

Conceitos de Estatística

→ Falar dos três problemas do livro

Podemos pensar em estatística com um modo estruturado ou sistematizado de abordar os problemas.

Módulos do Python

→ Scipy.stats;

→ Sympy.stats;

→ Stats Models.

Tipos de convergência

Em séries e sequências, temos a ideia de convergência

$$x_n \rightarrow x_0$$

no sentido que existe um exarbitrariamente pequeno e

no sentido que existe um inteiro m , tal que para $n > m$, temos

$$|x_n - x_0| < \epsilon$$

Do ponto de vista probabilístico temos diferentes noções de convergência de uma variável aleatória.

Convergência quase certa

Dada uma variável aleatória X_n , temos:

$$P[\forall \epsilon > 0 \text{ existe } n_\epsilon > 0 \mid \forall n > n_\epsilon, |X_n - X| < \epsilon] = 1$$

Nesse caso, escrevemos

$$X_n \xrightarrow{\text{as}} X$$

almost sure

(2)

Exemplo: máximo de um conjunto de números uniformemente distribuídos.

Seja $X_n \sim U[0,1]$ e

$$X^{(n)} = \max \{X_1, X_2, \dots, X_n\}$$

A ideia é que $X^{(n)}$ deve convergir para 1 se $n \rightarrow \infty$.

Assim,

$$P(|1 - X^{(n)}| < \epsilon \text{ quando } n > m) = 1$$

como $X^{(n)} < 1$, temos

$$\begin{aligned} P(|1 - X^{(n)}| < \epsilon) &= P(1 - X^{(n)} < \epsilon) \\ &= P(1 - \epsilon < X^{(n)}) \\ &= P(X^{(n)} > 1 - \epsilon) \\ &= 1 - P(X^{(n)} < 1 - \epsilon) \\ &= 1 - P(X_1 < 1 - \epsilon)P(X_2 < 1 - \epsilon) \dots \\ &= 1 - P(X_1 < 1 - \epsilon)^n \\ &= 1 - (1 - \epsilon)^n \end{aligned}$$

Logo, $n > m \rightarrow \infty$, temos

$$P(|1 - X^{(n)}| < \epsilon) \rightarrow 1$$

* Exemplo no notebook $P = 1 - (1 - \epsilon)^n$

$$1 - P = (1 - \epsilon)^n$$

$$n = \frac{\log(1 - P)}{\log(1 - \epsilon)}$$

Convergência em probabilidade

Uma forma mais fraca de convergência é a convergência em probabilidade, no sentido

$$P(|X_n - X| > \epsilon) \rightarrow 0$$

quando $n \rightarrow \infty$ para algum $\epsilon > 0$. Esse tipo de convergência costuma ser denotado por:

$$X_n \xrightarrow{P} X$$

Exemplo Seja

$$X_n = \begin{cases} 1/2^n & \text{prob } P_n \\ c & \text{prob } 1 - P_n \end{cases}$$

Nesse caso, $X_n \xrightarrow{P} 0$ se $P_n \rightarrow 1$. Note que a parte não convergente ($X_n = c$) diminui quando $P_n \rightarrow 1$, de modo que a convergência em probabilidade faz sentido.

Exemplo Seja X_n uma variável "individuar" de quando uma outra variável $X \sim U[0, 1]$ esteja nos intervalos

$$(0, 1], (0, \frac{1}{2}], (\frac{1}{2}, 1], (0, \frac{1}{3}], [\frac{1}{3}, \frac{2}{3}], (\frac{2}{3}, 1] \dots$$

$x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \dots$

Assim, $x_1 = 1$ pois X sempre está entre 0 e 1

$$x_2 = 1 \quad \text{se } 0 \leq X \leq 1/2 \quad P(x_2=1) = 1/2$$

$$x_3 = 1 \quad \text{se } 1/2 \leq X \leq 1 \quad P(x_3=1) = 1/2$$

$$x_4 = 1 \quad \text{se } 0 < X \leq 1/3 \quad P(x_4=1) = 1/3$$

(4)

Dado $m \in \mathbb{N}$, podemos calcular

$$P(X_n > \epsilon)$$

Vemos que a sequência em probabilidade é do tipo

$$\left\{ \frac{1}{1}, \frac{1}{2}, \frac{1}{2}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \dots \right\}$$

Logo $P(X_n > \epsilon) \rightarrow 0$ ou $X_n \xrightarrow{P} 0$

Por outro lado, na sequência de realizações da variável X_n é uma lista de zeros e uns. Desse modo, X_n não é limitado por uma ϵ , ou seja, a convergência quase certa

$$P(|X_n| < \epsilon) = 1$$

para algum $n > m$ não se aplica. A diferença entre as duas é que uma está ligada aos valores da variável aleatória (a.s.) enquanto a outra está relacionada aos valores de probabilidade.

Exemplo no notebook

Convergência em distribuição

Para a convergência em distribuição, temos

$$\lim_{n \rightarrow \infty} F_n(t) = F(t)$$

Aqui, $F_n(t)$ é a distribuição acumulada estimada a partir do conjunto $\{X_1, X_2, \dots, X_n\}$ e $F(t)$ é a distribuição acumulada da variável aleatória X .

Nesse caso, costuma-se escrever

$$X_n \xrightarrow{d} X$$

Exemplo: tome uma sequência de variáveis aleatórias de Bernoulli (0 com prob. P e 1 com prob. $1-P$). Suponha $X \rightsquigarrow$ Bernoulli (P), então, direitamente temos

$$X_n \xrightarrow{d} X$$

Além disso, se definirmos $Y = 1 - X$, temos que Y será também uma distribuição de Bernoulli, logo:

$$X_n \xrightarrow{d} Y$$

Por outro lado,

$$\begin{aligned} Y_n &= 1 - X_n \\ \Rightarrow |X_n - Y_n| &= 1 \quad \forall n, \end{aligned}$$

De modo que X_n não converga para Y nem em probabilidade nem de forma quase certa. A forma de convergência em probabilidade é a mais fraca das três que discutimos aqui.

