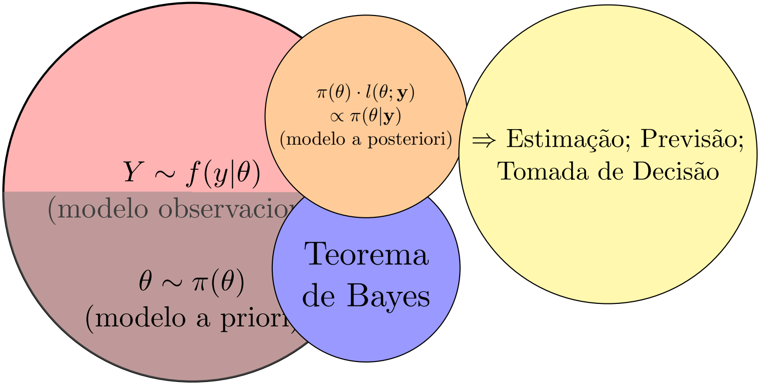
Como procedemos?



Como procedemos?



Como procedemos?



Conteúdo

Elementos de Inferência Bayesiana: Priori

Distribuição a priori $\pi(\theta)$: modelo probabilístico associado ao parâmetro θ, descrevendo a incerteza/informação sobre θ prévia à observação dos dados. A distribuição a priori:

- reflete as crenças de cada analista
- formalmente incorpora subjetivismo à análise.
- Variância a priori controla nível de incerteza (subjetiva) do analista.
- Variância a priori pequena $\Rightarrow$pouca incerteza.
- Quaisquer informações relevantes externas aos dados podem ser incorporadas à inferência por meio da distribuição a priori.

Uso de distribuição a priori - Ilustração

- Duas grandes marcas dividem o mercado de certo produto. Historicamente, cada uma detém uma fatia de aproximadamente 50% do mercado.

Uso de distribuição a priori - Ilustração

- Duas grandes marcas dividem o mercado de certo produto. Historicamente, cada uma detém uma fatia de aproximadamente 50% do mercado.

- Um analista é o encarregado de manter estimativas atualizadas sobre o volume de vendas da marca 1. Ele costuma manter suas estimativas atualizadas com base em um modelo Normal atualizado de forma dinâmica, alimentado pelas informações das vendas do produto da marca 1, dando maior peso às vendas de meses recentes.

Uso de distribuição a priori - Ilustração

- O analista acaba de ser informado que a marca 2 foi retirada do mercado e entende que a marca 1 precisa reagir rapidamente a essa informação, adequando seus estoques e distribuição para o mês seguinte.

Uso de distribuição a priori - Ilustração

- O analista acaba de ser informado que a marca 2 foi retirada do mercado e entende que a marca 1 precisa reagir rapidamente a essa informação, adequando seus estoques e distribuição para o mês seguinte.

- A informação de que a marca 2 foi retirada do mercado não está na série histórica de suas vendas. O analista precisa fazer uso de informação externa à série histórica de vendas e calibrar suas estimativas (sistema de aprendizagem) para incorporação dessa nova informação. Pode fazer uso de distribuição a priori, para isso.

Paradigma bayesiano de aprendizagem - quando usar?

- Só se utiliza o mecanismo bayesiano de aprendizagem quando se dispõe de informação sobre os parâmetros externa aos dados observados?

Paradigma bayesiano de aprendizagem - quando usar?

- Só se utiliza o mecanismo bayesiano de aprendizagem quando se dispõe de informação sobre os parâmetros externa aos dados observados?
- Não! Podemos fazer especificações de distribuições a priori bastante vagas (flat/dispersas), refletindo nossa falta de conhecimento sobre os parâmetros ou quaisquer quantidades não observáveis de interesse.

Paradigma bayesiano de aprendizagem - quando usar?

- Só se utiliza o mecanismo bayesiano de aprendizagem quando se dispõe de informação sobre os parâmetros externa aos dados observados?
- Não! Podemos fazer especificações de distribuições a priori bastante vagas (flat/dispersas), refletindo nossa falta de conhecimento sobre os parâmetros ou quaisquer quantidades não observáveis de interesse.
- Nesse caso, a informação vinda dos dados, resumida pela função de verossimilhança, dominará a inferência.

Conteúdo

- Função de verossimilhança $l(\theta; \mathbf{y})$ resume a informação amostral sobre θ.
- Tipicamente, quanto maior a amostra, mais concentrada a função de verossimilhança em torno da estimativa de máxima verossimilhança.

Conteúdo

Teorema de Bayes e Distribuição a posteriori

Priori e Verossimilhança resumem as informações de entrada no sistema Bayesiano de aprendizagem.

Deseja-se combinar as entradas, ou seja, a informação amostral e a informação a priori, para obter como saída modelo atualizado para descrição das incertezas sobre θ:

$$\pi(\theta|\mathbf{y}) \rightarrow \text{distribuição a posteriori}.$$

Teorema de Bayes e Distribuição a posteriori

A distribuição a posteriori de θ é dada por:

$$\pi(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)\pi(\theta)}{f(\mathbf{y})} \propto l(\theta;\mathbf{y})\pi(\theta). \tag{2}$$

e atualiza as crenças a priori do analista, à luz da informação observada nos dados amostrais. Note que o denominador

$$f(\mathbf{y}) = \int_{\Theta} f(\mathbf{y}|\theta)\pi(\theta)d\theta \tag{3}$$

é constante com respeito a θ, então, muitas vezes é possivel identificar a distribuição a posteriori somente através do produto entre priori e verossimilhança.

Conteúdo

Exemplo: análise bayesiana, modelo Normal

- Y é variável que denota nível de colesterol em certa população de adultos.
- Modelo observacional: $Y \sim N(\theta, \sigma^2)$.
- Modelo a priori para θ: $\theta \sim N(\mu_0, \tau_0^2)$.
- σ^2, μ_0 e τ_0^2 são supostos conhecidos.

Aplicando-se o teorema de Bayes, pode-se mostrar que

$$\theta | y \sim N(\mu_1, \tau_1^2)$$

$$\mu_1 = \omega\mu_0 + (1-\omega)y, \quad \tau_1^{-2} = \tau_0^{-2} + \sigma^{-2}, \quad \omega = \tau_0^{-2}/(\tau_0^{-2} + \sigma^{-2}).$$

Exemplo: análise bayesiana, modelo Normal

- Média a posteriori, μ_1 é uma média ponderada entre a média a priori, μ_0 e a observação y.
- Caso tivéssemos mais de uma observação, μ_1 seria uma média ponderada entre μ_0 e $\bar{y}$, a média amostral.
- Os pesos ω são governados pelas precisões dos modelos a priori e observacional.
- Caso tivéssemos mais de uma observação, a precisão observacional, σ^{-2}, apareceria na expressão de ω multiplicada pelo tamanho amostral n.
- Precisão do modelo a posteriori é dada pela soma das precisões a priori e observacional.

Exemplo: análise bayesiana, modelo Normal

Ilustração (Box & Tiao,1973): θ é grandeza física.

- Opiniões a priori sobre θ:
  Físico A (experiente): $\theta \sim N(900, (20)^2)$
  Físico B (menos experiente): $\theta \sim N(800, (80)^2)$.

- Valor observado na medição (sujeito a erro): $y = 850$.

- Distribuições a posteriori resultantes:
  Físico A: $(\theta|y = 850) \sim N(890, (17, 9)^2)$;
  Físico B: $(\theta|y = 850) \sim N(840, (35, 7)2)$.

- Para o físico A, posteriori e priori são próximas, mas para o Físico B, não.

Exemplo: análise bayesiana, modelo Normal

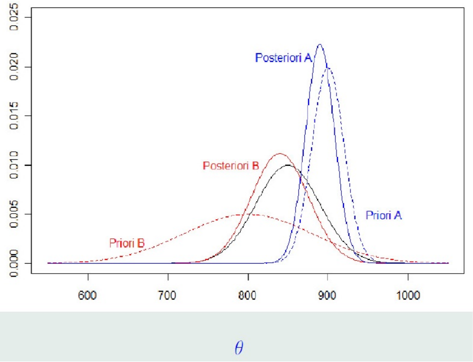


Figura 4: Verossimilhança (curva em preto) e distribuições a priori e a posteriori para os físicos A e B.

Exemplo: Fatia do mercado

- Companhia deve decidir se vai lançar ou não um novo produto.
- θ é a proporção do mercado que a companhia vai capturar.
- Espaço paramétrico (conjunto de valores possíveis para θ) é $\Theta = [0, 1]$.
- Informações a priori:
  Analista 1:
  $\theta \sim U[0, 1] \Rightarrow E[\theta] = 0,5; \quad V[\theta] = 0,083; \quad CV[\theta] = 58\%;$
  Analista 2:
  $\theta \sim U[0,1; 0,2] \Rightarrow E[\theta] = 0,15; \quad V[\theta] = 0,00008; \quad CV[\theta] = 19\%;$
  Analista 3:
  $\theta \sim Beta(1,2; 8) \Rightarrow E[\theta] = 0,13; \quad V[\theta] = 0,10; \quad CV[\theta] = 244\%..$

Exemplo: Fatia do mercado

- Informação amostral: n pessoas, selecionadas ao acaso, entrevistadas: compraria ou não produto?
- $Y_1, \ldots, Y_n \overset{iid}{\sim} Bern(\theta)$
- Função de verossimilhança: $l(\theta; \mathbf{y}) = f_n(\mathbf{y}|\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^t(1-\theta)^{n-t}, t = \sum_{i=1}^{n} y_i.$
- Entre $n = 10$ pessoas entrevistadas, $t = 2$ comprariam o produto, implicando que $l(\theta; \mathbf{y}) = \theta^2(1-\theta)^8.$

Exemplo: Fatia do mercado

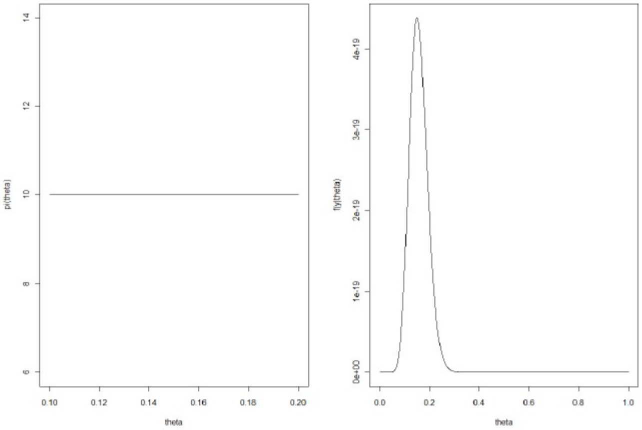


Figura 5: Informação a priori (analista 1) e verossimilhança (informação amostral) para θ.

Exemplo: Fatia do mercado

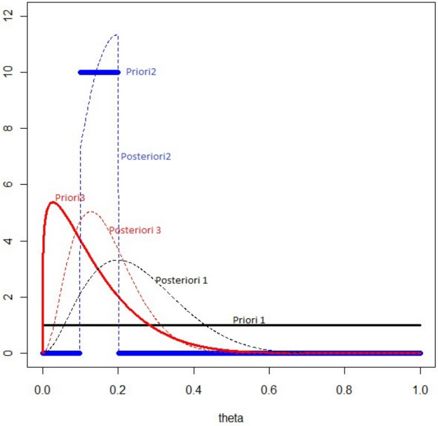


Figura 6: Distribuições a priori e a posteriori para θ, com base em $n = 10$ e $t = 2$, para os analistas 1, 2 e 3.

