LISTA 5: ESTIMAÇÃO PONTUAL E INTERVALO DE CONFIANÇA - SME0123

Intervalos aleatórios

Exercício 1. Seja X uma variável aleatória com distribuição uniforme Unif$(0, a)$. Notar que $E(X) = a/2$, mas a é, em princípio, desconhecido. É planejado o seguinte experimento: São extraídas duas instâncias X_1 e X_2 de X, independentes. Alguém fala:"É muito improvável que a seja maior que $X_1 + X_2$".

Qual é a probabilidade de $a > X_1 + X_2$?

Resposta: 1/2 (se fosse a soma de n instâncias, seria $1/n!$).

Exercício 2. Seja X uma variável aleatória com distribuição de Bernoulli, com probabilidade p de tomar o valor 1 e $1-p$ de tomar o valor 0. Notar que $E(X) = p$, mas p é, em princípio, desconhecido. É planejado o seguinte experimento: São extraídas n instâncias X_1, ..., X_n de X, independentes. Alguém fala:"É muito improvável que a média $\overline{X} = \sum_k X_k/n$ seja menor que 0.9 p".

Qual é a probabilidade de $\overline{X} \leq 0.9p$? (notar que a variabilidade de $\overline{X}$ é unicamente consequência da amostragem ser finita).

Resposta: A variável $n\overline{X}$ é binomial $\sim$ Binom(n, p). A pergunta é a probabilidade de $n\overline{X} \leq 0.9np$, que se calcula fazendo

```
prob=sss.binom.cdf(0.9*n*p,n,p)
```

Aqui temos o problema de que p é desconhecido. Suponhamos que $n = 100$ ou 1000 e vejamos o resultado para vários p.

p	0.01	0.1	0.3	0.5	0.7	0.9
`prob` $(n = 100)$	0.37	0.45	0.30	0.18	0.05	4E-3
`prob` $(n = 1000)$	0.46	0.02	9E-3	1E-6	8E-18	

Com isto, dependendo do valor estimado para p, é possível responder a pergunta. Por exemplo, se $n = 100$ e se estima que $p \simeq 0.7$,

$$0.05 = \text{Prob}(\overline{X} \leq 0.9p) = \text{Prob}(p \geq 1.11\overline{X})$$

Exercício 3. Seja X uma variável aleatória com distribuição Normal $N(a, 1)$. Notar que $E(X) = a$, mas a é, em princípio, desconhecido. É planejado o seguinte experimento: São extraídas duas observações X_1 e X_2 de X, independentes. Alguém fala:"É muito improvável que a média a seja menor que $\min(X_1, X_2)$".

Qual é a probabilidade de $a < \min(X_1, X_2)$?

Resposta: 1/4.

Lembrete: Propriedades da normal.
Se $X \sim N(\mu, \sigma^2)$, então $Y = X + c \sim N(\mu + c, \sigma^2)$.
Se $X \sim N(\mu, \sigma^2)$, então $Y = cX \sim N(c\mu, c^2\sigma^2)$.
Se $X \sim N(\mu, \sigma^2)$, então $(X - \mu)/\sigma \sim N(0, 1)$.

Lembrete: Soma de variáveis independentes com distribuição arbitrária de média μ_X e variança σ_X^2. Se $X_1, \ldots, X_n \sim X$ e $Y = \sum_i X_i$, então $\mu_Y = n\mu_X$ e $\sigma_Y^2 = n\sigma_X^2$.

Exercício 4. Seja X uma variável aleatória com distribuição Normal $N(\mu, 1)$, cuja média μ é desconhecida. Será extraída uma amostra de 4 elementos dessa variável: $X_1, \ldots, X_4$ e com ela serão calculadas as variáveis $A = (X_1 + X_2 + X_3 + X_4)/4 - 1$ e $B = (X_1 + X_2 + X_3 + X_4)/4 + 1$.

São A e B variáveis aleatórias? Tem A e B distribuição normal? Qual a média e qual o desvio padrão de cada uma? A e B são independentes?

Resposta: A e B são variáveis aleatórias (μ não!) com distribuição normal. $A \sim N(\mu - 1, 1/4)$ e $B \sim N(\mu + 1, 1/4)$. O desvio padrão delas é 1/2. e $B = A + 2$ (dependentes!). Notar que $Z_A = (A - (\mu - 1))/(1/2)$ tem distribuição $N(0, 1)$.