Teoremas Limites

Lei fraca dos grandes números

Seja $\{X_1, X_2, \dots, X_n\}$ um conjunto de números aleatórios identicamente distribuídos com média μ . Então, temos que

$$\bar{X}_n \xrightarrow{P} \mu \quad (\text{quando } n \rightarrow \infty)$$

sendo $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$. Em outras palavras, ⑥

a distribuição de probabilidade de \bar{X}_n tende para

$$\delta(X_n - \mu)$$

quando $n \rightarrow \infty$.

Para mostrar esse resultado é preciso revisar algumas ideias associadas à função característica. Dada uma distribuição de probabilidade, sua função característica é definida como:

$$\hat{f}(K) = \langle e^{iKX} \rangle = \int_{-\infty}^{\infty} e^{iKx} f(x) dx$$

ou seja, é a transformada de Fourier da distribuição. Logo, conhecendo $\hat{f}(K)$, temos

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iKx} \hat{f}(K) dK$$

Tal como a função geradora de momentos, a função característica pode ser usada para encontrar $\langle x^n \rangle$.

Para isso, note que

$$\begin{aligned} \hat{f}(K) &= \int_{-\infty}^{\infty} f(x) e^{iKx} dx = \int_{-\infty}^{\infty} f(x) \sum_{n=0}^{\infty} \frac{(iKx)^n}{n!} dx \\ &= \sum_{n=0}^{\infty} \frac{(iK)^n}{n!} \int_{-\infty}^{\infty} x^n f(x) dx \\ &= \sum_{n=0}^{\infty} \frac{(iK)^n}{n!} \langle x^n \rangle \end{aligned}$$

Calculando a derivada de ordem m , temos

$$\frac{\partial^m f}{\partial K^m} = \sum_{n=0}^{\infty} \frac{m! (i)^n K^{n-m}}{n!} \langle x^n \rangle$$

Tomando o valor em $K=0$,

$$\left. \frac{\partial^m f}{\partial K^m} \right|_{K=0} = (i)^m \langle x^m \rangle$$

ou ainda,

$$\langle x^m \rangle = (-i)^m \left. \frac{\partial^m f}{\partial K^m} \right|_{K=0}$$

Outra propriedade importante da função característica está associada a soma de variáveis aleatórias independentes. Sejam x_1, x_2, \dots, x_n variáveis aleatórias e

$$y = x_1 + x_2 + \dots + x_n$$

Nesse caso, é possível mostrar que

Suponha que a função característica de cada x_i é conhecida ($g_i(K)$), mas não sabemos que é a função característica de y ($G(K)$). Podemos

escrever

$$G(K) = \int_{-\infty}^{\infty} e^{iyK} P(y) dy$$

e pensar y como uma combinação de variáveis, isto é

$$P(y) = \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 \dots \int_{-\infty}^{\infty} dx_n \delta(y - (x_1 + x_2 + \dots + x_n)) P(x_1, x_2, \dots, x_n)$$

(8)

ou seja, estamos mudando de (x_1, x_2, \dots, x_n) para y .

Se as x_i são independentes, temos ainda

$$P(x_1, x_2, \dots, x_n) = P_1(x_1) P_2(x_2) \dots P_n(x_n)$$

Combinação desses dois resultados na definição de $G(k)$, ficamos com

$$\begin{aligned} G(k) &= \int_{-\infty}^{\infty} e^{iky} P(y) dy \\ &= \int_{-\infty}^{\infty} dy e^{iky} \left[\int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 \dots \int_{-\infty}^{\infty} dx_n \delta(y - (x_1 + x_2 + \dots + x_n)) P(x_1, x_2, \dots, x_n) \right] \end{aligned}$$

Integrando em y

$$G(k) = \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx_n e^{iK(x_1 + x_2 + \dots + x_n)} P(x_1, x_2, \dots, x_n)$$

usando a independência

$$G(k) = \int_{-\infty}^{\infty} dx_1 e^{ikx_1} P_1(x_1) \int_{-\infty}^{\infty} dx_2 e^{ikx_2} P_2(x_2) \dots$$

$$= g_1(k) g_2(k) \dots g_n(k)$$

$$G(k) = \prod_{i=1}^n g_i(k)$$

Assim, a função característica da soma de variáveis é o produto da função característica de cada variável.

No caso de serem independentes e identicamente distribuídas ($g_1 = g_2 = \dots = g(k)$), temos

$$G(k) = [g(k)]^n$$

Com isso, podemos mostrar a lei fraca dos grandes números. ⑨

$$\bar{X}_n = \sum_{j=1}^n X_j$$

seja $g(K)$ a função
característica de X_K .

Queremos mostrar que

$$\lim_{n \rightarrow \infty} \left\{ \frac{1}{n} \sum_{j=1}^n X_j \right\} \rightarrow \mu$$

sendo $\mu = \langle X \rangle$

O que é equivalente a mostrar que

$$P(y = \bar{X}_n) \xrightarrow[n \rightarrow \infty]{} \delta(y - \mu)$$

Podemos pensar \bar{X}_n como uma soma de v.a.i.i.d., logo

$$\begin{aligned} G(K) &= \langle e^{iK\bar{X}_n} \rangle \\ &= \langle \exp \left[\frac{iK}{n} \sum_{j=1}^n X_j \right] \rangle \\ &= \langle e^{iKx_1/n} e^{iKx_2/n} \dots e^{iKx_n/n} \rangle \\ &= \langle e^{iKx_1/n} \rangle \langle e^{iKx_2/n} \rangle \dots \langle e^{iKx_n/n} \rangle \\ &= g\left(\frac{K}{n}\right)^n \end{aligned}$$