Exemplo: Fatia do mercado

A amostra observada é pequena ($n = 10$) $\Rightarrow$ diferentes modelos a priori levam a modelos a posteriori bastante distintos, ainda que a informação observada na amostra seja a mesma.

Admita, agora, uma amostra bem maior, com $n = 100$ pessoas entrevistadas, das quais $t = 15$ comprariam o produto

$$l(\theta; \mathbf{y}) = \theta^{15}(1-\theta)^{85}.$$

Exemplo: Fatia do mercado

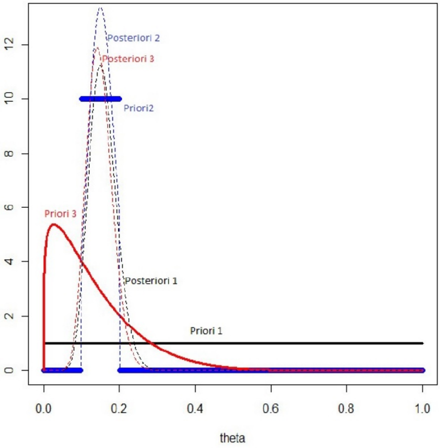


Figura 7: Distribuições a priori e a posteriori para θ, com base em $n = 100$ e $t = 15$, para os analistas 1, 2 e 3.

Conteúdo

Distribuições a Priori e Conjugação

- O analista é livre para escolher a distribuição a priori que julga refletir suas crenças.
- Pode-se escolher priori por conveniência analítica.
- Há distribuições de probabilidade a priori que, quando combinadas à função de verossimilhança, geram modelo a posteriori com forma analítica conhecida e pertencente à mesma família da priori.
- Diz-se então que a distribuição a priori é conjugada ao modelo observacional.

Conjugação: Bernoulli-Beta

Bernoulli-Beta

Se $Y_1, \ldots, Y_n \stackrel{iid}{\sim} Bernoulli(\theta)$ e assume-se, a priori, $\theta \sim Beta(\alpha, \beta)$, então, a posteriori, tem-se
$(\theta | y_1, \ldots, y_n) \sim Beta(\alpha + t, \beta + n - t), \quad t = \sum_{i=1}^{n} y_i$.

Sabemos, do teorema de Bayes, que

$$\begin{aligned} \pi(\theta|\mathbf{y}) &\propto l(\theta;\mathbf{y})\pi(\theta) \propto \theta^t(1-\theta)^{n-t} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{\alpha+t-1}(1-\theta)^{\beta+n-t-1} \end{aligned}$$

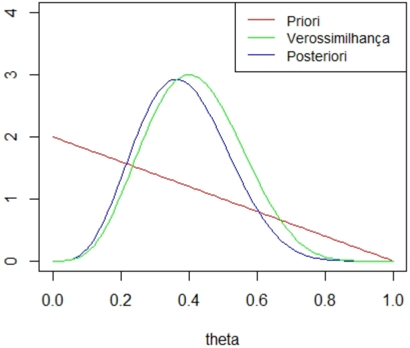
que, a menos de uma constante multiplicativa, é o núcleo de uma densidade $Beta(\alpha_1, \beta_1)$, com: $\alpha_1 = \alpha + t, \quad \beta_1 = \beta + n - t$. Então $(\theta|\mathbf{y}) \sim Beta(\alpha_1, \beta_1)$.

Conjugação Bernoulli-Beta: Exemplo

- Y é variável que indica se cada segurado vai aderir a um programa de exercícios físicos, caso tenha alguma redução na mensalidade de seu plano de saúde. $Y \sim Bernoulli(\theta)$, θ, a proporção de segurados que teriam interesse em aderir a um programa de exercícios físicos.
- Assuma, a priori: $\theta \sim Beta(\alpha, \beta)$.
- Informação a priori(subjetiva), baseada em experiência já adquirida com outra população: $E[\theta] = 1/3$ e probabilidade em torno de 10% de que a proporção θ seja superior a $0,7$.
- A partir das características associadas à distribuição Beta (veja a tabela de distribuições), é possível verificar que a escolha de $\alpha = 1$ e $\beta = 2$ retrata as crenças do analista sobre β.

Conjugação Bernoulli-Beta: Exemplo

- Pesquisa feita com $n = 10$ segurados, selecionados aleatoriamente e avaliados quanto ao interesse ou não em aderir ao programa: $Y_1, \ldots, Y_n \overset{iid}{\sim} Bernoulli(\theta)$ e observa-se que 4 segurados demosntrarm interesse no programa.



$\Rightarrow$ Posteriori: $\theta \sim Beta(1+4, 2+6) \equiv Beta(5, 8)$

Conjugação Bernoulli-Beta: Exemplo

Uma vez obtida a posteriori (distribuição que atualiza as crenças iniciais sobre θ à luz dos dados), pode-se fazer as mais diversas inferências sobre θ (proporção da população de segurados que aceitaria ingressar no programa de exercícios). Por exemplo:

- uma estimativa pontual, a posteriori, para θ, pode ser dada pela média da distribuição a posteriori:
  $\alpha^*/(\alpha^* + \beta^*) = 5/(5 + 8) = 0,385$;

Conjugação Bernoulli-Beta: Exemplo

Uma vez obtida a posteriori (distribuição que atualiza as crenças iniciais sobre θ à luz dos dados), pode-se fazer as mais diversas inferências sobre θ (proporção da população de segurados que aceitaria ingressar no programa de exercícios). Por exemplo:

- uma estimativa pontual, a posteriori, para θ, pode ser dada pela média da distribuição a posteriori:
  $\alpha^*/(\alpha^* + \beta^*) = 5/(5+8) = 0,385$;

- A proporção de segurados que aceitariam ingressar no programa está, com 95% de probabilidade, entre $0,15$ e $0,65$.

Conjugação Bernoulli-Beta: Exemplo

Uma vez obtida a posteriori (distribuição que atualiza as crenças iniciais sobre θ à luz dos dados), pode-se fazer as mais diversas inferências sobre θ (proporção da população de segurados que aceitaria ingressar no programa de exercícios). Por exemplo:

- uma estimativa pontual, a posteriori, para θ, pode ser dada pela média da distribuição a posteriori:
  $\alpha^*/(\alpha^* + \beta^*) = 5/(5+8) = 0,385$;
- A proporção de segurados que aceitariam ingressar no programa está, com 95% de probabilidade, entre $0,15$ e $0,65$.
- Observe que há menos incerteza sobre θ, ao se comparar as distribuições a priori e a posteriori, mas, ainda assim, há bastante incerteza sobre a proporção de segurados que aceitariam ingressar no programa.

Conjugação Bernoulli-Beta: Exercício para fixação de conteúdo

Exercício: Refaça a análise admitindo a mesma distribuição a priori, mas assuma que tenha sido observada uma amostra maior, composta por 100 segurados, os quais 40 respondem afirmativamente à pesquisa.

- Primeiramente, compare as curvas de distribuição a priori, verossimilhança e posteriori. Comente.
- Em seguida, explore a distribuição a posteriori, levantando fazendo afrmações inferenciais sobre θ, a partir desta.
- Compare seus resultados àqueles obtidos com base na amostra de 10 segurados.

Conjugação: Poisson-Gama

Poisson-Gama

Se $Y_1, \ldots, Y_n \overset{iid}{\sim} Poisson(\theta)$ e assume-se, a priori, $\theta \sim Gama(\alpha, \beta)$, então, a posteriori, tem-se
$(\theta|y_1, \ldots, y_n) \sim Gama(\alpha + t, \beta + n), \quad t = \sum_{i=1}^{n} y_i$.

Sabemos, do teorema de Bayes, que

$$\begin{aligned} \pi(\theta|\mathbf{y}) &\propto l(\theta; \mathbf{y})\pi(\theta) \\ &\propto e^{-n\theta}\theta^t \cdot \theta^{\alpha-1}e^{-\beta\theta} \\ &= e^{-(n+\beta)\theta}\theta^{\alpha+t-1} \end{aligned}$$

que, a menos de uma constante multiplicativa, é o núcleo de uma densidade $Gama(\alpha_1, \beta_1)$, com: $\alpha_1 = \alpha + t, \quad \beta_1 = \beta + n$. Então $(\theta|\mathbf{y}) \sim Gama(\alpha_1, \beta_1)$.

Conjugação Poisson-Gama: Exemplo

- Y é variável que registra a contagem de cancelamentos de contaratos em uma empresa, por semana. Assuma $Y \sim Poisson(\theta)$.
- Há interesse em fazer inferência sobre θ: número esperado de cancelamentos, por semana.
- Assuma, a priori: $\theta \sim Gama(\alpha, \beta)$.

Conjugação Poisson-Gama: Exemplo

- Informação a priori(subjetiva), baseada na experiência já adquirida por um analista, com produto semelhante:

  - Perguntado sobre qual valor esperaria para o número médio de cancelamentos por semana: $E[\theta] = \alpha/\beta = 5 \Rightarrow \alpha = 5\beta$.

  - Perguntado sobre a probabilidade de que o número médio de cancelamentos por semana, θ, seja superior a 10: “Não espero que essa probabilidade seja superior a 10%”.

- Com base nas informações fornecidas pelo analista, foi possível especificar um modelo a priori: $\theta \sim Gama(2, 0.4)$

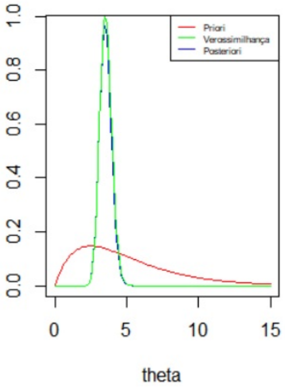
Conjugação Poisson-Gama: Exemplo

- Ao se observar os números de cancelamentos em 20 mercados aleatoriamente selecionados, ao longo de uma semana, registrou-se uma média $\bar{y} = 3,5$ cancelamentos, o que equivale a $t = \sum_{i=1}^{20} y_i = 70$ cancelamentos.

- Combinando-se a informação subjetiva elicitada com base na experiência do analista à informação obtida através dos dados coletados, chega-se à distribuição a posteriori: $(\theta|y_1, \ldots, y_{20}) \sim Gama(2 + 70, 0.4 + 20) \equiv Gama(72, 20.4)$

Conjugação Poisson-Gama: Exemplo

- Com base na distribuição a posteriori, atualiza-se a crença do analista: $P(\theta > 10|t = 70) \approx 0$.



Conjugação Poisson-Gamma: Exercício para fixação de conteúdo

Exercício: Refaça a análise admitindo a mesma distribuição a priori, mas assuma que tenha sido observada uma amostra menor, composta por 3 mercados, nos uais foi observado um total de 24 cancelamentos.

- Primeiramente, compare as curvas de distribuição a priori, verossimilhança e posteriori. Comente.
- Em seguida, explore a distribuição a posteriori, fazendo afrmações inferenciais sobre θ, a partir desta.
- Compare seus resultados àqueles obtidos com base na amostra de 20 segurados.

Conjugação: Exponencial-Gama

Exponencial-Gama

Se $Y_1, \dots, Y_n \overset{iid}{\sim} Exponencial(\theta)$ e assume-se, a priori, $\theta \sim Gama(\alpha, \beta)$, então, a posteriori, tem-se
$(\theta | y_1, \dots, y_n) \sim Gama(\alpha + n, \beta + t), \quad t = \sum_{i=1}^{n} y_i$.

