Amostragem e Distribuições por Amostragem

Tópicos de Estatística

José Passos

ISEG-UTL

2 de Novembro de 2009

Tabela de conteúdos

Conceitos

- Probabilidade: parte-se de um modelo probabilistico que se assume como correcto e calculam-se as probabilidades de certos acontecimentos. Neste caso o modelo (paramétrico) e os seus parâmetros são conhecidos.
- Inferência Estatística: parte-se dos dados e procura-se inferir sobre o modelo probabilistico que os gerou. Neste caso a natureza dos dados e o tipo de amostragem considerado permitem lançar alguma luz sobre o modelo paramétrico a considerar. Utilizando um procedimento adequado os seus parâmetros serão estimados a partir dos dados.

Conceitos

- Os dados (amostra) resultam da observação de uma característica de interesse (população).
- A amostra é um subconjunto da população obtido por um processo devidamente controlado.
- Modelo estatístico: corresponde à função de distribuição, F, da variável aleatória, X, que representa a característica de interesse (em estudo) no universo.
- O modelo estatístico tem que ser especificado *a priori* e pode ser paramétrico ou não paramétrico.

Conceitos

- O modelo estatístico paramétrico é geralmente definido por,

  $$\mathcal{F} = \{F(.|\theta) : \theta \in \Theta\},$$

  onde Θ é o espaço parâmetro. Isto significa que se conhece a forma de F, desconhecendo-se apenas o verdadeiro valor do parâmetro, isto é, o valor que indexa a função de distribuição.
- A especificação do modelo é uma fase essencial da inferência estatística e assenta geralmente nas seguintes características:
  - conhecimento do fenómeno em estudo
  - resultado de estudos anteriores
  - conhecimento da teoria das probabilidades
  - tipo de amostragem

Processo de amostragem

- A recolha da amostra deve obedecer a determinados critérios, existindo vários processos de amostragem.
- O processo de amostragem mais conhecido, com o recurso a métodos probabilísticos, é a amostragem casual simples.
- Outros processos de amostragem:
  - amostragem estratificada
  - amostragem de conglomerados
  - amostragem por etapas

Amostragem Casual Simples

- Numa amostra casual simples de dimensão n, proveniente de uma população X, as n v.a.'s que constituem a amostra são independentes e identicamente distribuídas. Simbolicamente, $X_1, \ldots, X_n \overset{iid}{\sim} X$
- Distribuição conjunta da amostra: se o modelo é definido por $\mathcal{F} = \{F(.|\theta) : \theta \in \Theta\}$ e a amostra é casual, $X_1, \ldots, X_n \overset{iid}{\sim} X$, tem-se,

$$\begin{aligned} F_{X_1,\ldots,X_n}(x_1, \ldots, x_n \mid \theta) &= \prod_{i=1}^{n} F_{X_i}(x_i \mid \theta) \\ &= \prod_{i=1}^{n} F(x_i \mid \theta) \end{aligned}$$

Amostragem Casual Simples: Exemplos

- Se $X_1, \ldots, X_n$ é uma amostra casual proveniente de uma população $X \sim Po(\lambda)$ tem-se,

$$
\begin{aligned}
P(X_1 = x_1, \ldots, X_n = x_n) &= f(x_1, \ldots, x_n) \\
&= \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^x}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum x}}{\prod x_i!}
\end{aligned}
$$

Amostragem Casual Simples: Exemplos

- Se $X_1, \ldots, X_n$ é uma amostra casual proveniente de uma população $X \sim N(\mu, 1)$ tem-se,

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x_i - \mu)^2\right]$$
$$= (2\pi)^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

Definição

- Estatística: é uma qualquer função da amostra que não depende de parâmetros desconhecidos.
- Exemplos: a amostra em si, a média amostral, a variância amostral, o máximo da amostra, etc.

Definição

- A distribuição por amostragem de uma estatística, $T = T(X_1, \dots, X_n)$, corresponde à sua distribuição de probabilidade, $P[T(X_1, \dots, X_n) < t(x_1, \dots, x_n)] = P[T < t]$, onde t é o valor observado da estatística para uma amostra concreta, $x_1, \dots, x_n$.