Usando a expansão em momentos para $g(K)$

$$g(K) = \sum_{l=0}^{\infty} \frac{(iK)^l}{l!} \langle X^l \rangle$$

$$g(K/n) = \sum_{l=0}^{\infty} \frac{\left(\frac{iK}{n}\right)^l}{l!} \langle X^l \rangle$$

$$= 1 + \frac{iK}{n} \langle X \rangle + O\left(\frac{K^2}{n^2}\right)$$

Assim,

$$G(k) = \left[1 + \frac{ik}{n} \langle x \rangle + O\left(\frac{k^2}{n^2}\right) \right]^n$$

No limite de $n \rightarrow \infty$

$$G(k) = e^{ik\langle x \rangle} = e^{ik\mu}$$

Tomando a transformada inversa

$$\begin{aligned} P(y) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iky} e^{ik\langle x \rangle} dk \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ik(y-\mu)} dk \end{aligned}$$

$$P(y) = \delta(y-\mu)$$

Assim, a variável $y = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n x_i$ tem distribuição $\delta(y-\mu)$.

Lei forte dos grandes números

Dado um conjunto de v:a.i.id. $\{x_1, x_2, \dots, x_n\}$ com média μ e variação σ , então

$$\bar{x}_n = \frac{1}{n} \sum_{k=1}^n x_k$$

converge para μ de forma quase certa, ou seja,

$$\bar{x}_n \xrightarrow{\text{a.s.}} \mu$$

no sentido que

$$P(|\bar{x}_n - \mu| < \epsilon) = 1 \quad \text{para } n > m$$

Para mostrar esse resultado, podemos usar a desigualdade de Chebyshov

$$P(|X - \mu| > t) \leq \sigma^2/t^2$$

sendo μ a média de X e σ^2 sua variância. (11)

Nesse caso, basta substituir

$$X \rightarrow \bar{X}_n$$

$$t \rightarrow \epsilon$$

$$\sigma^2 \rightarrow \sigma^2/n$$

sendo essa última uma consequência do fato de a variância da soma ser a soma da variâncias, isto é,

X_k tem variância σ^2

$$\tilde{X}_n = \sum_{k=1}^n X_k \text{ tem variância } n \sigma^2$$

Para ver isso, tome inicialmente o caso da média da soma:

$$y = \xi_1 + \xi_2 + \dots + \xi_n$$

$$\langle y \rangle = \langle \xi_1 \rangle + \dots + \langle \xi_n \rangle = n \langle \xi \rangle$$

$$\text{e } \tilde{y} = \frac{1}{n} \{ \xi_1 + \dots + \xi_n \}$$

$$\langle \tilde{y} \rangle = \frac{n \langle \xi \rangle}{n} = \langle \xi \rangle$$

No caso da variância:

$$y = \xi_1 + \dots + \xi_n$$

$$\text{Var}(y) = \text{Var} \{ \xi_1 + \dots + \xi_n \}$$

(12)

$$\text{Var}\{\tilde{y}\} = \text{Var}(\xi_1) + \dots + \text{Var}(\xi_n)$$

$$= n \text{Var}(\xi) = n \langle (\xi - \langle \xi \rangle)^2 \rangle$$

se

$$\tilde{y} = \frac{1}{n} \{ \xi_1 + \dots + \xi_n \}$$

$$\text{Var}\{\tilde{y}\} = \text{Var}\left\{ \frac{1}{n} [\xi_1 + \dots + \xi_n] \right\}$$

$$= \frac{1}{n^2} \text{Var}\{ \xi_1 + \dots + \xi_n \}$$

$$= \frac{n \text{Var}(\xi)}{n^2}$$

$$\text{Var}\{\tilde{y}\} = \frac{1}{n} \text{Var}(\xi)$$

Assim, no nosso caso

$$\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$$

e com X_k tem variância σ^2 , temos que
 \bar{X}_n tem variância σ^2/n , ou desvio padrão
 σ/\sqrt{n} . Usando esses resultados na desigualdade

de Chebyshov, temos

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

Sendo assim, existe um $\delta > 0$ tal que

$$\text{se } n > \frac{\sigma^2}{\delta \epsilon^2}$$

então

$$P(|X_n - \mu| < \epsilon) > 1 - \delta.$$

completando a prova. Além disso, a forma relacionada à desigualdade de Chebyshhev pode ser pensada como ocorre a convergência para o valor de μ .

Teorema Central do Límite

Dado o conjunto $\{X_1, X_2, \dots, X_n\}$ de v.a.i.i.d. e

$$\bar{X}_n = \frac{1}{n} \{X_1 + X_2 + \dots + X_n\}$$

então \bar{X}_n converge ~~para um~~ para uma distribuição para uma gaussiana com média μ e variância σ^2/n , sendo $\mu = \langle X \rangle$ e $\sigma^2 = \langle (X - \langle X \rangle)^2 \rangle$. Para mostrar esse resultado, considere a variável

$$\begin{aligned} Z_n &= \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}}{\sigma} \left\{ \frac{1}{n} \sum_{k=1}^n (X_k - \mu) \right\} \\ &= \frac{1}{\sigma\sqrt{n}} \left\{ \sum_{k=1}^n (X_k - \mu) \right\} \\ &= \frac{1}{\sigma\sqrt{n}} \left\{ \sum_{k=1}^n (X_k - \langle X \rangle) \right\} \end{aligned}$$

Nesse caso, temos que mostrar que

$$Z_n \xrightarrow{P} Z \sim N(0, 1)$$

Para isso, vamos precisar do conceito de cumulantes. Vimos que a função característica pode ser expandida em momentos:

$$P(k) = \sum_{n=0}^{\infty} \frac{(ik)^n}{n!} \langle X^n \rangle$$

(14)

Uma outra possibilidade é

$$P(K) = \exp \left\{ \sum_{n=1}^{\infty} \frac{(ik)^n}{n!} C_n \right\}$$

sendo C_n o chamado coeficiente de ordem n .