Sabemos, do teorema de Bayes, que

$$\begin{aligned} \pi(\theta|\mathbf{y}) &\propto l(\theta; \mathbf{y})\pi(\theta) \\ &\propto e^{\theta t}\theta^{n} \cdot \theta^{\alpha-1}e^{-\beta\theta} \\ &= e^{-(t+\beta)\theta}\theta^{\alpha+n-1} \end{aligned}$$

que, a menos de uma constante multiplicativa, é o núcleo de uma densidade $Gama(\alpha_1, \beta_1)$, com: $\alpha_1 = \alpha + n \quad \beta_1 = \beta + t$. Então $(\theta|\mathbf{y}) \sim Gama(\alpha_1, \beta_1)$.

Conjugação: Normal-Normal

Normal-Normal

Se $Y_1, \ldots, Y_n \overset{iid}{\sim} Normal(\theta, \sigma^2)$, com σ^2 conhecido e assume-se, a priori, $\theta \sim Normal(\mu_0, \tau_0^2)$, então, a posteriori, tem-se $(\theta | y_1, \ldots, y_n) \sim Normal(\mu_1, \tau_1^2)$, com:

$$\mu_1 = \omega\mu_0 + (1-\omega)\bar{y}, \quad \tau_1^{-2} = \tau_0^{-2} + n\sigma^{-2}, \quad \omega = \tau_0^{-2}/(\tau_0^{-2} + n\sigma^{-2}).$$

Conjugação: Normal/Normal-Gama

Normal/Normal-Gama

Se $Y_1, \ldots, Y_n \overset{iid}{\sim} Normal(\theta, \sigma^2)$, com $\sigma^2 = 1/\phi$ desconhecido e assume-se, a priori, $\theta|\phi \sim Normal(\mu_0, (c_0\phi)^{-1})$ e $n_0\sigma_0^2\phi \sim \chi^2_{n_0}$, então:

	Priori	Posteriori
$\theta \mid \phi$	$N(\mu_0, (c_0\phi)^{-1})$	$N(\mu_1, (c_1\phi)^{-1})$
ϕ	$n_0\sigma_0^2\phi \sim \chi^2_{n_0}$	$n_1\sigma_1^2\phi \sim \chi^2_{n_1}$
θ	$t_{n_0}(\mu_0, \sigma_0^2/c_0)$	$t_{n_1}(\mu_1, \sigma_1^2/c_1)$
$\phi \mid \theta$	$[n_0\sigma_0^2 + c_0(\theta - \mu_0)^2)]\phi \sim \chi^2_{n_0+1}$	$[n_1\sigma_1^2 + c_1(\theta - \mu_1)^2)]\phi \sim \chi^2_{n_1+1}$

em que:

$\mu_1 = (c_0\mu_0 + n\bar{y})/(c_0 + n)$; $\quad c_1 = c_0 + n$; $\quad n_1 = n_0 + n$; $\quad n_1\sigma_1^2 = n_0\sigma_0^2 + ns^2 + [c_0 n/(c_0 + n)](\mu_0 - \bar{y})^2$, $\quad s^2 = \sum(y_i - \bar{y})^2/n$.

Conteúdo

Estimação Pontual Bayesiana

Do ponto de vista Bayesiano, a estimação pontual de um parâmetro θ pode ser encarada como um problema de teoria da decisão, que envolve os seguintes elementos:

- θ: quantidade desconhecida que se deseja estimar. Denotaremos por Θ o espaço paramétrico, isto é, o conjunto de todos os possíveis valores de θ;

- $\mathcal{A}$: espaço de ações. Cada estimativa pontual de θ, $a \in \mathcal{A}$, é uma ação;

- $L(\theta, a)$: função de perda, que especifica a perda associada a uma estimativa a, quando o verdadeiro valor do parâmetro é θ.

Estimação Pontual Bayesiana

- Objetiva-se encontrar a estimativa a^* ótima. Há diferentes critérios de optimalidade. Adotaremoso critério de minimização da perda esperada com respeito à distribuição a posteriori de θ

- a^* que minimize:
$$r(a) = E[L(\theta, a)|\mathbf{y}].$$

- O valor mínimo de $r(a)$, $r(a^*)$, é denominado risco de Bayes.

- O valor ótimo para a, a^*, é denominado estimador de Bayes e será uma função do vetor amostral $(Y_1, \ldots, Y_n)$ e dos parâmetros da distribuição a priori.

Perda Quadrática

Se $L(\theta, a) = (\theta - a)^2$ o estimador de Bayes, $a*$, é dado pela média a posteriori de θ, isto é:

$$a^* = E[\theta|\mathbf{Y}].$$

Ainda, nesse caso, o risco de Bayes é dado pela variância a posteriori :

$$r(a^*) = V[\theta|\mathbf{Y}].$$

Exemplo: Perda Quadrática

Volte ao exemplo de definição da fatia de mercado de um produto.

- Priori do analista 3: $\theta \sim Beta(1,2;8) \Rightarrow E[\theta] = 0,13$, com coeficiente de variação bastante alto (244%), refletindo grande incerteza a priori.
- Informação amostral: $n = 100$ pessoas foram entrevistas, $t = \sum y_i = 15$ comprariam o produto.
- Posteriori: $(\theta|\mathbf{y}) \sim Beta(16,2;93)$
- Estimativa de Bayes, com respeito à perda quadrática: o valor esperado de uma distribuição Beta com parâmetros $16,2$ e 93 é $E[\theta|\mathbf{y}] = \frac{16{,}2}{16{,}2+93} \approx 0,148$, próxima à estimativa de máxima verossimilhança, $\hat{\theta} = t/n = 15/100 = 0,15$.

Exemplo: Perda Quadrática

- Assuma que se observasse uma amostra de tamanho 10, com $t = 5$ respostas afirmativas

- Posteriori: $(\theta|\mathbf{y}) \sim Beta(6,2;13)$.

- Estimativa de Bayes, com respeito à perda quadrática: $E[\theta|\mathbf{y}] = \frac{6{,}2}{6{,}2+13} \approx 0,32$, bem mais distante da estimativa de máxima verossimilhança que, nesse caso, seria $\hat{\theta} = t/n = 5/10 = 0,5$, mostrando a influência do modelo a priori sobre a estimativa de Bayes, já que a amostra é pequena.

Perda Linear

Se

$$L(\theta, a) = \begin{cases} k_0(\theta - a), & \theta - a \geq 0, \\ k_1(a - \theta), & \theta - a < 0, \end{cases},$$

então o estimador de Bayes é dado pelo quantil $\frac{k_0}{k_0+k_1}$ da distribuição a posteriori de θ. Observe que a escolha de $k_0 > k_1$ penaliza mais severamente subestimativas que superestimativas de θ. Se $k_0 = k_1$, a função de perda resultante equivale à perda absoluta:

$$L(\theta, a) = |\theta - a|.$$

Sob perda absoluta, o estimador de Bayes é dado pelo quantil $1/2$ da distribuição a posteriori, ou seja, a mediana.

Exemplo: Perda Absoluta

- No exemplo anterior, considerou-se, a priori, $\theta \sim Beta(1,2;8)$
- Poesteriori, ao se tomar uma amostra de 100 indivíduos, com 15 respostas favoráveis: $Beta((16,2;93)$.
- Sob perda absoluta, teríamos a mediana da posteriori como estimador de Bayes, que pode ser obtida aplicando-se o comando R:
  qbeta(0.5, 16.2,93),
  resultando em $0,278$, inferior à estimativa sob perda quadrática, já que a posteriori, nesse caso, é assimétrica positiva.

Perda Binária

Perda Binária

Se

$$L(\theta, a) = \begin{cases} 0, & a = \theta, \\ 1, & a \neq \theta, \end{cases}$$

então o estimador de Bayes é dado pela moda da posteriori de θ.

Exemplo: Perda Binária

- No exemplo anterior, considerou-se, a priori, $\theta \sim Beta(1,2;8)$.
- Ao se tomar uma amostra de 100 indivíduos, com 15 respostas favoráveis, a distribuição a posteriori foi uma $Beta((16,2;93)$.
- Sob perda binária, teríamos a moda da posteriori como estimador de Bayes, que pode ser obtida aplicando-se os comando R:
  library(modeest)
  mlv("beta",16.2,93)

  Estimativa de Bayes: $0,142$, bastante próxima à estimativa sob perdas quadrática e absoluta, dado o comportamento simétrico dessa distribuição a posteriori

Exemplo: Perda Binária

Ao se considerar a amostra de tamanho $n = 10$, com 5 respostas favoráveis, a distribuição a posteriori obtida foi uma

$$Beta(6,2;13).$$

Portanto, a estimativa de Bayes, sob perda binária, seria:
mlv("beta",6.2,13)
Estimativa de Bayes: $0,302$.

Conteúdo

Estimação Intervalar

Questões
a pensar:

Questões
a pensar:

Estimativa veio
de uma amostra.

Questões
a pensar:

Estimativa veio
de uma amostra.

Se a amostra
fosse outra?

Estimação Intervalar

Questões
a pensar:

Se a amostra fosse
maior (ou menor)?

Estimativa veio
de uma amostra.

Se a amostra
fosse outra?

Estimação Intervalar

Toda a posteriori
reduzida a um ponto?

Questões
a pensar:

Se a amostra fosse
maior (ou menor)?

Estimativa veio
de uma amostra.

Se a amostra
fosse outra?

E a incerteza? (Qtdd de informação amostral)

- Uma resposta "Sim" entre 10 entrevistados: uma estimativa pontual clássica da proporção populacional θ de "Sins" é $\hat{\theta} = 1/10 = 0,1$.

- Dez respostas"Sim" entre 100 entrevistados $\Rightarrow \hat{\theta} = 10/100 = 0,1$.

- 100 respostas"Sim" entre 1000 entrevistados: $\Rightarrow \hat{\theta} = 100/1000 = 0,1$.

A incerteza é a mesma?

E a variação amostral?

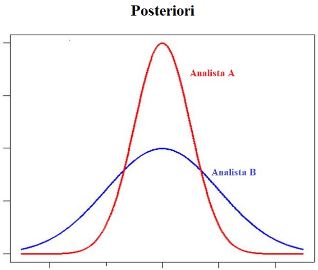
- Amostra aleatória de n indivíduos média amostral de indenizações $\bar{x}_1$ é uma estimativa pontual da indenização média populacional, μ.

- Outra amostra aleatória de n indivíduos, usando o mesmo método, de uma mesma população$\Rightarrow$ média amostral de indenizações $\bar{x}_2$, outra estimativa pontual da mesma indenização média populacional, μ.

Várias outras estimativas pontuais poderiam ter sido obtidas.

E a precisão da informação de outras fontes?

- Analista A tem mais experiência sobre o processo.
- Analista B tem menos experiência sobre o processo.
- A e B observam os mesmos dados - amostra pequena (mas cada um tem sua informação prévia aos dados - embora priori seja centrada em um mesmo ponto, incertezas distintas)



Estimação Intervalar - Objetivo

Produzir estimativas que levem em consideraçãoa incerteza associada ao processo de estimação.

A caracterização da incerteza depende do paradigma adotado, Clássico ou Bayesiano.

Conteúdo

Estimação Intervalar Bayesiana

- Dada a distribuição a posteriori de θ, o problema de estimação intervalar é naturalmente acomodado.
- A distribuição a posteriori de um parâmetro θ reflete a incerteza do analista sobre θ, atualizada à luz dos dados observados $\mathbf{y}$.
- Um intervalo de valores que sirvam como estimativas "razoáveis" para θ: basta tomar um intervalo delimitado por quantis l e u que, com razoável probabilidade, contenham θ.