Intervalo de confiança. Estimar média com variância conhecida.

Exercício 5. Sejam A e B as variáveis aleatórias definidas no exercício anterior.

Alguém fala:"É muito improvável que o intervalo $[A, B]$ que resulte desse procedimento não contenha a média populacional μ".

Qual é a probabilidade de $\mu \notin [A, B]$?

Resposta: A e B são variáveis aleatórias (μ não!) com distribuição normal. $A \sim N(\mu - 1, 1/4)$ e $B = A + 2$ (porquê?). Notar que $Z_A = (A - (\mu - 1))/(1/2)$ tem distribuição $N(0, 1)$. Agora vejamos o perguntado:

$$\text{Prob}\,(\mu \notin [A, B]) = \text{Prob}\,(A \geq \mu) + \text{Prob}\,(A + 2 \leq \mu)$$

é também igual a (somar e subtrair o necessário nas desigualdades)

$$= \text{Prob}\,(Z_A \geq 2) + \text{Prob}\,(Z_A \leq -2)$$

$$= 1 - \text{Prob}\,(-2 \leq Z_A \leq 2)$$

```
import scipy.stats as sss
sss.norm.cdf(-2,0,1)+1-sss.norm.cdf(2,0,1)
```
=0.0455

Distribuição χ^2 e teorema central do limite. Estimar variância com média conhecida.

Lembrete: Soma de quadrados de variáveis com distribuição normal. Se $Z_1, \ldots, Z_k$ são variáveis independentes com distribuição $N(0, 1)$, e $Q = \sum_{i=1}^k Z_i^2$, então $Q \sim \chi^2(k)$. Prova-se que $E(Q) = k$ e que $\text{Var}(Q) = 2k$.

Lembrete: Teorema Central do Limite. Sejam X_1, X_2, ... uma sequência de variáveis aleatórias independentes e identicamente distribuídas, sendo $E(X_i) = \mu$ e $\text{Var}(X_i) = \sigma^2 < \infty$. Seja $\overline{X}_n = (X_1 + \ldots + X_n)/n$ a média de n delas. Então, como distribuições,

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \to N(0, 1)$$

Exercício 6. Será extraída uma amostra de n elementos independentes, X_1, ..., X_n da variável $X \sim N(1, \sigma^2)$, da qual a variância σ^2 é desconhecida. A partir dessa amostra será computada a **estatística** (nesse caso, o **estimador**)

$$Y^2 = \frac{1}{n-1} \sum_i (X_i - 1)^2 \ .$$

Alguém fala: "É muito improvável que a variância da população σ^2 não esteja no intervalo $[Y^2/2, 2Y^2]$".

Qual é a probabilidade de $Y^2/2 \leq \sigma^2 \leq 2Y^2$?

Responder para n pequeno usando a distribuição χ^2 e para n grande usando aproximação pela distribuição normal. Comparar ambas probabilidades quando $n = 20$.

Resposta: $Z_i = (X_i - 1)/\sigma \sim N(0, 1)$, então $Q = \sum_{i=1}^n Z_i^2 \sim \chi^2(n)$. Re-escrevemos a pergunta como "qual é a probabilidade de $\sigma^2/2 \leq Y^2 \leq 2\sigma^2$", sabendo que $Y^2 = Q\sigma^2/(n-1)$. Assim, a pergunta se torna: "Qual é a probabilidade de $(n-1)/2 \leq Q \leq 2(n-1)$?". Agora, seja $n = 5$, nesse caso $(n-1)/2 = 2$ e $2(n-1) = 8$, e

```
scipy.stats.chi2.cdf(2,5)=0.1508
scipy.stats.chi2.cdf(8,5)=0.8438
```

Por tanto, a probabilidade pedida é 0.6929.