Comparando as duas definições, é possível encontrar as relações:

$$C_1 = \langle X \rangle$$

$$C_2 = \langle X^2 \rangle - \langle X \rangle^2 = \sigma^2$$

$$C_3 = \langle X^3 \rangle - 3\langle X \rangle \langle X^2 \rangle + 2\langle X \rangle^3$$

$$C_4 = \langle X^4 \rangle - 4\langle X \rangle \langle X^3 \rangle - 3\langle X^2 \rangle^2 + 12\langle X^2 \rangle \langle X \rangle^2 - 6\langle X \rangle^4$$

:

De modo geral, podemos escrever

$$C_n = (-i)^n \left. \frac{d^n}{dK^n} \ln P(K) \right|_{K=0}$$

Sendo assim, considere a função característica da variável X_K que compõe a soma Z_n

$$g(K) = \exp \left\{ \sum_{n=1}^{\infty} \frac{(ik)^n}{n!} C_n \right\}$$

$$= \exp \left\{ iK\mu - \frac{K^2 \sigma^2}{2} + O(K^3) \right\}$$

com Z_n é uma soma, temos

$$G(K) = \int_{-\infty}^{\infty} P(z_n) e^{ikz_n} dz_n = \langle e^{ikz_n} \rangle$$

$$\begin{aligned}
 G(K) &= \left\langle \exp \left\{ \frac{iK}{\sqrt{n}\sigma} \sum_{ij=1}^n (x_{ij} - \mu) \right\} \right\rangle \\
 &= \prod_{j=1}^n \left\langle \exp \left\{ \frac{iK}{\sqrt{n}\sigma} (x_j - \mu) \right\} \right\rangle \\
 &= \prod_{j=1}^n \exp \left\{ \frac{-iK\mu}{\sqrt{n}\sigma} \right\} \left\langle \exp \left\{ \frac{iKx_j}{\sqrt{n}\sigma} \right\} \right\rangle \\
 &= \left[\exp \left\{ \frac{-iK\mu}{\sqrt{n}\sigma} \right\} \left\langle \exp \left\{ \frac{iKx_j}{\sqrt{n}\sigma} \right\} \right\rangle \right]^n \\
 &= \left[\exp \left\{ \frac{-iK\mu}{\sqrt{n}\sigma} \right\} g \left(\frac{K}{\sqrt{n}\sigma} \right) \right]^n
 \end{aligned}$$

Definindo $g(K)$ a função característica de x_j . Usando a expansão em armadeiros

$$g(K) = \exp \left\{ iK\mu - \frac{K^2\sigma^2}{2} + \mathcal{O}(K^3) \right\}$$

$$g \left(\frac{K}{\sqrt{n}\sigma} \right) = \exp \left\{ \frac{iK\mu}{\sqrt{n}\sigma} - \frac{K^2\cancel{\sigma^2}}{2n\cancel{\sigma^2}} + \mathcal{O} \left[\left(\frac{K}{\sqrt{n}\sigma} \right)^3 \right] \right\}$$

Substituindo de volta na $G(K)$, temos

$$\begin{aligned}
 G(K) &= \left[\exp \left\{ \frac{-iK\mu}{\sqrt{n}\sigma} \right\} \exp \left\{ \frac{iK\mu}{\sqrt{n}\sigma} - \frac{K^2}{2n} + \mathcal{O} \left[\left(\frac{K}{\sqrt{n}\sigma} \right)^3 \right] \right\} \right]^n \\
 &= \left[\exp \left(- \frac{K^2}{2n} + \mathcal{O} \left[\left(\frac{K}{\sqrt{n}\sigma} \right)^3 \right] \right) \right]^n
 \end{aligned}$$

Tomando o limite de $n \rightarrow \infty$, ficamos com

$$G(K) = e^{-K^2/2}$$

Tomando, finalmente, a transformada inversa de Fourier, temos

$$P(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-K^2/2} e^{-iKz} dK$$

$$P(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}$$

A expressão anterior é uma gaussiana

$$P(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

com média μ e variância unitária, completando nossa prova.

Estimativas via máxima verossimilhança

Dado um modelo composto de parâmetros, que supostamente descreve um conjunto de dados, chamamos de inferência os estimadores dos parâmetros. O processo de encontrar o melhor conjunto de parâmetros que descrevem os dados. Podemos dividir os procedimentos em paramétricos e não paramétricos. De maneira simplificada, podemos pensar no caso paramétrico como aqueles modelos que apenas o conhecimento dos parâmetros permite extrapolar nosso conhecimento sobre os dados. Já no caso de um modelo não paramétrico, precisamos

dos parâmetros e do estudo atual dos dados. (17)

Vamos nos preocupar inicialmente com o caso paramétrico, por exemplo, um modelo linear do tipo

$$Y = \alpha X + \epsilon$$

sendo Y o resultado da medida e X a variável preditora do comportamento de Y e ϵ um erro associado as quantidades não consideradas que podem afetar Y .