Métodos

- Métodos para obter a distribuição por amostragem de uma estatística, $T(X_1, \ldots, X_n)$:
  - mudança de variável: se X é contínua,

$$F_T(t \mid \theta) = \int_{A(t)} \prod_{i=1}^{n} f(x_i \mid \theta) dx_1 \ldots dx_n$$

    onde $A(t) = \{(x_1, \ldots, x_n) : T(x_1, \ldots, x_n) \leq t\}$
  - função geradora de momentos (ou função característica) de T
  - propriedades conhecidas da distribuição de X
  - aproximação pela distribuição assintótica com o recurso ao Teorema do Limite Central
  - aproximação por simulação

Métodos: exemplos

Considere uma amostra casual (X_1, X_2) de uma população, X, com função de distribuição $F(x)$ e densidade $f(x)$, com $-\infty < x < \infty$. A estatística $T = X_1 + X_2$ tem função de distribuição,

$$\begin{aligned} F_T(t) = P(X_1 + X_2 \leq t) &= \iint\limits_{x_1+x_2 \leq t} f_{X_1}(x_1) f_{X_2}(x_2) dx_1 dx_2 \\ &= \int_{-\infty}^{+\infty} \left\{ \int_{-\infty}^{t-x_1} f(x_2) dx_2 \right\} f(x_1) dx_1 \\ &= \int_{-\infty}^{+\infty} F(t - x_1) f(x_1) dx_1 \end{aligned}$$

Métodos: exemplos (cont)

ou, dado que $X_1 \sim X$,

$$F_T(t) = \int_{-\infty}^{+\infty} F(t-x) f(x) dx$$

e função densidade,

$$f_T(t) = \int_{-\infty}^{+\infty} f(t-x) f(x) dx$$

Métodos: exemplos (cont)

Considere uma amostra casual (X_1, X_2) de uma população exponencial com parâmetro θ, $f_X(x) = \theta e^{-\theta x}$. A estatística $T = X_1 + X_2$ tem função densidade,

$$
\begin{aligned}
f_T(t) &= \int_0^t f_{X_1}(t-x) f_{X_2}(x) dx \\
&= \int_0^t \theta e^{-\theta(t-x)} \theta e^{-\theta x} dx \\
&= \int_0^t \theta^2 e^{-\theta t} dx \\
&= \theta^2 t e^{-\theta t}, \quad t \geq 0
\end{aligned}
$$

Métodos: exemplos (cont)

Seja $X_1, \ldots, X_n$ uma amostra casual e $T = \sum_{i=1}^{n} X_i$, a estatística de interesse.

- Seja $(X_1, \ldots, X_n)$ uma amostra casual de uma população $X \sim Po(\theta)$. Sabendo que a soma de Poissons é uma Poisson, temos que $T \sim Po(n\theta)$.
- Seja $(X_1, \ldots, X_n)$ uma amostra casual de uma população $X \sim B(1, \theta)$. Sabendo que a soma de Bernoulli's é uma Binomial, temos que $T \sim B(n, \theta)$.

Definição

- A simulação Monte Carlo é um procedimento computacional que nos permite aproximar a distribuição de probabilidade de uma v.a. (estatística)
- Seja $X \sim F$. Para uma dimensão da amostra fixa, n, gera-se computacionalmente G amostras $(x_1^g, \ldots, x_n^g)$ com $g = 1, \ldots, G$, a partir de F. Para cada amostra calcula-se o valor da estatística $t^g = T(x_1^g, \ldots, x_n^g)$. Os G valores de t^g permitem-nos caracterizar a distribuição da estatística, $T(X_1, \ldots, X_n)$

Definição

Seja $(X_1, \ldots, X_n)$ uma amostra casual de dimensão n proveniente de uma população $X \sim F$. Para $k \in \mathbb{N}$ definem-se momentos amostrais:

- momento ordinário amostral de ordem k

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

- momento central amostral de ordem k

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

Definição

Nota: não confundir os momentos amostrais com os momentos populacionais e/ou os momentos da amostra

- momentos populacionais de ordem k

$$\mu_k^{'} = E[X^k]$$
$$\mu_k = E[(X - E(X))^k]$$

- momentos da amostra de ordem k

$$m_k^{'} = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$
$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