As mesmas contas, quando $n = 20$, levam à probabilidade de $19/2 \leq Q \leq 38$, que é 0.9675 (neste caso devemos usar a $\chi^2(20)$.

```
scipy.stats.chi2.cdf(9.5,20)=0.02364
scipy.stats.chi2.cdf(38,20)=0.99114
```

Para aproximar com a normal, vemos que a média de Q é n e $\text{Var}(Q) = 2n$. Como Q é soma das n variáveis identicamente distribuídas $Y_i = Z_i^2$ e de variância finita, ela é aproximadamente $N(n, 2n)$. A probabilidade aproximada é 0.9493.

```
scipy.stats.norm.cdf(9.5,20,np.sqrt(40))=0.04844
scipy.stats.norm.cdf(38,20,np.sqrt(40))=0.99779
```

Para n maior, a probabilidade é praticamente 1.

Distribuição t de Student. Estimar média com variância desconhecida.

Lembrete: Se $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, e consideramos as **estatísticas** (de fato, **estimadores**)

$$\overline{X} = \frac{1}{n}\sum_i X_i, \qquad S^2 = \frac{1}{n-1}\sum_i \left(X_i - \overline{X}\right)^2 \ .$$

então $T = (\overline{X} - \mu)/(S/\sqrt{n})$ tem distribuição t_{n-1}. Prova-se que $E(t_k) = 0$, $\mathrm{Var}(t_k) = k/(k-2)$.

Exercício 7. Será extraída uma amostra de n elementos independentes, X_1, ..., X_n da variável $X \sim N(\mu, \sigma^2)$, da qual os parâmetros $\theta = (\mu, \sigma^2)$ são desconhecidos. A partir dessa amostra serão computadas as estatísticas $\overline{X}$ e S^2.

Alguém fala: "É muito improvável que a média da população μ não esteja no intervalo $[\overline{X} - S/10, \overline{X} + S/10]$".

Qual é a probabilidade de $\overline{X} - S/10 \le \mu \le \overline{X} + S/10$?

Responder para n pequeno usando a distribuição t de Student e para n grande usando aproximação pela distribuição normal. Comparar ambas probabilidades quando $n = 10$, 100, 1000.

Resposta: As desigualdades correspondem a

$$\frac{\overline{X} - \mu}{S} - \frac{1}{10} \le 0 \le \frac{\overline{X} - \mu}{S} + \frac{1}{10}$$

ou seja $T/\sqrt{n} - 0.1 \le 0 \le T/\sqrt{n} + 0.1$ ou seja $-0.1\sqrt{n} \le T \le 0.1\sqrt{n}$. Ademais sabemos que T tem distribuição t_{n-1}, então, considerando $n = 10$,

```
scipy.stats.t.cdf(-0.1*np.sqrt(n),n-1)
```
=0.3795
```
scipy.stats.t.cdf(0.1*np.sqrt(n),n-1)
```
=0.6205

A probabilidade é a diferença, 0.2409.

Por outro lado, a média de T é 0 e a variança é $(n-1)/(n-3)$, e quando n é grande pode ser aproximada por $N(0, (n-1)/(n-3))$. Isto nos fornece uma probabilidade aproximada de 0.2197:

```
scipy.stats.norm.cdf(-0.1*np.sqrt(n),0,np.sqrt((n-1)/(n-3)))
```
=0.3902
```
scipy.stats.norm.cdf(0.1*np.sqrt(n),0,np.sqrt((n-1)/(n-3)))
```
=0.6098

A probabilidade de o intervalo aleatório $[\overline{X} - 0.1S, \overline{X} + 0.1S]$ conter a média μ (desconhecida), é assim 0.24 para $n = 10$, 0.68 para $n = 100$ e 0.998 se $n = 1000$. Esse intervalo, cuja largura não depende de n, tem mais e mais confiabilidade a medida que n cresce.

O intervalo aleatório $[\overline{X} - 0.1S, \overline{X} + 0.1S]$ (ambos extremos dele são variáveis aleatórias) é um **intervalo de confiança** de 99.8% para a média μ quando $n = 1000$. A confiabilidade cai para 68% se $n = 100$ e para 24% se $n = 10$.

Se quisermos um intervalo de confiança de 95%, colocaríamos o intervalo $[\overline{X} - 0.062S, \overline{X} + 0.062S]$ (um pouco menor que o anterior, o que é lógico já a confiabilidade é menor).

```
scipy.stats.t.ppf(0.975,n-1)/np.sqrt(n)
```
$=1.962/\sqrt{n}$=0.062.