Tome o caso geral de um modelo com um parâmetro θ o qual está contido numa ampla classe de valores Θ . Para estimar o melhor conjunto de dados \bar{x} , o chamado $\hat{\theta}$, é comum definir uma função risco

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$$

Aqui, θ é o suposto melhor parâmetro e $\hat{\theta}$ é o parâmetro obtido dos dados \bar{x} . Note que $\hat{\theta}$ é uma variável aleatória, já que depende dos dados. Dessa modo, podemos calcular o valor esperado da função risco, isto é,

$$R(\theta, \hat{\theta}) = \int L(\theta, \hat{\theta}) f(\bar{x}; \hat{\theta}) d\bar{x}$$

sendo $f(\bar{x}; \hat{\theta})$ a distribuição dos valores de $\hat{\theta}$ em função do conjunto de dados. Usando a forma quadrática

$$E_{\theta}[(\theta - \hat{\theta})^2] = \int (\theta - \hat{\theta})^2 f(\bar{x}; \theta) d\bar{x}$$

Embora não seja útil para estimar θ , essa expressão (18) teria a chamada variance and bias relationship:

$$E_{\hat{\theta}}[(\theta - \hat{\theta})^2] = E_{\hat{\theta}}[(\theta - E_{\hat{\theta}}[\hat{\theta}] + E_{\hat{\theta}}[\hat{\theta}] - \hat{\theta})^2]$$

$$= E_{\hat{\theta}}[(\theta - E_{\hat{\theta}}[\hat{\theta}])^2 + 2(\theta - E_{\hat{\theta}}[\hat{\theta}])(E_{\hat{\theta}}[\hat{\theta}] - \hat{\theta}) + (E_{\hat{\theta}}[\hat{\theta}] - \hat{\theta})^2]$$

$$= E_{\hat{\theta}}[\underbrace{(\theta - E_{\hat{\theta}}[\hat{\theta}])^2}_{\text{não dependem de } \hat{\theta}}] + 2 E_{\hat{\theta}}[\underbrace{(\theta - E_{\hat{\theta}}[\hat{\theta}])(E_{\hat{\theta}}[\hat{\theta}] - \hat{\theta})}_{\text{não dependem de } \hat{\theta}}] + E_{\hat{\theta}}[(E_{\hat{\theta}}[\hat{\theta}] - \hat{\theta})^2]$$

$\text{Var}(\hat{\theta})$

$$= \text{Var}(\hat{\theta}) + (\theta - E_{\hat{\theta}}[\hat{\theta}])^2 + 2(\theta - E_{\hat{\theta}}[\hat{\theta}])(E_{\hat{\theta}}[\hat{\theta}] - E_{\hat{\theta}}[\hat{\theta}])$$

$$E_{\hat{\theta}}[(\theta - \hat{\theta})^2] = \text{Var}(\hat{\theta}) + (\theta - E_{\hat{\theta}}[\hat{\theta}])^2$$

$$= \text{Var}(\hat{\theta}) + \text{bias}(\theta)^2$$

Note que o termo $\text{bias}(\theta) = (\theta - E_{\hat{\theta}}[\hat{\theta}])$ representa um possível viés do estimador, ou seja, não importa a quantidade de dados, se esse termo não for nulo, o estimador $\hat{\theta}$ é enviesado.

Note ainda que se a quantidade $E_{\hat{\theta}}[(\theta - \hat{\theta})^2]$ (mean square error) estiver fixa, quando menor for $\text{bias}(\hat{\theta})^2$ maior deve ser a variância.

Existe assim um trade-off em os dois tipos
de erro para minimizar $E_{\theta}(\hat{\theta} - \theta)^2$. (19)

Na tentativa de encontrar o melhor estimador
para θ , $\hat{\theta}_{\text{mmx}}$ podemos procurar por aquele que
obtem o menor risco, ou seja,

$$\sup_{\theta} R(\theta, \hat{\theta}_{\text{max}}) = \inf_{\theta} \sup_{\theta} R(\theta, \hat{\theta})$$

ou seja, ainda que consideremos o pior θ que
maximiza $R(\theta, \hat{\theta})$, teremos o $\hat{\theta}_{\text{mmx}}$ que minimiza
o risco.

Máxima verossimilhança para distribuição
de Bernoulli

A distribuição de Bernoulli pode ser expressa como

$$P(x) = \begin{cases} P & \text{se } x=1 \\ 1-P & \text{se } x=0 \end{cases}$$

ou ainda como

$$\phi(x) = P^x (1-P)^{1-x} \quad x \in \{0,1\}$$

Supondo que temos um conjunto de dados

$$\bar{x} = \{x_1, x_2, \dots, x_n\}$$

associado a n realizações independentes de um processo
de Bernoulli, a distribuição conjunta é

$$\phi(\bar{x}) = \prod_{i=1}^n P^{x_i} (1-P)^{1-x_i}$$

Podemos pensar $\phi(\bar{x})$ como a probabilidade de encontrar um dado conjunto de valores \bar{x} , supondo que a distribuição de x_i seja uma Bernoulli com parâmetro p . Essa é a chamada função de verossimilhança, geralmente denotada por

$$\ell(p; \bar{x}) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}$$

Supondo que p não seja conhecido, vamos usar o princípio da máxima verossimilhança para determinar p , ou melhor, um estimador para p , \hat{p} . Esse estimador é aquele que maximiza $\ell(p; \bar{x})$. Note que, nesse caso, devemos pensar na verossimilhança como uma função de p , cujo máximo \hat{p} é o estimador de máxima verossimilhança do parâmetro p da distribuição de Bernoulli.

Para maximizar $\ell(p; \bar{x})$, vamos considerar o seu logaritmo,

$$\begin{aligned}\ell(p; \bar{x}) &= \ln \ell(p; \bar{x}) \\ &= \ln \left\{ \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} \right\} \\ &= \sum_{i=1}^n \ln \left[p^{x_i} (1-p)^{1-x_i} \right] \\ &= \ln p \sum_{i=1}^n x_i + \ln(1-p) \sum_{i=1}^n (1-x_i) \\ &= \ln p \sum_{i=1}^n x_i + \ln(1-p) \left[n - \sum_{i=1}^n x_i \right]\end{aligned}$$

Derivando com relação a P , temos

$$\frac{\partial \lambda}{\partial P} = \frac{1}{P} \sum_{i=1}^n x_i + \frac{(-1)}{1-P} \left[n - \sum_{i=1}^n x_i \right]$$

Igualando a zero em $P = \hat{P}$, temos

$$\left. \frac{\partial \lambda}{\partial P} \right|_{P=\hat{P}} = 0 \rightarrow \frac{1}{\hat{P}} \sum_{i=1}^n x_i = \frac{1}{1-\hat{P}} \left[n - \sum_{i=1}^n x_i \right]$$

$$S - \hat{P}S = \hat{P}n - \hat{P}S$$

$$\hat{P} = S/n$$

$$\hat{P} = \frac{1}{n} \sum_{i=1}^n x_i$$

Assim, o estimador de verossimilhança para P é igual ao valor esperado de x_i .

