Capítulo 1

Distribuição Amostral

1.1 Introdução

O conhecimento das distribuições amostrais é fundamental para a aplicação de técnicas de inferências estatísticas. Inferência é o processo de obter informações sobre uma população a partir de uma amostra.

Sendo uma população com N elementos, da qual se deseja retirar todas as amostras possíveis de tamanho n, ou seja, podem ser obtidas N^n amostras (no caso de retiradas com reposição) ou $C_{N,n}$ amostras (no caso de retiradas sem reposição). Em cada amostra pode-se calcular o valor de um certo estimador $\hat{\theta}$ $(\bar{X},\ S^2,\ etc.)$; o conjunto de valores das estimativas de $\hat{\theta}$, forma uma distribuição de probabilidades denominada distribuição amostral (ou por amostragem) de $\hat{\theta}$.

Seja X uma variável populacional que se deseja estudar. Uma amostra aleatória de X é o conjunto de n variáveis aleatórias independentes $(X_1,\ X_2, ..., X_n)$ tal que cada $X_i,\ i = 1, 2, ..., n$ tem a mesma distribuição da variável X. Se a população for infinita, os valores da amostra aleatória também serão igualmente distribuidos pois, neste caso, a retirada de alguns elmentos não modificará a distribuição de probabilidade da população.

Para formalizar s idéias que serão apresentadas neste capítulo precisamos definir alguns conceitos que seguem abaixo.

1.1.1 Parâmetros, Estimativas e Estimadores

Def.: Parâmetro é a quantidade da população, em geral desconhecida, que temos interesse. Usualmente representadas pelas letras gregas como θ, μ, σ, entre outras.

Def.: Estimador é a combinação dos elementos da amostra construida com a finalidade de estimar um parâmetro de interesse da população. Em geral, denotada por $\hat{\theta}, \hat{\mu}, \hat{\sigma}$, entre outras.

Def.: Estimativas são os valores numéricos assumidos pelos estimadores.

Notamos que um estimador, digamos $\hat{\theta}$, é uma função das variáveis aleatórias da amostra, isto é, $\hat{\theta} = f(X_1, X_2, ..., X_n)$. Logo, um estimador também é uma variável aleatória.

O problema da estimação é, então, determinar uma função $\hat{\theta} = f(X_1, X_2, ..., X_n)$ que seja próxima de θ, segundo algum critério. Tais critérios (propriedade) são apresentados a seguir.

Prop.1 $\Rightarrow$ Vício: Um estimador $\hat{\theta}$ é não viciado ou não viesado para um parâmetro θ se $E(\hat{\theta}) = \theta$.

Prop.2 $\Rightarrow$ Consistência: Um estimador $\hat{\theta}$ é consistente se, à medida que o tamanho da amostra aumenta, seu valor esperado converge para o parâmetro de interesse e sua variância converge para zero. Ou seja,

$$i) \quad \lim_{n\to\infty} E(\hat{\theta}) = \theta \qquad ii) \quad \lim_{n\to\infty} Var(\hat{\theta}) = 0.$$

Quando dois estimadores forem consistentes e não viciados para um mesmo parâmetro precisamos de definir qual é mais preciso. Neste caso, temos o conceito de eficiência do estimador, como segue.

Prop.3 $\Rightarrow$ Eficiência: Dado dois estimadores $\hat{\theta}_1$ e $\hat{\theta}_2$, não viciados para um parâmetro θ, dizemos que $\hat{\theta}_1$ é mais eficiente do que $\hat{\theta}_2$ se $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$.

Exemplo: No caso de Distribuição Normal, verifica-se que os estimadores $\hat{\mu}_1 = \bar{X}$ e $\hat{\mu}_2 =$ $mediana(X_1, ..., X_n)$ são não viciados e suas variâncias são

$$Var(\hat{\mu}_1) = \frac{\sigma^2}{n} \quad \text{e} \quad Var(\hat{\mu}_2) = \frac{(\pi/2)\sigma^2}{n}.$$

Então,

$$\frac{Var(\hat{\mu}_1)}{Var(\hat{\mu}_2)} = \frac{\sigma^2/n}{(\pi/2)\sigma^2/n} = \frac{2}{\pi} = 0,63 < 1 \Rightarrow Var(\hat{\theta}_1) < Var(\hat{\theta}_2),$$

concluimos que $\hat{\mu}_1$ é mais eficiente do que $\hat{\mu}_2$.