Tomar, na distribuição a posteriori de θ, quantis l e u tais que

$$P(l \leq \theta \leq u|\mathbf{y}) = \gamma, \tag{4}$$

$0 < \gamma < 1$, γ representando o nível de credibilidade associado à estimativa intervalar.

Estimação Intervalar Bayesiana - Escolha dos limites

- Há inúmeras escolhas possíveis de valores (l, u).

- Idealmente, deseja-se escolher o intervalo de menor amplitude possível.

- Se a posteriori é simétrica e unimodal, o intervalo de menor amplitude, satisfazendo ao nível de credibilidade γ, é o intervalo centrado na modaa posteriori.

- Se a distribuição a posteriori não for simétrica, uma escolha usual é tomar:

$$l \text{ tal que } P(\theta < l|\mathbf{y}) = \alpha/2,$$

$$u \text{ tal que } P(\theta \leq u|\mathbf{y}) = 1 - \alpha/2,$$

com $\alpha = 1 - \gamma$.

Estimação Intervalar Bayesiana - Ilustração

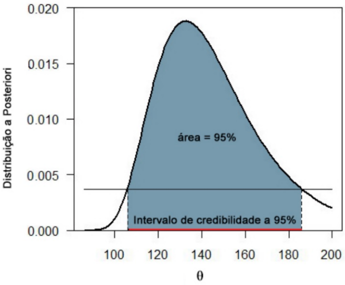


Figura 8: Intervalo de credibilidade.

O conceito de intervalo de credibilidade estende-se para regiões de credibilidade, quando θ é um vetor paramétrico de dimensão p e, portanto, sua distribuição a posteriori é um modelo p-dimensional.

Estimação Intervalar Bayesiana - Exemplo

- $t = 15$ opiniões favoráveis em $n = 100$ entrevistados
- priori (crenças iniciais do analista 3): $\theta \sim Beta(1,2;8)$
- posteriori: $(\theta|\mathbf{y}) \sim Beta(16,2;93)$.
- Estimativa intervalar para θ, ao nível de credibilidade 95%, e refletindo as crenças iniciais do analista 3 pode ser obtida tomando-se os quantis $2,5\%$ e $97,5\%$ da distribuição a priori, dados, no R, respectivamente, por:

Código R

```
qbeta(0.025,1.2,8) qbeta(0.975,1.2,8)
```

e retornando a estimativa intervalar $(0,006;0,397)$.

Estimação Intervalar Bayesiana - Exemplo

Ou seja, antes da observação dos dados, o analista acreditava que, com 95% de probabilidade, a proporção de pessoas que comprariam o produto está contida no intervalo $(0,006; 0,397)$.

Após a observação da amostra, a estimativa intervalar ao nível de credibilidade 95% é dada por quantis adequados na distribuição a posteriori:

Código R

```
qbeta(0.025,16.2,93) qbeta(0.975,16.2,93)
```

que fornece a estimativa intervalar $(0,088; 0,220)$.

Ou seja, após a observação dos dados, o analista acredita que, com 95% de probabilidade, a proporção de pessoas que comprariam o produto seja algo entre $8,8\%$ e 22%.

Estimação Intervalar Bayesiana - Exemplo

- intervalo de credibilidade a priori, ao nível de credibilidade 95%: $(0,006; 0,397)$.

- intervalo de credibilidade a posteriori, ao nível de credibilidade 95%:$(0,088; 0,220)$.

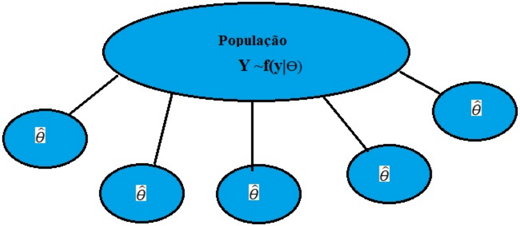
- A amplitude do intervalo de credibilidade a posteriori é menor que a do intervalo a priori, ao mesmo nível de credibilidade, refletindo o ganho de informação (e redução de incerteza), ao se passar do modelo a priori (anterior à observação dos dados) ao modelo a posteriori.

Conteúdo

Estimação Intervalar Clássica

- Enfoque Clássico não admite tratamento probabilístico associado a parâmetros.

  A estimação intervalar clássica busca refletir a variabilidade que um estimador sofreria, caso fosse observado em diferentes amostras aleatórias de mesmo tamanho.

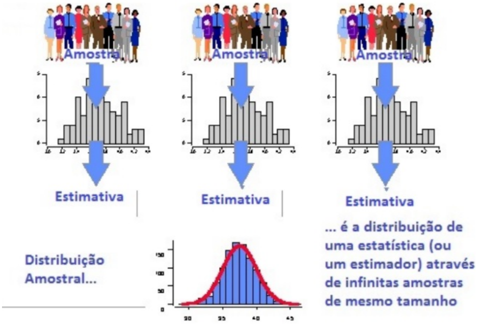


Estimação Intervalar Clássica

- Do ponto de vista clássico, um estimador é uma função somente de informação amostral.
- Estimadores clássicos são funções, apenas, de $\mathbf{Y} = (Y_1, \ldots, Y_n) \overset{iid}{\sim} f(y|\theta)$
- Estimador é uma variável aleatória e, assim, tem uma distribuição de probabilidade, denominada distribuição amostral.

Estimação Intervalar Clássica

Distribuição amostral reflete a variação teórica que o estimador sofreria, se observado em inúmeras amostras aleatórias de mesmo tamanho, oriundas de uma mesma população (de fato, de um mesmo modelo observacional $f(y|\theta)$).



Estimação Intervalar Clássica

- Na prática trabalha-se com uma única amostra e adistribuição amostral é modelo teórico para a variação que o estimador sofreria.

- Dado um valor para θ, poderíamos utilizar tal distribuição, por exemplo, para determinar a probabilidade de que um estimador produza estimativas em um determinado intervalo (mas θ é desconhecido - podemos apenas fixar cenários hipotéticos)

- Bastante usual utilizarmos tais distribuições para para determinar o tamanho amostral necessário para que, com razoável probabilidade, não tenhamos estimativas distantes do "verdadeiro" valor de θ.

- Além dessas aplicações, a partir das distribuições amostrais de estimadores, podemos construir intervalos de confiança (estimativas intervalares clássicas).

Distribuição da média amostral - População Normal, variância conhecida

Admita o modelo observacional $Y \sim N(\theta, \sigma^2)$. Por ora, assuma σ^2 conhecido

Interesse na estimação da média populacional (ou teórica) θ.

Amostra aleatória $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\theta, \sigma^2)$.

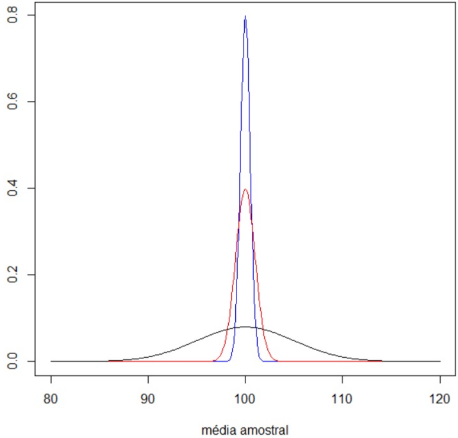
Um estimador usual para θ, por suas boas propriedades (não tendencioso, consistentente) é $\bar{Y}_n = \sum_{i=1}^{n} Y_i / n$.

Pode-se mostrar que a distribuição amostral de $\hat{\theta} = \bar{Y}_n$ é:

$$\bar{Y}_n \sim N(\theta, \sigma^2/n);$$

Distribuição da média amostral - População Normal, variância conhecida

Para qualquer que seja θ, se Y segue distribuição normal, as médias de diferentes amostras aleatórias de tamanho n, oscilariam simetricamente em torno de θ, com variabilidade que depende diretamente da incerteza intrínseca associada a Y e a suas mensurações (σ^2) e depende inversamente do tamanho das amostras coletadas (n).



Distribuição da média amostral - Simulação empírica

Podemos observar empiricamente o comportamento do estimador média amostral, utilizando o seguinte código R, que gera médias amostrais de 200 amostras aleatórias de tamanho 10 da distribuição $N(100, 25)$:

Código R

Gerando 200*n valores aleatórios com distribuição N(100,25)

```
n=10
vetorsim< −rnorm(200*n,100,sqrt(25))
```

Alocando os valores gerados em uma matriz com n linhas e 200 colunas:

```
mat.amostras¡-matrix(vetorsim,ncol=200,byrow=F)
```

Tomando a média de cada uma das 200 amostras de tamanho n:

```
vetor.medias< −apply(mat.amostras,2,mean)
```

Distribuição empírica das médias amostrais:

```
hist(vetor.medias)
```

Distribuição da média amostral - Simulação empírica

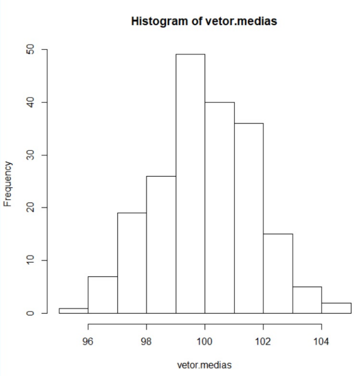


Figura 9: Distribuição empírica da média amostral. Amostras geradas de população $N(100, 25)$, para $n = 10$

Distribuição da média amostral - Simulação empírica

Vejamos o comportamento empírico da distribuição de $\bar{Y}_n$ ao aumentarmos o tamanho amostral para $n = 100$:

Código R

Gerando 200*n valores aleatórios com distribuição N(100,25)
n=100
vetorsim< −rnorm(200*n,100,sqrt(25))
Alocando os valores gerados em uma matriz com n linhas e 200 colunas:
mat.amostras¡-matrix(vetorsim,ncol=200,byrow=F)
Tomando a média de cada uma das 200 amostras de tamanho n:
vetor.medias¡-apply(mat.amostras,2,mean)
Distribuição empírica das médias amostrais:
hist(vetor.medias)

Distribuição da média amostral - Simulação empírica

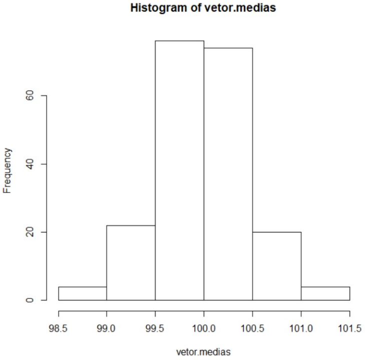


Figura 10: Distribuição empírica da média amostral. Amostras geradas de população $N(100, 25)$, para $n = 100$

.

Distribuição amostral da média: cálculo de tamanho amostral

Y : Certo tipo de desembolso feito usualmente por uma companhia.

Admita $Y \sim N(\theta; 1,2^2)$ (centenas de unidades monetárias).

Objetivo: estimar o desembolso médio θ.

Desejamos dimensionar o tamanho da amostra aleatória a ser coletada, de tal forma que a margem de erro absoluta, $|\bar{Y}_n - \theta|$, ao se usar a média amostral como estimador do desembolso médio θ, não ultrapasse $0,1$ (centena de unidades monetárias), com probabilidade $0,95$.