Podemos verificar se nosso estimador é biased.

Lembrando que

$$\text{bias} = P - E(\hat{P})$$

Podemos calcular

$$\begin{aligned} E(\hat{P}) &= E \left\{ \frac{1}{n} \sum_{i=1}^n x_i \right\} \\ &= \frac{1}{n} \sum_{i=1}^n E(x_i) \end{aligned}$$

$$= \frac{1}{n} n E(x_i)$$

O valor esperado de x_i é

$$E(x_i) = p \cdot 1 + (1-p) \cdot 0 = p$$

Assim,

$$E(\hat{P}) = P$$

e, portanto, bias = 0.

Podemos calcular ainda a variância de \hat{P} . Para isso, precisamos de encontrar $E(\hat{P}^2)$:

$$E(\hat{P}^2) = E\left\{\left[\frac{1}{n} \sum_{i=1}^n x_i\right]^2\right\}$$

$$= \frac{1}{n^2} E\left\{\left(\sum_{i=1}^n x_i\right)^2\right\}$$

$$= \frac{1}{n^2} E\left\{\sum_{i=1}^n x_i^2 + 2 \sum_{j=1}^n \sum_{i=1}^{j-1} x_i x_j\right\}$$

$$= \frac{1}{n^2} \left\{ \sum_{i=1}^n \underbrace{E(x_i^2)}_{\rightarrow P(1)^2 + (1-P)(0)^2 = P} + 2 \sum_{j=1}^n \sum_{i=1}^{j-1} \cancel{E(x_i x_j)} \right\}$$

$$= \frac{1}{n^2} \left\{ nP + 2 \frac{n(n-1)}{2} P^2 \right\}$$

$$E(\hat{P}^2) = \frac{P}{n} + P^2 - \cancel{\frac{P^2}{n}}$$

Assim, a variância de \hat{P} fica

$$\text{Var}(\hat{P}) = E(\hat{P}^2) - E(\hat{P})^2$$

$$= \frac{P}{n} - \frac{P^2}{n}$$

$$\text{Var}(\hat{P}) = \frac{P(1-P)}{n}$$

Note que essa quantidade diminui com o aumento de n , ou seja, quando mais dados, melhor nossa

estimativa para p . Observe também que o pior cenário para estimar p ocorre quando $p = 1/2$, que maximiza o valor da variância. Entretanto, sozinha, essa fórmula é praticamente inútil para estimar a variância de \hat{p} , já que ela depende do valor de p (que não é conhecido). Uma possibilidade é usar o chamado plug-in principle e substituir p por \hat{p} , encontrando assim uma estimativa para a variância de \hat{p} .

Nesse caso, ainda podemos encontrar a distribuição de probabilidade de \hat{p} . Para isso, podemos perguntar inicialmente a probabilidade de $\hat{p} = 0$, ou seja,

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i = 0 \quad x_i \in \{0, 1\}$$

Isso ocorre apenas quando $x_i = 0 \forall i$. Podemos então, usar a $L(p; \bar{x})$ para calcular a probabilidade dessa configuração ocorrer

$$L(p; \bar{x}) \Big|_{\substack{x_i=0 \\ \forall i}} = \prod_{i=1}^n p^0 (1-p)^{1-0} = (1-p)^n$$

Podemos perguntar também qual a probabilidade de $\hat{p} = 1/n$. Nesse caso, devemos ter $x_i = \delta_{ii}$,

cuja probabilidade é

$$L(p; \bar{x}) \Big|_{\substack{x_i=\delta_{ii} \\ \forall i}} = \prod_{i=1}^n p^{\delta_{ii}} (1-\delta_{ii})^{1-\delta_{ii}} = p \prod_{i=2}^n (1-p) = p(1-p)^{n-1}$$

Porém, com existem n possibilidades para escolher qual X_i é diferente de 0, a probabilidade de encontrar $\hat{P} = 1/n$ é

$$np(1-p)^{n-1}$$

Podemos prosseguir e perguntar qual a probabilidade de $\hat{P} = 2/n$. Nesse caso, dois valores de X_i devem ser não nulos. Ou, de modo geral, podemos perguntar a probabilidade de $\hat{P} = K/n$, que é

$$\mathcal{L}(P, \bar{X}) \Big|_{\sum_{i=1}^n X_i = K} = \frac{n!}{K!(n-K)!} P^K (1-P)^{n-K}$$

$\binom{n}{K}$ representam as possíveis combinações para escolher K termos não nulos em n

As maneiras de termos em n passos, K para à direita e $(n-K)$ para à esquerda ↴

Note que essa distribuição depende de K e não de \hat{P} . Porém, $K = n\hat{P}$.