Exercício 8. Seja X uma variável aleatória com distribuição de Bernoulli, com probabilidade p de tomar o valor 1 e $1-p$ de tomar o valor 0. Notar que $E(X) = p$, mas p é, em princípio, desconhecido. É planejado o seguinte experimento: São extraídas n observações X_1, ..., X_n de X, independentes. Construa um intervalo de confiança de 95% para p a partir da estatística $f = \overline{X}$ (frequência amostral?).

Resposta: Vemos que $nf = \sum_k X_k$ tem distribuição binomial e tenderá a ter distribuição normal. Sabemos que $E(X_k) = p$ e que $\mathrm{Var}(X_k) = p(1-p)$. O intervalo de confiança da binomial se calcula com

```
n=1000
f=0.5,
intervalo=sss.binom.interval(0.95,n,f)
```

```
(469.0 531.0)
```

Onde utilizamos f como estimativa de p. Por isto, se a frequência amostral é igual a $f = 0.5$, concluímos que um intervalo de 95% de confiança para p é $0.469 \le p \le 0.531$.

Por outro lado, sabemos que f tem média $\mu_f = p$, variância $\sigma_f^2 = p(1-p)/n$ e vemos que satisfaz as hipóteses do TCL. Assim, $Z = (f-p)/\sqrt{p(1-p)/n}$ é aproximadamente $N(0,1)$. Na abordagem "otimista" aproximamos $p(1-p) \simeq f(1-f)$ e por tanto definimos $S^2 = f(1-f)/n$. Na abordagem "conservativa" definimos $S^2 = 1/(4n)$. Calculamos o intervalo com a aproximação normal como segue:

```
f=0.5
n=1000
S=np.sqrt(f*(1-f)/n)
z95=sss.norm.ppf(0.975,0,1)
print(z95,f-z95*S,f+z95*S)
```

```
1.9599 0.46901 0.53099
```

Exercício 9. Dado que a população de homens de certa cidade tem pesos distribuídos normalmente com média 78, 47Kg e desvio-padrão 13, 61Kg, determinar a probabilidade de:

(a) Um homem escolhido aleatoriamente pesar mais de 81, 65Kg.
(b) Em 36 homens escolhidos aleatoriamente, o peso médio ser superior a 81, 65Kg.

Resposta: 0, 9999.

Exercício 10. Calcule o intervalo de confiança para a média de uma $N(\mu, \sigma^2)$ em cada um dos casos abaixo:

(a) $\bar{x} = 170$cm; $n = 100$; $\sigma = 15$cm; $\alpha = 5\%$.
(b) $\bar{x} = 165$cm; $n = 184$; $\sigma = 30$cm; $\alpha = 15\%$.

Resposta: (a) $[167,06; 172,94]$, (b) $[161,82; 168,18]$.

Exercício 11. Por analogia a produtos similares, o tempo de reação de um novo medicamento pode ser considerado como tendo distribuição Normal com desvio padrão igual a 2 min. Vinte pacientes foram sorteados, receberam a medicação e tiveram seu tempo de reação anotado. Os dados foram os seguinte (em min):

2,9	3,4	3,5	4,1	4,6	4,7	4,5	3,8	5,3	4,9
4,8	5,7	5,8	5,0	3,4	5,9	6,3	4,6	5,5	6,2

Obtenha um intervalo de confiança para o tempo médio de reação. Fixe o coeficiente de confiança do intervalo em 96%.

Resposta: $\bar{X} = 4.745$ e $IC_\mu(0.96) = [3.826; 5.664]$

Exercício 12. Será coletada uma amostra de uma população Normal com variância igual a 81. Para uma confiança de 90%, determine a amplitude do intervalo de confiança para a média populacional nos casos em que o tamanho da amostra é 30, 50 e 100. Comente as diferenças.

Resposta: $n = 30$, amplitude do $IC_\mu(0.90) = 5.406$, $n = 50$, amplitude do $IC_\mu(0.90) = 4.187$, $n = 100$, amplitude do $IC_\mu(0.90) = 2.961$.

Exercício 13. Numa pesquisa com 50 eleitores, o candidato X obteve 0, 34 da preferência dos eleitores. Construa, para a confiança 94%, o intervalo de confiança para a proporção de votos a serem recebidos pelo candidato mencionado, supondo que a eleição fosse nesse momento.

Resposta: $IC_p(0.94) = [0.214; 0.466]$