Distribuição amostral da média: cálculo de tamanho amostral

Ou seja, deseja-se obter n tal que

$$\begin{aligned} P(|\bar{Y}_n - \theta| \leq 0,1) &= 0,95 \\ P(-0,1 \leq \bar{Y}_n - \theta \leq 0,1) &= 0,95 \\ P(-\frac{0,1\sqrt{n}}{1,2} \leq \frac{\bar{Y}_n - \theta}{\sigma/\sqrt{n}} \leq \frac{0,1\sqrt{n}}{1,2}) &= 0,95 \\ \Leftrightarrow \frac{\sqrt{n}}{1,2} &= \frac{1,96}{0,1} \Leftrightarrow n \approx 267. \end{aligned}$$

Obs: Para se obter o quantil $1,96$ da curva Normal-padrão, utlizou-se o comando $qnorm(0.975, 0, 1)$ no R.

Estimação clássica intervalar da média de população Normal - σ^2 conhecida

Como já vimos,

$$\bar{Y}_n \sim N(\theta, \sigma^2/n) \Rightarrow \frac{\bar{Y}_n - \theta}{\sqrt{\sigma^2/n}} \sim N(0,1).$$

Sendo assim, se tomarmos $0 < \gamma < 1$ e $\alpha = 1 - \gamma$:

$$\begin{aligned} P\left(\Phi^{-1}(\alpha/2) \leq \frac{\bar{Y}_n - \theta}{\sigma/\sqrt{n}} \leq \Phi^{-1}(1-\alpha/2)\right) &= 1-\alpha \\ \Leftrightarrow P(\bar{Y}_n - \Phi^{-1}(1-\alpha/2)\sigma/\sqrt{n} \leq \theta \leq \bar{Y}_n + \Phi^{-1}(1-\alpha/2)\sigma/\sqrt{n}) &= \gamma, \quad (5) \end{aligned}$$

$\Phi^{-1}(1-\alpha/2)$ denota o percentil $100(1-\alpha/2)\%$ da distribuição $N(0,1)$.

Interpretação do nível de confiança

Após observar a amostra, γ é denominado nível de confiança e sua interpretação é frequentista: esperaríamos que aproximadamente $100\gamma\%$ dos intervalos de confiança com limites

$$(\bar{y}_n - \Phi^{-1}(1-\alpha/2)\sigma/\sqrt{n}; \quad \bar{y}_n + \Phi^{-1}(1-\alpha/2)\sigma/\sqrt{n}) \tag{6}$$

contivessem o valor real de θ, se inúmeras diferentes amostras fossem tomadas, cada uma produzindo um valor $\bar{y}_n$ e, portanto, um intervalo de confiança ao nível γ.

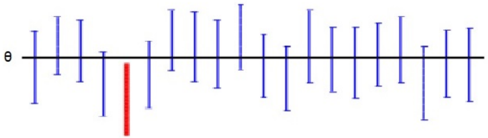


Figura 11: Diferentes intervalos de confiança para θ, obtidos com base em amostras de mesmo tamanho n. Cada intervalo é centrado na estimativa pontual $\bar{y}_n$.

Exemplo

Semanalmente, uma loja de departamentos seleciona uma amostra aleatória de 25 clientes para saber qual quantia Y gastam, em cada ida às compras.

Com base em dados históricos, assuma-se $\sigma = US\$20$.

Ainda: $Y \sim N(\theta, \sigma^2)$.

Na última semana, a pesquisa entre os 25 clientes selecionados resultou em um gasto médio $\bar{y}_{25} = US\$82,00$. Deseja-se obter um intervalo ao nível de confiança 95% para o gasto médio θ.

Exemplo (cont.)

O quantil $\Phi^{-1}(1-\alpha/2)$ da curva normal-padrão são obtidos usando os comandos R:
gamma=0.95
alpha=1-gamma
qnorm(1-alpha/2)
retornando um valor $\Phi^{-1}(1-\alpha/2) \approx 1,96$. O intervalo de confiança é, então, dado por:

$$\left(82 - 1,96\frac{20}{\sqrt{25}}; \quad 82 + 1,96\frac{20}{\sqrt{25}}\right) \equiv (74,16; \quad 89,84),$$

que é uma estimativa intervalar clássica, ao nível de confiança 95% e com base em uma amostra aleatória de tamanho $n = 25$, para o gasto médio por cliente da loja.

Exemplo (cont.)

Poderíamos contruir uma estimativa intervalar bayesiana com base nos mesmos dados.

Admita quetenhamos uma distribuição a priori bastante vaga sobre θ, algo como $\theta \sim N(50, 10000)$.

Sob tais condições teríamos distribuição a posteriori normal para θ, com precisão:

$$\tau_1^{-2} = \tau_0^{-2} + n\sigma^{-2} = \frac{1}{10000} + 25 \times \frac{1}{400} = 0,0626,$$

portanto, a variância a posteriori é $\frac{1}{0,0626} = 15,97444$.

Exemplo (cont.)

A média a posteriori é uma ponderação entre média a priori e média observada. Seja

$$\omega = \tau_0^{-2}/(\tau_0^{-2} + n\sigma^{-2}) = 0,0001/(0,0001 + 25 \times 400) = 10^{-8}.$$

Então, a média a posteriori é

$$\mu_1 = \omega\mu_0 + (1-\omega)\bar{y} = 10^{-8} \times 50 + (1 - 10^{-8}) \times 82 = 82.$$

Tomando os quantis $2,5\%$ e $97,5\%$ da distribuição $N(82; 15,97)$ temos:
gamma=0.95
alpha=1-gamma
qnorm(alpha/2,82,sqrt(15.97))
qnorm(1-alpha/2,82,sqrt(15.97))
resultando na estimativa intervalar $(74,17; \quad 89,83)$, muito próxima à estimativa intervalar clássica, uma vez que adotamos distribuição a prori vaga.

Cálculo do tamanho da amostra

No exemplo, a margem de erro do intervalo de confiança clássico foi de $7,84$ dólares.

Poderíamos, por exemplo, desejar determinar o tamanho da amostra a ser coletada, de forma a produzir intervalos com nível de confiança $\gamma = 1 - \alpha$ para o gasto médio, respeitando uma margem de erro de US\$ ϵ.

Seja $z = \Phi^{-1}(1 - \alpha/2)$. Basta tomar:

$$\begin{aligned} z\sigma/\sqrt{n} &= \epsilon \\ \Leftrightarrow n &= \left(\frac{z\sigma}{\epsilon}\right)^2. \end{aligned} \tag{7}$$

Exemplo

Se desejarmos aumentar o nível de confiança, no exemplo anterior, para 99%, teremos o quantil $\Phi^{-1}(1-\alpha/2)$ da curva normal-padrão dado pelos comandos R:

gamma=0.99
alpha=1-gamma
qnorm(1-alpha/2)
que retornam um valor $\Phi^{-1}(1-\alpha/2) \approx 2,58$, elevando a margem de erro para $\Phi^{-1}(1-\alpha/2)\sigma/\sqrt{n} \approx 10,32$.

Se desejarmos preservar o nível de confiança $\gamma = 0,99$, associado a uma margem de erro de $\epsilon = US\$2$, deveríamos ter:

$$n = \left(\frac{2,58 \cdot 20}{2}\right)^2 \approx 664.$$

Distribuição da média amostral - população Normal - σ^2 desconhecida

Situação prática real: além da média, a variância teórica σ^2 é desconhecida.

Se $Y_1, \dots, Y_n \overset{iid}{\sim} N(\theta, \sigma^2)$, então

$$\bar{Y}_n \sim N(\theta, \sigma^2/n) \Rightarrow Z = \frac{\bar{Y}_n - \theta}{\sqrt{\sigma^2}/n} \sim N(0, 1).$$

Quando o valor de σ^2 não é conhecido, substituiremos σ^2 por seu estimador não tendencioso:

$$\hat{\sigma}^2 = s^2 = \sum_{i=1}^{n} \frac{(Y_i - \bar{Y}_n)^2}{n-1}$$

e, ao invés de operarmos com a estatística Z, teremos a estatística:

$$T = \frac{\bar{Y}_n - \theta}{\sqrt{S^2}/n}.$$

Distribuição t de Student

A estatística T está sujeita à incerteza envolvida em duas variáveis aleatórias $\bar{Y}_n$ e S^2. Portanto, sua incerteza é maior que a de Z e T não terá distribuição normal-padrão.

Pode-se mostrar que T segue uma distribuição t de Student com $(n-1)$ graus de liberdade.

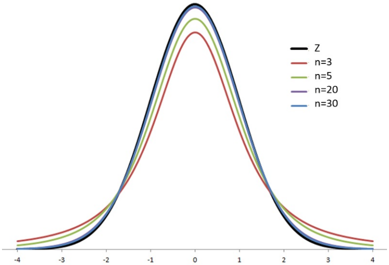


Figura 12: Distribuição t, comparada à Normal padrão.

Intervalo de Confiança para a média - População Normal - σ^2 desconhecida

A expressão do intervalo de confiança para a média θ do modelo normal passa a ser dada por:

$$(\bar{y}_n - T^{-1}(1-\alpha/2)\sqrt{s^2/n}; \quad \bar{y}_n + T^{-1}(1-\alpha/2)\sqrt{s^2/n}), \tag{8}$$

em que em que $T^{-1}(1-\alpha/2)$ denota o percentil $100(1-\alpha/2)\%$ da distribuição t de Student com $n-1$ graus de liberdade.

Exemplo

Voltemos ao exemplo da estimação da quantia média gasta por cliente em um estabelecimento.

Admita, como antes, que $Y \sim Normal(\theta, \sigma^2)$, mas considere agora a situação mais realista em que tanto θ quanto σ^2 são desconhecidos.

Suponha que a amostra aleatória de 25 clientes tenha fornecido um gasto médio de $\bar{y}_{25} = US\$82,00$, com desvio-padrão (amostral) $s = US\$20$.

Vejamos como fica, agora, o intervalo de confiança ao nível 95% para θ.

Exemplo (cont.)

O quantil $T^{-1}(1-\alpha/2)$, neste exemplo, é dado pelos comandos R:

n=25
gamma=0.95
alpha=1-gamma
qt(1-alpha/2,n-1)

retornando o valor aproximado 2.064.

O intervalo ao nível de confiança 95% paara o gasto médio θ, por cliente, é dado por:

$$\left(82 - 2,064\frac{20}{\sqrt{25}}; \quad 82 + 2,064\frac{20}{\sqrt{25}}\right) \equiv (73,74; \quad 90,26).$$

Este intervalo tem aplitude maior, se comparado ao intervalo ao nível 95% obtido admitindo-se $\sigma = 20$ conhecido, pois incorpora (devidamente) a incerteza adicional decorrente da estimação de σ.

Distribuição assintótica da média amostral

Teorema Central do Limite - TCL:

Se $Y_1, \ldots, Y_n$ são variáveis aleatórias independentes e identicamente distribuídas, com média θ e variância σ^2 finitas, então:

$$Z = \frac{\bar{Y}_n - \theta}{\sqrt{\sigma^2/n}} \overset{a}{\sim} N(0,1).$$

Intervalo de Confiança para a proporção

$Y \sim Bernoulli(\theta)$, em que $Y = 1$ denota a presença e $Y = 0$, a ausência da característica de interesse.

Então que $E[Y] = \theta$, $V[Y] = \theta(1-\theta)$.

Se tomarmos uma amostra aleatória de indivíduos, avaliando para cada um a presença ("sucesso") ou ausência ("fracasso") da característica de interesse, teremos $Y_1, \ldots, Y_n \overset{iid}{\sim} Bernoulli(\theta)$.