Usando essa distribuição podemos determinar intervalos de confiança para os valores de \hat{P} . Por exemplo, podemos perguntar pela

$$P(|\hat{P} - P| \leq \epsilon_p) = P(P - \epsilon_p < \hat{P} < P + \epsilon_p)$$

Ou seja, a probabilidade de \hat{P} estar dentro de um intervalo de uma fração ϵ do valor de P

Vejamos um caso específico. Suponha que $p=1/2$ (25)

e $\epsilon = 1/100$. Usando

$$P_n(K, p) = \binom{n}{K} p^K (1-p)^{n-K} \text{ e que } K = \hat{N}$$
$$\hat{N} = \frac{1}{n} \sum_{i=1}^n X_i$$
$$K = \sum_{i=1}^n X_i$$

Temos que

$$P(p - \epsilon p < \hat{N} < p + \epsilon p)$$

$$= P(np - n\epsilon p < \sum_{i=1}^n X_i < np + n\epsilon p)$$

Usando os valores, temos

$$P\left(\frac{99}{200}n < \sum_{i=1}^n X_i < \frac{101}{200}n\right)$$

Tome, ainda, $n = 101$

$$P\left(\frac{9999}{200} < \sum_{i=1}^n X_i < \frac{10201}{200}\right)$$
$$= P(49,995 < \sum_{i=1}^n X_i < 51,005)$$
$$= P_{101}(K=50, \frac{1}{2}) = \frac{101!}{50!(51!)} \left(\frac{1}{2}\right)^{50} \left(1-\frac{1}{2}\right)^{51}$$
$$= 0,0788$$

Portanto, a probabilidade de encontrarmos \hat{N} entre $p - \epsilon p$ e $p + \epsilon p$ ($0,495$ e $0,505$) é aproximadamente 8%.

Podemos perguntar de outra maneira. Dado que $n=100$, o quão ~~perto~~ próximo do valor de P o estimador pode chegar. Nesse caso,

$$P(np - n\epsilon p < \sum_{i=1}^n x_i < np + n\epsilon p) = \text{"nível de confiança"}$$

ou seja, queremos encontrar o ϵ que garante que a probabilidade de \hat{p} estar entre $np - n\epsilon p$ e $np + n\epsilon p$ é igual ao nível de confiança. No caso particular, escolhendo o nível de confiança de 95%, temos

$$P(50 - 50\epsilon < \sum_{i=1}^{100} x_i < 50 + 50\epsilon) = 0,95$$

e devemos resolver para ϵ . Isso envolve resolver e

$$\text{"nível de confiança"} = \sum_{K=K_{\min}(\epsilon)}^{K_{\max}(\epsilon)} \frac{n!}{K!(n-K)!} p^K (1-p)^{n-K} \quad \text{sendo} \quad \begin{cases} K_{\min}(\epsilon) = np - n\epsilon p \\ K_{\max}(\epsilon) = np + n\epsilon p \end{cases}$$

Porém, não existe forma fechada para essa equação.

Exemplo no notebook

Estimador para uma distribuição uniforme $[0, \theta]$

Nesse caso, dado um conjunto n v.a.i.i.d $X \sim [0, \theta]$,

$$p(\theta) = \begin{cases} 1/\theta & \text{se } 0 < \theta < \theta \\ 0 & \text{tora} \end{cases}$$

Logo

$$\mathcal{L}(\theta) = \prod_{i=1}^n \frac{1}{\theta} = \frac{1}{\theta^n}$$

Note que essa função não tem derivada nula

para nenhum valor de θ . Note ainda que a função verossimilhança é monotonamente decrescente com θ . Além disso, como $x_i \in [0, \theta]$ e $n > 1$,

$$\hat{\theta} = \max_i(x_i)$$

Para ver isso melhor, pense que $L(\theta)$ deve ser nula se nula se algum x_i estiver fora do intervalo, ou seja

$$L(\theta) = \frac{1}{\theta^n} I(x_1, x_2, \dots, x_n \in [0, \theta])$$

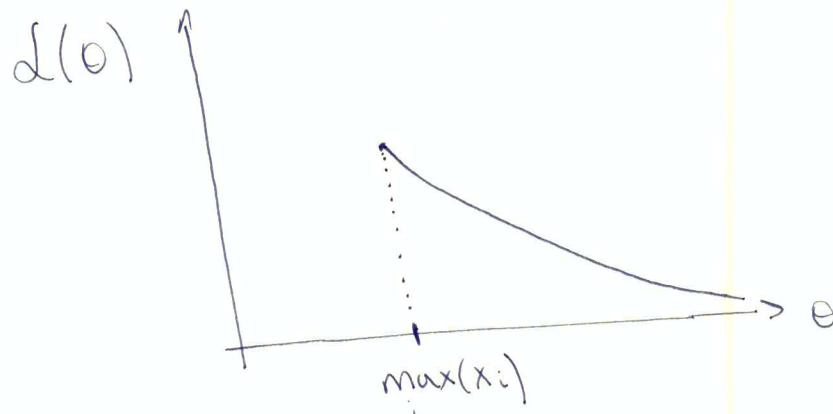
sendo $I(\dots)$ uma função indicadora que é zero quando a condição não é satisfeita e 1 quando é. Assim

$$L(\theta) = \frac{1}{\theta^n} I(\max_i(x_i) \leq \theta)$$

Podemos escrever também

$$L(\theta) = \begin{cases} 0 & \text{se } \theta < \max_i(x_i) \\ \frac{1}{\theta^n} & \text{se } \theta \geq \max_i(x_i) \end{cases}$$

Num gráfico terímos:



Logo o máximo de $L(\theta)$ é

$$\hat{\theta} = \max_i(x_i)$$

Podemos encontrar distribuições para $\hat{\theta}$. Para isso, (28)