1.2 Distribuição Amostral da Média $\bar{X}$

Seja $(X_1,\ X_2, ..., X_n)$ uma amostra aleatória extraida de uma população X com média μ e variância σ^2. Sabe-se que $\bar{X} = (\sum_{i=1}^{n} x_i)/n$ representa a média amostral. Como a amostra é aleatória, cada X_i tem a mesma distribuição da população, ou seja,

$$E(X_i) = \mu \qquad \text{e} \qquad Var(X_i) = \sigma^2, \quad i = 1, 2, ..., n.$$

Como a amostra foi extraida ao acaso, $X_1,\ X_2, ..., X_n$ são independentes. Então,

$$E(\bar{X}) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n\mu}{n} = \mu.$$

Portanto, $E(\bar{X}) = \mu \ \rightarrow$ é um estimador justo de μ.

$$Var(\bar{X}) = Var\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Portanto, $Var(\bar{X}) = \frac{\sigma^2}{n}$ $\rightarrow$ é consistente, pois $Var(\bar{X}) \rightarrow\ 0,\ n\ \rightarrow\ \infty$.

Assim, se $X \stackrel{\mathrm{d}}{=} N(\mu, \sigma^2) \ \Rightarrow \ \bar{X} \stackrel{\mathrm{d}}{=} N\left(\mu, \frac{\sigma^2}{n}\right)$.

Então, se a distribuiçãoda população for normal, a distribuição amostral de $\bar{X}$ também será normal, para qualquer tamanho da amostra.

No caso de amostragem sem reposição de população finita, pode-se demonstrar que:

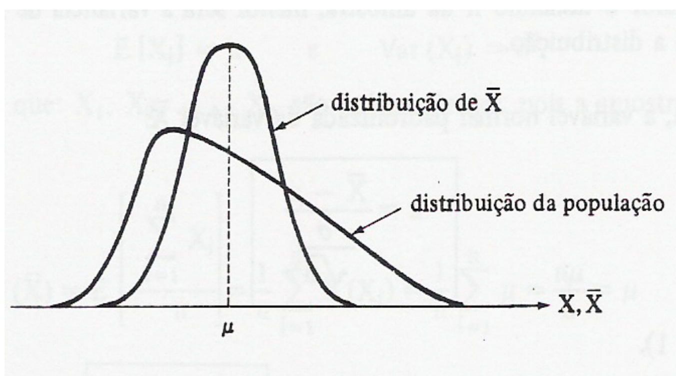
$$X \stackrel{\mathrm{d}}{=} N(\mu, \sigma^2) \;\Rightarrow\; \bar{X} \stackrel{\mathrm{d}}{=} N\left[\mu, \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)\right],$$

onde a fração $\left(\frac{N-n}{N-1}\right)$ é chamada fator de correção da população finita.

Se a distribuição da população não for normal, mas a amostra for suficientemente grande, resultará do Teorema do Limite Central que a distribuição amostral de $\bar{X}$ será aproximadamente normal no caso de população infinita ou amostragem com reposição.

O **Teorema do Limite Central** garante que *"sob condições bastante gerais, uma variável aleatória, resultante de uma soma de n variáveis aleatórias independentes, tem distribuição normal, no limite para n tendendo a infinito".*

Na figura abaixo tem-se uma distribuição populacional não-normal e a correspondente distribuição amostral de $\bar{X}$, para um n suficientemente grande.



Exemplo: Seja X uma população constituida dos seguintes elementos 2, 3, 4, 5. Extrair todas as amostras de 2 elementos dessa população, com reposição, e determinar:

a) Média e variância da população.

b) Média e variância da distribuição das médias.

Solução: *Temos que:*

a) $\mu(X) = \frac{2+3+4+5}{4} = 3,5.$

$\sigma^2(X) = \frac{(2-3,5)^2+(3-3,5)^2+(4-3,5)^2+(5-3,5)^2}{4} = 1,25.$

b) $E(\bar{X}) = \mu = 3,5; \quad Var(\bar{X}) = \frac{\sigma^2(X)}{n} = \frac{1,25}{2} = 0,625.$

Podemos verificar a validade dos resultados obtidos. Para isso, listar todas as amostras de tamanho 2, com reposição; a seguir calcular a média de cada amostra e, em seguida, calcular a média e variância das médias amostrais.