Um estimador pontual não tendencioso e consistente para θ é dado pela proporção amostral de sucessos:

$$\hat{\theta} = \sum_{i=1}^{n} \frac{Y_i}{n}. \tag{9}$$

Intervalo de Confiança para a proporção

Para tamanho amostral n suficientemente grande, então, de acordo com o TCL, $\hat{\theta}$ terá distribuição aproximadamente normal:

$$\hat{\theta} \overset{a}{\sim} N(E[Y], V[Y]/n), \tag{10}$$

em que $E[Y] = \theta$ e $V[Y] = \theta(1-\theta)$, concluindo-se, finalmente, que:

$$Z = \frac{\hat{\theta}_n - \theta}{\sqrt{\theta(1-\theta)}/n} \overset{a}{\sim} N(0,1). \tag{11}$$

Intervalo de Confiança para a proporção

Um intervalo aproximado, ao nível de confiança γ para a proporção populacional de sucessos θ pode ser obtido observando-se que, para n suficientemente grande:

$$P\left(\Phi^{-1}(\alpha/2) \leq \frac{\hat{\theta}-\theta}{\sqrt{\theta(1-\theta)/n}} \leq \Phi^{-1}(1-\alpha/2)\right)$$

$$= P(\hat{\theta}-\Phi^{-1}(1-\alpha/2)\sqrt{\theta(1-\theta)/n} \leq \theta \leq \hat{\theta}+\Phi^{-1}(1-\alpha/2)\sqrt{\theta(1-\theta)/n})$$

$$= 1-\alpha = \gamma$$

em que $\Phi^{-1}(1-\alpha/2)$ denota o percentil $100(1-\alpha/2)\%$ da distribuição $N(0,1)$.

Intervalo de Confiança para a proporção

Como os limites do intervalo acima dependem de θ, que é justamente o parâmetro que se deseja estimar, e como, aqui, admitimos amostras suficientementemente grandes, é usual que se substitua θ por $\hat{\theta}$ nos limites do intervalo, já que $\hat{\theta}$ é um estimador não tendencioso e consistente. Teremos, então, um intervalo aproximado, ao nível de confiança $\gamma = 1 - \alpha$, para θ, dado por:

$$\left(\hat{\theta} - \Phi^{-1}(1-\alpha/2)\sqrt{\hat{\theta}(1-\hat{\theta})/n}; \quad \hat{\theta} + \Phi^{-1}(1-\alpha/2)\sqrt{\hat{\theta}(1-\hat{\theta})/n}\right).$$

Exemplo

Admita que a proporção θ de ocorrências de certo evento gerador de indenizações, na população segurada, seja desconhecida e que haja interesse em estimar tal proporção.

Ao se observar uma amostra aleatória de $n = 200$ expostos a esse tipo de risco, verificou-se 10 ocorrências desse evento, resultando em uma estimativa pontual $\hat{\theta} = 10/200 = 0,05$.

Um intervalo de confiança aproximado para θ ao nível de confiançaa 99%, é dado por:

$$\left(\hat{\theta} - \Phi^{-1}(1-\alpha/2)\sqrt{\hat{\theta}(1-\hat{\theta})/n}; \quad \hat{\theta} + \Phi^{-1}(1-\alpha/2)\sqrt{\hat{\theta}(1-\hat{\theta})/n}\right)$$

$$\equiv (0,05 - 2,58\sqrt{0,05(0,95)/200}; \quad 0,05 + 2,58\sqrt{0,05(0,95)/200})$$

$$\equiv (0,01024; \quad 0,08976),$$

Exemplo (cont.)

O quantil 2,58 foi obtido usando-se os comandos R:
gamma=0.99
alpha=1-gamma
qnorm(1-alpha/2)

A leitura do intervalo acima é: tem-se 99% de confiança de que a proporção populacional do evento gerador de indenizações, θ, seja algo entre 1% e 9%.

Não é correto afirmar que θ está no intervalo $(0,01; 0,09)$ com *probabilidade* 99%, visto que não se admite tratamento probabilístico para parâmetros, no enfoque clássico.

Exemplo (cont.)

Podemos, alternativamente, obter um intervalo bayesiano, ao mesmo nível de credibilidade 99%, para θ.

Vimos que Se $Y_1, \ldots, Y_n \overset{iid}{\sim} Bernoulli(\theta)$ e assume-se, a priori, $\theta \sim Beta(\alpha, \beta)$, então, a posteriori, tem-se $(\theta | y_1, \ldots, y_n) \sim Beta(\alpha + t, \beta + n - t), \quad t = \sum_{i=1}^{n} y_i$.

Temos $n = 200$, observando-se $t = 10$ sucessos e $n - t = 190$ fracassos na amostra.

Exemplo (cont.)

Assuma que, a priori, não tenhamos conhecimento sobre o valor de θ. Podemos traduzir essa ausência de conhecimento adotando um modelo a priori $Beta(1,1) \equiv U(0,1)$, resultando em uma distribuição a posteriori para θ: $(\theta|\mathbf{y}) \sim Beta(11,191)$.

Um intervalo de credibilidade ao nível 99% para Beta é dado pelos quantis $0,05$ e $0,995$ da distribuição a posteriori, obtidos com os seguintes comandos R:

gamma=0.99
alpha=1-gamma
qbeta(alpha/2,11,191)
qbeta(1-alpha/2,11,191)

Estimativa intervalar bayesiana:$(0,02;0,10)$, bem próxima à estimativa clássica, dada a priori vaga utilizada. Para o intervalo Bayesiano, é correta a leitura: a probabilidade, a posteriori, de que θ esteja entre 2% e 9% é 99%.

endframe

Conteúdo

Conteúdo

Previsão Bayesiana

Observarmos uma amostra de valores $\mathbf{y} = (y_1, \ldots, y_n)$.

Assumimos que tais valores tenham sido originados a partir de um modelo probabilístico $f(y|\theta)$.

A partir dos dados observados, aprendemos sobre θ. Atualizamos nossas crenças sobre o parâmetro (ou vetor paramétrico) θ, passando:

- da distribuição a priori $\rightarrow \pi(\theta)$
- à distribuição a posteriori $\rightarrow \pi(\theta|\mathbf{y})$.

Previsão Bayesiana

Agora, admita que desejamos caracterizar o comportamento de uma observação futura y^f, dado o conhecimento de seus valores passados $\mathbf{y}$. Ou seja, desejamos obter um modelo preditivo:

$$\pi(y^f|\mathbf{y}) = \int f(y^f, \theta|\mathbf{y})d\theta = \int f(y^f|\theta)\pi(\theta|\mathbf{y})d\theta.$$

Como pode ser visto na equação acima, O modelo preditivo é a resultante de tomar uma "média" dos modelos observacionais $f(y|\theta)$, "passeando" por todo o espaço de valores possíveis para θ, com a posteriori $\pi(\theta|\mathbf{y})$ indicando a relevânca de cada ponto θ na composição desse modelo.

Previsão Bayesiana

Podemos aproximar a distribuição preditiva por amostragem. Para tomar uma amostra de k pontos da preditiva, repita, para $i = 1, \ldots, k$:

- gere $\theta^{(i)}$ da posteriori $\pi(\theta|\mathbf{y})$;
- gere $y^{(i)}$ do modelo observacional $f(y|\theta^{(i)})$.

Como é uma amostra de Monte Carlo, podemos:

- ter ideia do formato da distribuição preditiva: p. ex., através de histograma da amostra;
- obter estimativas para probabilidades: p. ex., $\hat{P}(Y^f > l|\mathbf{y}) = \frac{\#\{y^{(i)}:y^{(i)}>l\}}{k}$
- obter estimativas para o valor esperado de Y^f: $\hat{E}(Y^f) = \frac{\sum_{i=1}^{k} y^{(i)}}{k}$
- obter amostras de qualquer função $Z = g(Y^f)$: basta tomar $g(y^{(1)}, \ldots, g(y^{(k)}$
- etc

Universidade Federal do Rio de Janeiro

Instituto de Matemática
Universidade Federal do Rio de Janeiro

Previsão Bayesiana - Ex: Poisson-Gama

Veja o código R disponibilizado: *Preditiva_Poisson_Gama.R*

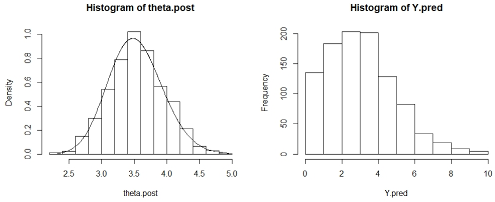


Figura 13: Esquerda: Distribuição a posteriori para θ. Direita: Distribuição preditiva para valor futuro da contagem de cancelamento de contratos.

O modelo preditivo leva em conta tanto a incerteza associada ao valor futuro de Y quanto a incerteza associada ao valor do parâmetro θ. Veja o código disponibilizado, para maiores detalhes.

Conteúdo

Previsão Clássica

Sob o paradigme clássico ou frequentista, é usual que o próprio modelo observacional $f(y|\theta)$ seja utilizado para fazer previsões.

Como o valor de θ é desconhecido, utiliza-se uma estimativa pontual, por exemplo a estimativa de máxima verossimilhança $\hat{\theta}$ e calculam-se quantidades de interesse para predição de comportamento futuro (quantis, probabilidades, valores esperados etc) com base em $f(y|\hat{\theta})$.

Questão: Essa forma de fazer previsão não leva em conta a incerteza associada a θ e trata a estimativa pontual como se fosse valor "verdadeiro " do parâmetro.

Conteúdo

Universidade Federal do Rio de Janeiro

Instituto de Matemática
Universidade Federal do Rio de Janeiro

Testes de Hipóteses

Assuma $Y \sim f(y|\theta)$.

Frequentemente, temos que decidir entre duas (contraditórias) afirmações envolvendo parâmetros de uma distribuição.

Uma hipótese estatística é qualquer afirmação feita sobre características da população de interesse, usualmente expressas em termos de valores especificados para θ ou alguma função de θ.

As hipóteses estatísticas envolvem características latentes, não observáveis.

Tipos de Hipóteses: Simples e Composta

Uma hipótese é dita simples se especifica um único valor para θ; caso contrário, a hipótese é dita composta.

Sob uma hipótese simples, a distribuição $f(y|\theta)$ fica completamente especificada.

(1) Seja Y a log-severidade de sinistros (em milhares de unidades monetárias) e admita que $Y \sim N(\mu, \sigma^2)$. Pode haver interesse nas hipóteses:

$$H_0 : \mu = 0 \quad vs \quad H_1 : \mu \neq 0$$

ou, ainda, em:

$$H_0 : \mu \leq 0 \quad vs \quad H_1 : \mu > 0.$$

(2) Seja Y a frequência de ocorrências de certo tipo de sinistro, em uma população de segurados. Admitra $Y \sim Poisson(\theta)$ e interesse em testar:

$$H_0 : \theta \leq 1/4 \quad vs \quad H_1 : \theta > 1/4.$$

Hipótese Nula e Alternativa

H_0, tomada como hipótese canônica é denominada hipótese nula.

H_1, que se opõe a H_0, é denominada hipótese alternativa.

É usual que H_0 seja favorecida, no sentido de que não seja rejeitada em favor da alternativa (H_1), a menos que exista forte evidência estatística na direção de H_1. Na ausência de tais evidências, H_0 não é rejeitada.

Estudos científicos tipicamente adotam uma abordagem conservativa e identificam a teoria ou prática corrente, estabelecida, como H_0 e a proposta alternativa do pesquisador como H_1.

Exemplo

Suponha que se saiba, a partir de dados históricos, que certo evento gerador de indenizações ocorra com probabilidade 10%.

Uma seguradora implementa uma política de redução do prêmio para segurados que adotem medidas de proteção contra esse tipo de evento.