Propriedades da média amostral

Seja $(X_1, \ldots, X_n)$ uma amostra casual de dimensão n proveniente de uma população $X \sim F$, onde existem os momentos populacionais μ, μ_2, μ_3, μ_4. Prova-se o seguinte:

$$\mu_1'(\bar{X}) = E[\bar{X}] = E[X] = \mu$$

$$\mu_2(\bar{X}) = Var[\bar{X}] = \frac{Var(X)}{n} = \frac{\sigma^2}{n}$$

$$\mu_3(\bar{X}) = \frac{\mu_3}{n^2}$$

$$\mu_4(\bar{X}) = \frac{3\mu_2^2}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}$$

Propriedades da média amostral

Notas:

- Os resultados anteriores são válidos qualquer que seja a distribuição da população
- A distribuição de $\bar{X}$ está centrada na média populacional, μ
- $lim_{n\to\infty} Var(\bar{X}) = 0$

Propriedades da variância amostral

Seja $(X_1, \ldots, X_n)$ uma amostra casual de dimensão n proveniente de uma população $X \sim F$, onde existem os momentos populacionais μ, μ_2, μ_3, μ_4. Prova-se o seguinte:

$$\mu_1'(S^2) = E[S^2] = \frac{n-1}{n}\sigma^2$$

$$\mu_2(S^2) = Var[S^2] = \frac{\mu_4 - \mu_2^2}{n} + 2\frac{\mu_4 - 2\mu_2^2}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}$$

Propriedades da variância amostral corrigida

Como $E(S^2) < \sigma^2$ trabalha-se em geral com a variância amostral corrigida,

$$S'^2 = \frac{n}{n-1} S^2$$

que tem as seguintes propriedades,

$$\mu_1'(S'^2) = E[S'^2] = \sigma^2$$
$$\mu_2(S'^2) = Var[S'^2] = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\mu_2^2\right)$$

Distribuição assintótica da média amostral

Seja $(X_1, \ldots, X_n)$ uma amostra casual de dimensão n proveniente de uma população $X \sim F$. Se existir $Var(X)$ tem-se como consequência do TLC, que a média amostral, $\bar{X}$, tem distribuição assintótica normal,

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

Distribuição assintótica da média amostral

Notas:

- Para n fixo, a distribuição normal é uma aproximação da distribuição exacta de $\bar{X}$
- A qualidade desta aproximação depende da distribuição da população
- Populações com distribuições simétricas e unimodais contribuem para que a aproximação seja boa

Definição

- Seja $(X_1, X_2, \ldots, X_n)$ uma amostra casual de dimensão n. Estas v.a.s dispostas por ordem crescente definem novas v.a.s,

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

que se designam por estatísticas de ordem.

- Ao contrário da amostra casual (iid) as estatísticas de ordem não são independentes.

Exemplos

- menor e maior valores amostrais: $X_{(1)} = \min\{X_i\}$ e $X_{(n)} = \max\{X_i\}$
- mediana amostral: $M_e = X_{((n+1)/2)}$ se n ímpar e $M_e = [X_{(n/2)} + X_{((n+1)/2)}]/2$ se n par
- amplitude amostral: $R = X_{(n)} - X_{(1)}$

Distribuição por amostragem das estatísticas de ordem

- Vamos considerar apenas o caso contínuo e para simplificar faça-se $(Y_1, \ldots, Y_n) = (X_{(1)}, \ldots, X_{(n)})$
- As estatísticas de ordem $(Y_1, \ldots, Y_n)$ têm função densidade conjunta,

$$g(y_1, \ldots, y_n) = n! \prod_{i=1}^{n} f(y_i), \quad \text{para} \quad y_1 < y_2 < \cdots < y_n$$

Distribuição por amostragem das estatísticas de ordem

- O conhecimento da função densidade conjunta permite obter a distribuição marginal de qualquer estatística de ordem.
- A estatística de ordem Y_v, $v = 1, 2, \ldots, n$, tem densidade marginal,

$$g_v(y) = \frac{n!}{(v-1)!1!(n-v)!} \times$$
$$[F(y)]^{v-1}[1-F(y)]^{n-v} f(y),$$

e distribuição marginal,

$$G_v(y) = \sum_{j=v}^{n} \binom{n}{j} [F(y)]^j [1-F(y)]^{n-j}.$$

Distribuição por amostragem das estatísticas de ordem

- Se $u < v$, a densidade marginal conjunta de (Y_u, Y_v) é dada por,

$$g_{u,v}(y,z) = \frac{n!}{(u-1)!(v-u-1)!(n-v)!} \times [F(y)]^{u-1}[F(z)-F(y)]^{v-u-1}[1-F(z)]^{n-v} f(y) f(z),$$

para $y < z$.