Assim, teremos:

$$\begin{aligned} Amostras &= \{(2,2),(2,3),(2,4),(2,5),(3,2),(3,3),(3,4),(3,5), \\ &\quad (4,2),(4,3),(4,4),(4,5),(5,2),(5,3),(5,4),(5,5),(5,6)\} \\ Médias\ Amostrais &= \{2;2,5;3;3,5;2,5;3;3,5;4;3;3,5;4;4,5;3,5;4;4,5;5\} \end{aligned}$$

Com reposição $\Rightarrow \quad N^n = (4)^2 = 16.$

$E(\bar{X}) = \frac{2+2,5+...+5}{16} = 3,5$

$Var(\bar{X}) = \frac{(2-3,5)^2+...+(5-3,5)^2}{16} = \frac{10}{16} = 0,625.$

Podemos desenvolver o item (b) supondo agora amostragem sem reposição:

$E(\bar{X}) = \mu = 3,5; \quad Var(\bar{X}) = \frac{\sigma^2(X)}{n}\left(\frac{N-n}{N-1}\right) = \frac{1,25}{2}\left(\frac{4-2}{4-1}\right) = 0,417.$

ou, considerando $C_{N,n} = C_{4,2} = 6$

$$\begin{aligned} Amostras &= \{(2,3),(2,4),(2,5),(3,4),(3,5),(4,5)\} \\ Médias\ Amostrais &= \{2,5;3;3,5;3,5;4;4,5\} \end{aligned}$$

$E(\bar{X}) = \frac{2,5+3+...+4,5}{6} = 3,5 \quad$ *e* $\quad Var(\bar{X}) = \frac{(2,5-3,5)^2+...+(5-3,5)^2}{6} = 0,417.$

1.3 Distribuição Amostral da Frequência Relativa

Seja X uma população infinita, p a probabilidade de sucesso de certo evento e $q = 1 - p$ a probabilidade de insucesso.

Seja $(x_1, x_2, ..., x_n)$ uma amostra aleatória de n elementos dessa população e x o número de sucessos na amostra. Então, temos uma variável com distribuição Binomial com média (np) e variância (npq). Assim, a distribuição amostral da frequência relativa (proporção) $f = \frac{x}{n}$ será dada por:

$$E(f) = E\left(\frac{x}{n}\right) = \frac{np}{n} = p \ \text{ e } \ Var(f) = Var\left(\frac{x}{n}\right) = \frac{1}{n^2}npq = \frac{pq}{n}.$$

Para $n \geq 30$ a distribuição amostral de f será normal

$$f \stackrel{d}{=} N(p, \frac{pq}{n}).$$

Assim, a distribuição padronizada será

$$Z_i = \frac{f_i - p}{\sqrt{\frac{pq}{n}}}.$$

1.4 Distribuição Amostral de Variâncias S^2

Sabemos que a variância populacional é designada por σ^2. Então, a variância amostral S^2 é o estimador de σ^2. Assim, S^2 tem distribuição qui-quadrado (χ^2) com $(n-1)$ graus de liberdade (ouseja, a soma de $(n-1)$ variáveis normais padronizadas e independentes) dada por:

$$\frac{(n-1)S^2}{\sigma^2} \stackrel{d}{=} \chi^2_{(n-1)}.$$

Aqui $(n-1)$ e σ^2 são constantes.

Os parâmetros da distribuição amostral de S^2 são:

$$E(S^2) = \sigma^2 \qquad \text{e} \qquad Var(S^2) = \frac{2\sigma^4}{(n-1)}.$$

1.5 Distribuição Amostral da Soma ou Diferença de Duas Médias

Queremos obter a distribuição amostral do estimador $(\bar{X}_1 \pm \bar{X}_2)$.

Sabemos que: $\bar{X}_1 \stackrel{d}{=} N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$ e $\bar{X}_2 \stackrel{d}{=} N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$.