é mais fácil encontrar primeiros a acumulada

$$P(\hat{\theta} < v) = P(x_0 < v \wedge x_1 < v \dots \wedge x_n < v)$$

ou seja $x_i < v$ $\forall i$ simultaneamente. Como x_i é uniforme de zero a θ :

$$P(\hat{\theta} < v) = \left(\frac{v}{\theta}\right) \cdot \left(\frac{v}{\theta}\right) \cdots \left(\frac{v}{\theta}\right)$$

$$P(\hat{\theta} < v) = \left(\frac{v}{\theta}\right)^n = \frac{v^n}{\theta^n}$$

Derivando podemos encontrar a distribuição de $\hat{\theta}$

$$P(\hat{\theta} < v) = \int_0^v f(\hat{\theta}) d\hat{\theta}$$

$$\frac{dP}{dv} = \frac{d}{dv} \int_0^v f(\hat{\theta}) d\hat{\theta} = f(v)$$

$$f(v) = \frac{dP}{dv} = nv^{n-1} \theta^{-n}$$

Voltando para a variável θ :

$$f(\hat{\theta}) = n \hat{\theta}^{n-1} \theta^{-n} \quad 0 < \theta < \theta$$

Usando essa distribuição podemos calcular

$$E(\hat{\theta}) = \int_0^\theta \hat{\theta} n \hat{\theta}^{n-1} \theta^{-n} d\hat{\theta} = n \theta^{-n} \int_0^\theta \hat{\theta}^n d\hat{\theta}$$

$$= \frac{n \theta^{-n} \theta^{n+1}}{n+1} = \frac{n \theta}{n+1} \quad \underline{n > 1}$$

$$E(\hat{\theta}^2) = \int_0^\theta \hat{\theta}^2 n \hat{\theta}^{n-1} \theta^n d\theta$$

$$= \frac{n\hat{\theta}^2}{2+n}$$

(29)

$n > 2$

Logo, a $\text{Var}(\hat{\theta})$ fica

$$\text{Var}(\hat{\theta}) = E(\hat{\theta}^2) - E(\hat{\theta})^2$$

$$= \frac{n\hat{\theta}^2}{(1+n)^2(2+n)}$$

Exemplo no notebook.

O Método Delta

Esse procedimento está baseado no Teorema central do limite e é uma maneira de produzir aproximações para uma função de uma variável aleatória, cuja distribuição é assintoticamente normal. Para apresentar o método, considere uma variável aleatória X que seja assintoticamente normal, isto é, dada uma sequência dessas variáveis $\{x_1, x_2, \dots, x_n\}$ então

$$\sqrt{n}(x_n - a) \xrightarrow{d} \text{Normal}(0, \sigma^2)$$

na qual a é a média assintótica de X e σ^2 a variância assintótica. Considere agora uma função $g(x)$ e uma sequência de variáveis aleatórias na forma $\{g(x_1), g(x_2), \dots, g(x_n)\}$. O método delta afirma que

$$\sqrt{n} (g(x_i) - g(a)) \rightarrow \text{Normal} \left(0, \left(\frac{dg}{dx} \Big|_{x=a} \right)^2 \sigma^2 \right) \quad (30)$$

ou seja, a variável aleatória $g(x_i)$ tem distribuição normal (do ponto de vista assintótico) com média $g(a)$ e variância $\left(\frac{dg}{dx} \Big|_{x=a} \right)^2 \sigma^2$.

Para mostrar esse resultado, tome inicialmente uma função g simples $g(x) = Ax + B$. Nesse caso,

$$E(x) \rightarrow a$$

$$E(g(x)) = AE(x) + B \rightarrow Aa + B = g(a)$$

de modo ~~analogamente~~ similar

$$\begin{aligned} E(g^2(x)) &= E[(Ax+B)^2] = E[A^2x^2 + 2ABx + B^2] \\ &= A^2 E(x^2) + 2AB E(x) + B^2 \end{aligned}$$

e, portanto,

$$\begin{aligned} \text{Var}(g(x)) &= E(g^2(x)) - E(g(x))^2 \\ &= A^2 E(x^2) + 2AB E(x) + B^2 \\ &\quad - A^2 E(x)^2 - 2AB E(x) - B^2 \end{aligned}$$

$$\begin{aligned} \text{Var}(g(x)) &= A^2 (E(x^2) - E(x)^2) \\ &= A^2 \text{Var}(x) \end{aligned}$$

Logo

$$\text{Var}(g(x)) \rightarrow A^2 \sigma^2$$

Assim a condição

$$\sqrt{n} (x_i - a) \xrightarrow{d} \text{Normal}(0, \sigma^2)$$

implica

$$\sqrt{n} (g(x_i) - g(a)) \rightarrow \text{Normal}(0, A^2 \sigma^2)$$

Suponha agora uma função $g(x)$ genérica que admitta expansão em série de Taylor

$$g(x) \approx g(a) + g'(a)(x-a)$$

Nesse caso, $g(x)$ fica linear e temos

$$\sqrt{n} (g(x_i) - g(a)) \rightarrow \text{Normal}(0, (g'(0))^2 \sigma^2)$$

Exemplo: Bernoulli Suponha um conjunto de variáveis aleatórias de Bernoulli $\{X_1, X_2, \dots, X_n\}$ com $X_k = 1$ com probabilidade P e $X_k = 0$ com probabilidade $1-P$. Suponha que desejamos estimar a razão $P/(1-P)$. Já vimos que estimador de máxima verossimilhança para P é

$$\hat{P} = \frac{1}{n} \sum X_k$$

e que

$$E(\hat{P}) = P$$

$$\text{Var}(\hat{P}) = \frac{P(1-P)}{n}$$

Sabemos ainda que \hat{P} deve ser assintoticamente distribuído de acordo com uma gaussiana. Então, estamos em condições de usar o método delta.

Para isso, pensamos a razão como ~~uma~~ uma função de p : (32)

$$g(p) = p / (1-p)$$

$$g'(p) = 1 / (1-p)^2$$

sendo assim, esperamos que $g(p)$ seja assintoticamente distribuído como uma gaussiana de média

$$g(E(\hat{p})) = p / (1-p)$$

e variância

$$\text{Var}(g(\hat{p})) = [g'(p)]^2 \text{Var}(\hat{p}) \\ = \left[\frac{1}{(1-p)^2} \right] \frac{p(1-p)}{n}$$

$$\text{Var}(g(\hat{p})) = \frac{p}{n(1-p)^3}$$

Exemplo notebook