Distribuição por amostragem das estatísticas de ordem

Casos particulares:

- função densidade e distribuição do mínimo, Y_1, da amostra:

$$g_1(y) = n[1 - F(y)]^{n-1} f(y)$$
$$G_1(y) = 1 - [1 - F(y)]^n$$

- função densidade e distribuição do máximo, Y_n, da amostra:

$$g_n(z) = n[F(z)]^{n-1} f(z)$$
$$G_n(z) = [F(z)]^n$$

Distribuição por amostragem das estatísticas de ordem

- função densidade conjunta do mínimo e máximo, (Y_1, Y_n):

$$g_{1,n}(y,z) = n(n-1)[F(z) - F(y)]^{n-2} f(y) f(z), \; y < z$$

Distribuição por amostragem das estatísticas de ordem

Exemplo: considere uma amostra casual de dimensão n de uma população com distribuição de Pareto, $X \sim Pa(c, \theta)$,

$$f(x) = \frac{\theta}{c}\left(\frac{c}{x}\right)^{\theta+1},$$

$$F(x) = 1 - \left(\frac{c}{x}\right)^{\theta}, \; x > c, \; c > 0, \; \theta > 0$$

Distribuição por amostragem das estatísticas de ordem

Exemplo (cont):

- função densidade do mínimo,

$$g_1(y) = n\frac{\theta}{y}\left(\frac{c}{y}\right)^{n\theta}$$

- função densidade do máximo,

$$g_n(z) = n\frac{\theta}{z}\left[1 - \left(\frac{c}{z}\right)^{\theta}\right]^{n-1}\left(\frac{c}{z}\right)^{\theta}$$

Distribuição por amostragem das estatísticas de ordem

Exemplo: considere uma amostra casual de dimensão n de uma população com distribuição exponencial, $X \sim Ex(\theta)$,

$$f(x) = \theta e^{-x\theta},$$
$$F(x) = 1 - e^{-x\theta}, \; x > 0, \; \theta > 0$$

Distribuição por amostragem das estatísticas de ordem

Exemplo (cont):

- função distribuição do mínimo,

$$G_1(y) = 1 - e^{-ny\theta}$$

- função distribuição do máximo,

$$G_n(z) = (1 - e^{-z\theta})^n$$

Distribuição por amostragem das estatísticas de ordem

Quantis: dado um qualquer número $0 < p < 1$, o p-ésimo quantil de uma distribuição $F(x)$ designa-se por ζ_p e define-se como o valor que satisfaz as desigualdades,

$$P(X \leq x) \geq p, \; P(X > x) \geq 1 - p$$

Casos particulares importantes são a mediana com $p = 0.5$ e os quartis com $p = s/4$ com $s = 1, 2, 3$

Distribuição por amostragem das estatísticas de ordem

Considere uma amostra casual de dimensão n de uma população com distribuição contínua F. O quantil de ordem p da amostra, Z_p, tem distribuição assintótica normal,

$$\sqrt{n} f(\zeta_p) \frac{Z_p - \zeta_p}{\sqrt{p(1-p)}} \overset{a}{\sim} N(0,1)$$

onde ζ_p é o quantil de ordem p da população.

Exemplo

Distribuição da mediana de uma população normal: $X \sim N(\mu, \sigma^2)$.
Temos neste caso $p = 0.5$, $\zeta_{0.5} = \mu$, $f(\zeta_{0.5}) = (2\pi\sigma^2)^{-1/2}$ e

$$\sqrt{\frac{2n}{\pi\sigma^2}}(Z_{0.5} - \mu) \overset{a}{\sim} N(0, 1)$$

Definição

- A função de distribuição empírica $\hat{F}_n(x)$ é a função de distribuição cumulativa com massa $1/n$ em cada ponto x_i com $i = 1, \ldots, n$

$$\hat{F}_n(x) = \frac{1}{n}\#\{i : x_i \leq x\} = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(x_i)$$

onde,

$$I_{(-\infty,x]}(x_i) = \begin{cases} 1 & \text{se } x_i \leq x \\ 0 & \text{se } x_i > x \end{cases}$$