Seja θ a proporção de ocorrências desse evento, após implantação da política de redução de prêmios. Deseja-se testar as seguintes hipóteses sobre θ:

$$H_0 : \theta = 0,10 \quad vs \quad H_1 : \theta < 0,10.$$

Aqui, temos uma hipótese nula simples, em contraposição a uma hipótese alternativa composta.

Problema Geral

O problema geral de teste de duas hipóteses pode ser expresso como:

$$H_0 : \theta \in \Theta_0 \quad vs \quad H_1 : \theta \in \Theta_1,$$

com $\Theta_0 \cup \Theta_1 = \Theta$

Θ é o espaço paramétrico, ou seja, conjunto de todos os possíveis valores de Θ em um dado problema.

Conteúdo

Testes de Hipóteses Clássicos

Devemos ter uma regra de decisão, para escolha entre H_0 e H_1.

Do ponto de vista clássico, tal regra é pautada somente pelo que se observa em uma amostra.

Utilizamos evidência amostral, resumida em estatísticas, para tomar decisões sobre quantidades populacionais (parâmetros) a que se referem as hipóteses.

Exemplo (cont.)

Suponha que examinemos 200 segurados, aleatoriamente selecionados, quanto à ocorrência ou não de sinistro após implantação das novas medidas.

Sejam $Y_1, \ldots, Y$ as observações na amostra ($y_i = 1$ se houve sinistro para o i-ésimo elemento; $y_i = 0$ em caso contrário).

A estatística $T = \sum_{i=1}^{200}$ representa o número de ocorrências de sinistros. Se fossem observadas $t = 19$ ocorrências do sinistro na amostra, você rejeitaria H_0?

Poderíamos adotar a seguinte regra de decisão arbitrária: "Rejeita-se H_0 se $t \leq 15$". Aqui, T é a estatística de teste e o conjunto $\{15, 14, \ldots, 1, 0\}$ é a região de rejeição.

Região Crítica

Uma estatística de teste é uma função da amostra aleatória $(Y_1, \ldots, Y_n)$ que deverá ser usada para decidirmos pela rejeição ou não de H_0.

A região de rejeição (ou região crítica) é o conjunto de valores da estatística de teste para os quais H_0 deve ser rejeitada.

Tipos de Erros

Ao se tomar uma decisão com base nos dados observados, pode-se estar cometendo um erro.

A decisão \ A hipótese nula	(V) VERDADEIRA	(F) FALSA
Rejeitar a Hipótese nula	Erro tipo I	acerto
Não rejeitar a Hipótese nula	acerto	Erro tipo II

Figura 14: Tipos de erros em um teste de hipóteses

O erro do tipo I consiste em rejeitar H_0, quando de fato H_0 ?é verdadeira.

O erro do tipo II consiste em não rejeitar H_0, quando de fato H_1 é verdadeira.

Tipos de Erros

- Procedimentos que conduzam a probabilidades de erro tipo I mais baixas levarão a probabilidades de erro tipo II mais altas.
- Tipicamente, considera-se o erro tipo I (rejeição indevida de H_0) como a situação mais grave.
- Nesse caso, buscam-se testes que atendam a uma probabilidade máxima de erro tipo I (julgada tolerável) e, se possível, minimizem probabilidades de erro tipo II, se comparados a outros testes.

Nível de Significância

O(s) valor(es) crítico(s) de uma regra de decisão é (são) o(s) limiar(es) que determina(m) a região crítica (ou de rejeição) do teste.

Uma vez fixada essa região, tem-se, decorrência desta, as probabilidades de erro tipo I e II implicadas:

$$\alpha(\theta_0) = P(\text{erro tipo I}|\theta_0) = P(\text{rejeitar} H_0|\theta = \theta_0), \quad \text{p/ cada } \theta_0 \in \Theta_0.$$

$$\beta(\theta_1) = P(\text{erro tipo II}|\theta_1) = P(\text{não rejeitar } H_0|\theta = \theta_1), \quad \text{p/ cada } \theta_1 \in \Theta_1.$$

O valor máximo de $\alpha(\theta_0)$, entre todos os valores θ_0 no domínio da hipótese nula é o tamanho do teste e tipicamente adotaremos nível de significância igual ou o mais próximo possível ao tamanho do teste.

Exemplo (cont.)

Prosseguindo com o exemplo anterior, tem-se interesse em testar:

$$H_0 : \theta = 0,10 \quad vs \quad H_1 : \theta < 0,10.$$

- Erro tipo I (rejeição indevida de H_0), significaria concluir que a nova política de prêmios trouxe redução na proporção de ocorrências do sinistro estudado, quando de fato a medida foi inócua.

Exemplo (cont.)

Prosseguindo com o exemplo anterior, tem-se interesse em testar:

$$H_0 : \theta = 0,10 \quad vs \quad H_1 : \theta < 0,10.$$

- Erro tipo I (rejeição indevida de H_0), significaria concluir que a nova política de prêmios trouxe redução na proporção de ocorrências do sinistro estudado, quando de fato a medida foi inócua.
- Erros tipo II ocorreriam se a medida tiver acarretado redução na proporção θ, mas os dados amostrais levarem à conclusão de que não houve alteração.

Exemplo (cont.)

Considere a regra de decisão (arbitrária) baseada na estatística de teste T: número de "sucessos" observados na amostra, com região crítica $\{15, 14, \ldots, 1, 0\}$.

Como só há um ponto no domínio da hipótese nula ($\theta_0 = 0,10$), este será o tamanho (e o nível de significância do teste), já que a probabilidade de erro tipo I é única. Observando que $\hat{\theta} = T/n$ e que, de acordo com o TCL, $\hat{\theta} \overset{a}{\sim} N(\theta, \theta(1-\theta)/n)$, temos:

$$
\begin{aligned}
\alpha &= P(\text{erro tipo I}) = P(\text{rejeitar } H_0|\theta = 0,10) = P(T \leq 15|\theta = 0,10) \\
&= P(\hat{\theta} \leq 15/200|\theta = 0,10) = P\left(\frac{\hat{\theta} - \theta_0}{\sqrt{\theta_0(1-\theta_0)/n}} \leq \frac{0,075 - 0,10}{\sqrt{0,10 \times 0,90/200}}\right) \\
&= \Phi(-1,18).
\end{aligned}
$$

Exemplo (cont.)

Usando o comando R *pnorm(-1.18)*, temos $\alpha \approx 0,12$, ou seja, há 12% de probabilidade de, ao se utilizar esta regra de decisão, decidirmos que houve redução na taxa de sinistros, quando de fato não houve.

Em contrapartida, se tiver havido uma redução para $\theta_1 = 0,07$, a probabilidade de que a regra de decisão adotada não detecte essa mudança é:

$$\begin{aligned}
\beta(0,7) &= P(\text{erro tipo II}|\theta_1) = P(\text{não rejeitar } H_0|\theta = 0,07) \\
&= P(\hat{\theta} > 0,075|\theta = 0,07) = 1 - \Phi\left(\frac{0,075 - 0,07}{\sqrt{0,07 \times 0,93/200}}\right) = 1 - \Phi(0,277).
\end{aligned}$$

Usando o comando R *1-pnorm(0.277)*, temos $\beta(0,07) \approx 0,39$, ou seja, se houver redução para uma taxa de sinistros $0,07$, há 39% de probabilidade de que, com esta regra, julguemos que a taxa permanece em $0,10$.

Exemplo (cont.)

Digamos que desejemos reduzir a probabilidade máxima de erro tipo I para 0,01.

Devemos alterar o valor crítico do teste e, consequentemente, sua região crítica.

Assuma que continuemos a adotar a regra genérica: “Rejeite H_0 se $T \leq c' \leftrightarrow T/n = \hat{\theta} \leq c$”.

A questão, então, é: qual é o valor de c tal que $P(\text{erro tipo I}) = 0,01$? Vejamos:

$$
\begin{aligned}
\alpha &= P(\text{erro tipo I}) = P(\text{rejeitar } H_0 | \theta = 0,10) = P(T \leq c' | \theta = 0,10) \\
&= P(\hat{\theta} \leq c | \theta = 0,10) = P\left(\frac{\hat{\theta} - \theta_0}{\sqrt{\theta(1-\theta)/n}} \leq \frac{c - 0,10}{\sqrt{0,10 \times 0,90/200}}\right) = 0,01. \\
\Leftrightarrow & \quad \frac{c - 0,10}{\sqrt{0,10 \times 0,90/200}} = -2,33 \Leftrightarrow c \approx 0,05 \Leftrightarrow c' \approx 10.
\end{aligned}
$$

Exemplo (cont.)

Ou seja, para ter probababilidade de erro tipo I igual a $0,01$, deveríamos rejeitar H_0 se $T \leq 10$, em que T é o número de segurados que tiveram sinistros, entre os observados.

Obs: no desenvolvimento acima, obtivemos o quantil $-2,33$ aplicando o comando R *qnorm(0.01)*.

Suponha que, na amostra, tivéssemos observado 8 sinistros, ou seja uma proporção 4%. Ao nível de significância 1%, a região crítica é $\{10, 9, 8, \ldots, 1, 0\}$ e, portanto, como o valor observado para a estatística de teste pertence à região crítica, rejeitaríamos H_0.

Exemplo (cont.)

Outra questão relevante é: com base nestes mesmos dados e tendo observado o mesmo número de sinistros ($t = 8$), para quais níveis de significância H_0 seria rejeitada?

Observe que, na solução acima, nos comprometemos com um nível de significância e depois observamos se o valor gerado para a estatística de teste, nos dados observados, pertencia à região crítica.

Agora, queremos saber qual seria o menor nível de significância que geraria regiões críticas contendo o valor observado $t = 8$.

Exemplo (cont.)

Resultados de um teste reportados dessa forma dão liberdade a cada usuário do teste para escolher seu próprio nível de significância (mas baseando-se nos mesmos dados experimentais).

Note que, para qualquer nível de significância α maior ou igual a:

$$p = P(T/n \leq 8/200|\theta = 0,10) = \Phi\left(\frac{0,04 - 0,10}{\sqrt{0,10 \times 0,90/200}}\right)$$

$=\Phi(-2,83) \approx 0,002.$
teremos uma região crítica que compreende o valor observado e, portanto, rejeitaremos H_0.

p-valor

O p-valor de um teste de hipóteses é o menor nível de significância, dada a amostra observada, que conduziria à rejeição de H_0.

No exemplo acima, o p-valor obtido foi $0,002$, ao se observar $t = 8$ sinistros em 200 segurados (ou seja, uma proporção amostral de $0,04$ sinistros). Qualquer analista que adote nível de significância maior ou igual a $0,002$ será levado a rejeitar H_0.

Teste de Hipóteses Clássico para Proporção

Seja θ a probabilidade de ocorrência de uma certa característica, na população de interesse.

Podemos construir testes a respeito de θ baseados na observação de uma amostra aleatória de tamanho n, registrando, para cada elemento amostral, a ocorrência ou não da característica de interesse.

Desde que n seja pequeno em relação ao tamanho da população, o número T de indivíduos que apresentam a característica estudada, na amostra, tem uma distribuição aproximadamente Binomial.

Para n grande (ainda que pequeno, comparado ao tamanho populacional), o estimador $T/n = \hat{\theta}$ (proporção amostral de sucessos) tem distribuição aproximadamente normal, de acordo com o Teorema Central do Limite.

Suponha que se deseje testar:

$$H_0 : \theta = \theta_0. \tag{12}$$

Considere a estatística de teste padronizada:

$$Z_0 = \frac{\hat{\theta} - \theta_0}{\sqrt{\theta_0(1-\theta_0)/n}} \overset{a}{\sim} N(0,1), \tag{13}$$

para tamanho amostal n suficientemente grande.