Considerando-se as amostras independentes de duas populações tem-se:

$$(\bar{X}_1 \pm \bar{X}_2) \stackrel{d}{=} N(\mu_1 \pm \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}).$$

1.6 Distribuição amostral da Soma ou Diferença de duas Frequências

Se $f_1 \stackrel{\mathrm{d}}{=} N(p_1, \frac{p_1 q_1}{n_1})$ e $f_2 \stackrel{\mathrm{d}}{=} N(p_2, \frac{p_2 q_2}{n_2})$ são válidas quando $n \geq 30$, então a distribuição amostral das diferenças ou somas será aproximadamente normal com:

$$(f_1 \pm f_2) \stackrel{\mathrm{d}}{=} N(p_1 \pm p_2, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}).$$

1.7 Exercícios

1- Uma populaçãoconsiste de cinco números:2,3,6,8,11. Consideremos todas as amostras possíveis de 2 elementos que dela podemos retirar, com reposição. Determinar:

a) a média e o desvio padrão da população;

b) a média e o desvio padrão da distribuição amostral das médias.

2- Resolver o problema anterior, no caso da amostragem sem reposição.

3- Admitindo-se que o comprimento das peças produzidas por uma máquina (produção de 3.000 peças) é normalmente distribuida com uma média de 172,72 mm e desvio padrão de 7,62 mm, e obtendo-se 80 amostras, de 25 peças cada uma, quais serão a média e o desvio padrão esperdados da distribuição amotral das médias, supondo que a amostragem é feita:

a) com reposição; b) sem reposição.

4- Em quantas amostras do problema anterior pode-se esperar que a média se encontre:

a) entre 169,67 mm e 173,48 mm; b) abaixo de 169,65 mm.

5- Certos amortecedores fabricados por uma empresa têm uma vida média de 800 dias e desvio padrão de 60 dias. Determinar a probabilidade de uma amostra aleatória de 16 amortecedores, retirados do grupo, ter a vida média:

a) entre 770 e 830 dias; c) superior a 820 dias;

b) menor que 785 dias; d) entre 790 e 810 dias.

6- Os pesos dos pacotes recebidos por um depósito têm uma média de 150 kg e um desvio padrão de 25 kg. Qual é a probabilidade de 25 pacotes, retirados ao acaso e carregados em, um elevador, não excederem o limite de segurança deste, que é de 4.100 kg?

7- Determinar a probabilidade de, em 120 lances de uma moeda honesta, ocorrerem caras:

a) entre 40 e 60%; b) em 5/8 ou mais.

8- Uma pesquisa de opinião pública numa comunidade mostrou que 46% das pessoas são favoráveis a um projeto de lei. Determinar a probabilidade de que a maioria das pessoas, de um conjunto amostral de 1.000 pessoas, seja favorável a tal projeto.

9- Entre 1.000 amostras, de 200 crianças cada uma, em quantas espera-se encontrar:

a) menos de 40% de meninos;

b) 53% ou mais de meninas.

10- Seja X_1 uma variável que representa qualquer dos elementos da população:3,7,8. E X_2 que representa qualquer dos elementos da população: 2 e 4. Calcular:

a) $E(X_1)$ e $Var(X_1)$;

b) $E(X_2)$ e $Var(X_2)$;

c) $E(X_1 - X_2)$ e $Var(X_1 - X_2)$.

Respostas:

1- a) $\mu = 6,0$; $\sigma = 3,29$ b) $E(\bar{X}) = 6,0$; $S(\bar{X}) = 2,32$.

2- a) $\mu = 6,0$; $\sigma = 3,29$ b) $E(\bar{X}) = 6,0$; $S(\bar{X}) = 2,01$.

3- a) $E(\bar{X}) = 172,72$; $S(\bar{X}) = 1,524$ b) $E(\bar{X}) = 172,72$; $S(\bar{X}) = 1,518$.

4- a) 54 b) 2.

5- a) 95,44% b) 15,87% c) 9,18% d) 49,72%.

6- 99,74%.

7- a) 97,08% b) 0,32%.

8- 0,62%.

9- a) 2 b) 200

10- a) $\mu(X_1) = 6$; $\sigma(X_1) = 2,16$

b) $\mu(X_2) = 3$; $\sigma(X_2)) = 1$

c) $\mu(X_1 - X_2) = 3$; $\sigma(X_1 - X_2) = 2,38$.