Definição

- Não confundir entre a função de distribuição empírica definida para uma particular amostra $x_1, \ldots, x_n$ e a função de distribuição da amostra definida para as v.a.'s $X_1, \ldots, X_n$,

$$F_n(x) = \frac{1}{n}\#\{i : X_i \leq x\} = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i)$$

Definição

- Considerando as estatísticas de ordem $X_{(1)}, \ldots, X_{(n)}$ tem-se a definição equivalente,

$$F_n(x) = \begin{cases} 0 & \text{se } x < X_{(1)} \\ \frac{i}{n} & \text{se } X_{(i)} \leq x \leq X_{(i+1)}, \; i = 1, \ldots, n-1 \\ 1 & \text{se } x \geq X_{(n)} \end{cases}$$

- $F_n(x)$ é para cada x uma v.a., função da amostra casual, e portanto é uma estatística; $\hat{F}_n(x)$ é o correspondente valor observado

Propriedades

- A expressão $P\left[F_n(x) = i/n\right]$ é a probabilidade de na amostra casual de dimensão n haver i variáveis inferiores ou iguais a x e $n-i$ variáveis superiores a x.
- Portanto, para cada $x \in \Re$ tem-se $nF_n(x) \sim B(n, F(x))$,

$$P(F_n(x) = i/n) = \binom{n}{i}[F(x)]^i[1-F(x)]^{n-i}, \; i = 0, \ldots, n$$

com momentos,

Propriedades

$$E[F_n(x)] = F(x)$$
$$Var[F_n(x)] = \frac{F(x)[1 - F(x)]}{n}$$

Propriedades

Pela lei forte dos grandes números, para cada $x \in \Re$,

$$F_n(x) \stackrel{q.c.}{\longrightarrow} F(x)$$

Pelo Teorema do Limite Central, para cada $x \in \Re$,

$$\sqrt{n}\frac{F_n(x) - F(x)}{\sqrt{F(x)[1 - F(x)]}} \stackrel{a}{\sim} N(0, 1)$$

Funcionais estatísticas

Uma funcional estatística $T(F)$ é uma qualquer função de F.
Exemplos:

- média: $\mu = \int x dF(x)$
- variância: $\sigma^2 = \int (x - \mu)^2 dF(x)$
- mediana: $m = F^{-1}(1/2)$

Funcionais estatísticas

Definição: Estimador *plug-in* de $\theta = T(F)$ defini-se como,

$$\hat{\theta} = T(\hat{F}_n)$$

Definição: T é um funcional linear se $T(F) = \int r(x)dF(x)$ para uma qualquer função $r(x)$

Funcionais estatísticas

Definição: Estimador *plug-in* para um funcional linear $T(F) = \int r(x)dF(x)$ é dado por,

$$T(\hat{F}_n) = \int r(x)d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} r(X_i)$$

Funcionais estatísticas

Exemplos:

- média: $\mu = T(F) = \int x dF(x)$. O estimador *plug-in* é $\hat{\mu} = T(\hat{F}_n) = \int x d\hat{F}_n(x) = \bar{X}$
- variância: $\sigma^2 = T(F) = Var(X) = \int x^2 dF(x) - (\int x dF(x))^2$. O estimador *plug-in* é

$$\hat{\sigma}^2 = \int x^2 d\hat{F}_n(x) - \left(\int x d\hat{F}_n(x)\right)^2$$
$$= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Funcionais estatísticas

- assimetria:

$$\kappa = \frac{E[(X-\mu)^3]}{\sigma^3} = \frac{\int (x-\mu)^3 dF(x)}{[\int (x-\mu)^2 dF(x)]^{3/2}}$$

Para obter o estimador *plug-in* note que $\hat{\mu} = \bar{X}$ e $\hat{\sigma}^2 = n^{-1}\sum_i (X_i - \hat{\mu})^2$

$$\hat{\kappa} = \frac{\int (x-\mu)^3 d\hat{F}_n(x)}{[\int (x-\mu)^2 d\hat{F}_n]^{3/2}} = \frac{\frac{1}{n}\sum_i (X_i - \hat{\mu})^3}{\hat{\sigma}^3}$$