Após observarmos a amostra, teremos o valor observado para $\hat{\theta} = T/n$. Ao substituir este valor na expressão da estatística de teste (13), obtemos o valor observado desta estatística, o qual denotaremos por z.

Teste de Hipóteses Clássico para Proporção

Denote por z_p o quantil $1-p$ da distribuição normal-padrão, ou seja, aquele que deixa, à sua direita, probabilidade p sob a curva normal-padrão.

A forma da região crítica (ou seja, a região de rejeição de H_0) do teste para H_0 dependerá da forma da hipótese alternativa H_1 e do nível de signisicância α especificado.

Assuma que Z seja uma variável aleatória com distribuição $N(0,1)$. O p-valor do teste, dado o valor observado z para Z_0 em (13) depende da forma de H_1.

Teste de Hipóteses Clássico para Proporção

H_1	Rejeita-se H_0 se:	p-valor
$\theta < \theta_0$ (unilateral à esquerda)	$z < -z_\alpha$	$P(Z \leq z)$
$\theta > \theta_0$ (unilateral à direita)	$z > z_\alpha$	$P(Z \geq z$
$\theta \neq \theta_0$ (bilateral)	$z < -z_{\alpha/2}$ ou $z > z_{\alpha/2}$	$2 \times P(Z \geq \lvert z \rvert)$

Código R para testes de hipóteses sobre proporção

Admita que tenhamos um vetor de comprimento n, com entrada nulas representando a ausência da característica de interesse e entrada unitária representando a presença dessa característica.

No código a seguir, vamos usar a contagem de sucessos observados na amostra, mas pode-se usar o vetor de dados original.

Teste de Hipóteses Clássico para Proporção

```
#Informe o tamanho da amostra
n=200
#Informe o valor $\theta_0$ especificado para a hipótese nula
theta_0=0.10
#Informe o nível de significância a ser adotado
alpha=0.05
gamma=1-alpha
#Informe o vetor de dados (amostra observada) ou número de sucessos
x=8 #número de sucessos observados na amostra de tamanho n
######################################################################
# Acima, informamos o número de sucessos observados,
# mas seu vetor de dados pode ser utilizado como entrada, com o comando:
# x<-c(0,0,0,1,0,0,0,0,0,0,1,...1) #exemplo de vetor x de dados
#######################################################################
# Abaixo, para testar:
# H1: theta < theta_0, use "less"
# H1: theta > theta_0, use "greater"
# H1: theta<> theta_0, use "two.sided"
#######################################################################
prop.test(x, n, p = theta_0,alternative ="less",
          conf.level = gamma,correct='FALSE')
```

Teste de Hipóteses Clássico para Proporção

O código acima retorna a seguinte saída:

```
        1-sample proportions test without continuity correction

data:  x out of n, null probability theta_0
X-squared = 8, df = 1, p-value = 0.002339
alternative hypothesis: true p is less than 0.1
95 percent confidence interval:
 0.00000000 0.06959661
sample estimates:
   p
0.04
```

Note que, na saída do R, o valor da estatística "X-squared" é o quadrado do valor observado para a estatística em 13, que denotamos por z.

Temos, na saída do R, o p-valor obtido no exemplo 4.4.

A saída fornece ainda um intervalo ao nível de confiança 95% para θ, com limites inferiores a $0,10$, levando-nos à rejeição de H_0 e conclusão por H_1 (ao nível de significância 5%).

Teste de Hipóteses Clássico para Média de População Normal

Suponha $Y \sim N(\mu, \sigma^2)$, em que ambos os parâmetros sejam desconhecidos, e que seja possível obter uma amostra aleatória de mensurações de Y: $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$.

Suponha que desejemos testar hipóteses sobre a média μ. Em particular, admita interesse em testar:

$$H_0 : \mu = \mu_0. \qquad (14)$$

Teste de Hipóteses Clássico para Média de População Normal

Sabe-se que um estimador não tendencioso e consistente para μ é $\bar{Y}_n$ e já vimos, no capítulo 3, que se $\hat{\sigma}^2 = S^2 = \sum_{i=1}^{n} \frac{Y_i - \bar{Y}_n)^2}{n-1}$, então a estatística

$$T = \frac{\bar{Y}_n - \mu}{\sqrt{S^2}/n} \sim t_{n-1}.$$

Considere, então, a estatística de teste condicionada a H_0:

$$T_0 = \frac{\bar{Y}_n - \theta_0}{\sqrt{S^2}/n} \sim t_{n-1}. \tag{15}$$

Teste de Hipóteses Clássico para Média de População Normal

Após observarmos a amostra, teremos o valor observado para $\bar{Y}_n$ e S^2. Ao substituir estes valores na expressão da estatística de teste (15), obtemos o valor observado desta estatística, o qual denotaremos por t.

Denote por t_p o quantil $1-p$ da distribuição t de Student com $n-1$ graus de liberdade, ou seja, aquele que deixa, à sua direita, probabilidade p.

A forma da região crítica (ou seja, a região de rejeição de H_0) do teste para 14 dependerá da forma da hipótese alternativa H_1 e do nível de signisicância α especificado.

Teste de Hipóteses Clássico para Média de População Normal

Seja T uma variável aleatória com distribuição $t_(n-1)$. O p-valor do teste, dado o valor observado t para T_0 em (15) depende da forma de H_1, como ilustra a tabela a seguir.

H_1	Rejeita-se H_0 se:	p-valor
$\mu < \mu_0$ (unilateral à esquerda)	$t < -t_\alpha$	$P(T \leq t)$
$\mu > \mu_0$ (unilateral à direita)	$t > t_\alpha$	$P(T \geq t$
$\mu \neq \mu_0$ (bilateral)	$t < -t_{\alpha/2}$ ou $t > t_{\alpha/2}$	$2 \times P(T \geq \|t\|)$

Exemplo

Uma analista de uma seguradora acredita que a indenizção média paga pelos sinistros retidos pela companhia esteja em torno de 30 centenas de unidades monetárias. Admite-se, ainda, que tais indenizacões retidas sigam uma distribuicão Normal.

Uma amostra aleatória de 25 indenizações resultou em valor médio indenizado $\bar{y}_{25} = 31,37$, com desvio-padrão $s = 2,49$. Ao nível de significância $0,01$, pode-se rejeitar a crença da analista?

Observe que desejamos testar:

$$H_0 : \mu = 30 \quad vs \quad H_1 : \mu \neq 30.$$

Código R para testes de hipóteses sobre a média de população Normal

```
#Informe o valor $\mu_0$ especificado para a hipótese nula
mu_0=30
#Informe o nível de significância a ser adotado
alpha=0.01
gamma=1-alpha
#Informe o vetor de dados (amostra observada) ou média
x<-c(34.16, 28.19, 34.72, 32.70, 27.46, 31.44, 37.47, 33.58,
32.06, 32.36, 28.74, 31.73, 28.76, 28.94, 33.16, 30.35, 30.01,
26.87, 29.21, 33.75, 31.96, 31.77, 32.74, 31.51, 30.62)
##########################################################################
# Abaixo, para testar:
# H1: mu < mu_0, use "less"
# H1: mu > mu_0, use "greater"
# H1: mu<> mu_0, use "two.sided"
##########################################################################
t.test(x, mu=mu_0,alternative ="two.sided",
          conf.level = gamma)
```

Exemplo

O código acima retorna a seguinte saída:

```
        One Sample t-test

data:  x
t = 2.7487, df = 24, p-value = 0.01118
alternative hypothesis: true mean is not equal to 30
99 percent confidence interval:
 29.97596 32.76484
sample estimates:
mean of x
  31.3704
```

em que $t = 2.7487$ é o valor observado, t, para a esttaística T_0.

O p-valor do teste foi igual a $1,118\%$, indicando que H_0 só deva ser rejeitada ao se adotar níveis de significância superiores a este.

Ao nível de significância $0,01$, não se rejeita a hipótese nula de que $\mu = 30$, o que está de acordo com o intervalo de confiança ao nível 99% para μ, que contempla o valor $\mu_0 = 30$.

Conteúdo

Testes de Hipóteses Bayesianos

Admita, ainda, que tenhamos interesse nas hipóteses

$$H_0 : \theta \in \Theta_0 \quad vs \quad H_1 : \theta \in \Theta_1,$$

com $\Theta_0 \cup \Theta_1 = \Theta$, Θ o espaço paramétrico, ou seja, conjunto de todos os possíveis valores de Θ em um dado problema.

Deseja-se avaliar o problema sob a abordagem Bayesiana, ou seja, assume-se que, além da observação de uma amostra aleatória $Y_1, \ldots, Y_n$, disponha-se de um modelo a priori $\pi(\theta)$ para o parâmetro θ, refletindo as crenças do analista.

Do ponto de vista Bayesiano, ao invés de especificarmos um nível de significância, especificam-se os custos (ou perdas) associados a decisões erradas.

Seja d_0 a decisão por H_0 e d_1, a decisão por H_1. Admita a seguinte tabela de perdas associadas a decisões erradas:

	d_0	d_1
H_0 $(\theta \in \Theta_0)$ verdadeira	0	ω_0
H_1 $(\theta \in \Theta_1)$ verdadeira	ω_1	0

com ω_0, $\omega_1 > 0$.

Seja $\pi(\theta|\mathbf{y})$ a distribuição a posteriori de θ. Deseja-se encontrar o procedimento de teste que minimize a perda esperada a posteriori. Tal procedimento de teste é demnominado teste de Bayes e pode-se mostrar que é dado pela seguinte regra de decisão:

" Rejeite H_0 se $P(\theta \in \Theta_0|\mathbf{y}) \leq \frac{\omega_1}{\omega_0+\omega_1}$".

Observe que o valor crítico para rejeiçao de H_0 passa a depender da razão entre os custos ω_0 e ω_1 associados às decisões erradas, e não mais de um nível de significância.

Exemplo

Voltemos ao exemplo em que θ representa a proporção de segurados que sofrem certo tipo de sinistro, após uma política de redução de prêmios para aqueles que tomassem medidas preventivas contra esse tipo de evento. Admita que desejemos testar:

$$H_0 : \theta \geq 0,10 \quad vs \quad H_1 : \theta < 0,10.$$

Continuamos a assumir que tenhamos acesso a uma amostra aleatória de 10 segurados, dos quais 8 sofreram o sinistro em questão.

Uma priori não informativa para θ seria a $U(0,1) \equiv Beta(1,1)$.

Sob tais condições, teríamos, a posteriori,
$\theta \sim Beta(8+1, 192+1) \equiv Beta(9, 193)$ (veja, no capítulo 3, o resultado de conjugação Bernoulli-Beta). Sob tal modelo, temos:

$$P(\theta \geq 0,10|\mathbf{y}) \approx 0,0013.$$

Exemplo

A questão é: essa probalididade nos leva diretamente à rejeição de H_0?

Essa decisão dependerá da escolha dos valores de ω_0 e ω_1.

Digamos que se assuma que a rejeição indevida de H_0 (assumir que a nova política trouxe redução na proporção de sinistros, quando de fato isso não aconteceu) seja 10 vezes mais grave que a não rejeição indevida de H_0.

Pode-se traduzir tais custos adotando-se $\omega_0 = 10$, $\omega_1 = 1$, resultando em $\frac{\omega_1}{\omega_0+\omega_1} = \frac{1}{11} = 0,0909 > 0,0013$.

Portanto, rejeitaríamos H_0.

Sigam sempre aprendendo!!

Obrigada!