INFERÊNCIA ESTATÍSTICA (EST0035)

AULA02: DISTRIBUIÇÕES AMOSTRAIS

Frederico Machado Almeida
frederico.almeida@unb.br

Departamento de Estatística
Instituto de Exatas
Universidade de Brasília (UnB)

- Vimos anteriormente que, uma estatística (ou estimador) $T_n = T(X_1, \cdots, X_n)$ é qualquer função de uma a.a., e portanto, também é uma v.a.
- Consequentemente, T_n pode estar associada a uma distribuição de probabilidades (distribuição amostral da estatística).

Exemplo 6: um jogo consiste em lançar uma moeda honesta 3 vezes. Para cada lançamento, se sair cara o jogador ganha 1 ponto, caso saia coroa, ele perde 1 ponto. Nisso, podemos modelar essa situação através de uma v.a.

$$X_i = \begin{cases} +1 & \text{com} \quad p = 1/2, \\ -1 & \text{com} \quad q = 1/2. \end{cases}$$

Como o experimento consiste em lançar uma a moeda honesta 3 vezes, vamos denotar por (X_1, X_2, X_3), a nossa a.a.

Distribuições Amostrais

Objetivo: encontrar as distribuições de probabilidades para os estimadores $T_{1n} = \bar{X}_n$ e $T_{2n} = S^2$, que denotam a média e variância amostral, respectivamente.

- Fazendo os devidos cálculos, obtemos os seguintes resultados:

$\bar{x}_n$	-1	-1/3	1/3	1
$P(\bar{X}_n = \bar{x}_n)$	1/8	3/8	3/8	1/8

Tabela: Distribuição amostral da média $T_{1n} = \bar{X}_n$.

s_n^2	0	4/3
$P(S^2 = s_n^2)$	1/4	3/4

Tabela: Distribuição amostral da variância $T_{2n} = S^2$.

- Por fim, podemos provar que T_{1n} e T_{2n} são estimadores não-viesados para μ e σ^2, respectivamente.

Distribuições Amostrais

- Observe que, no exemplo anterior, foi possível enumerar todas as amostras necessárias, e de seguida conseguimos obter as distribuições de probabilidades de T_{1n} e T_{2n}.
- No entanto, na maioria das vezes não é viável enumerar todos resultados possíveis. Portanto, precisamos de ferramentas que possam nos auxiliar a encontrar as distribuições dos estimadores, nos casos em que não conseguimos enumerar todas as possíveis a.a.'s.
- Obter a distribuição amostral de uma estatística T_n é um processo tão (mais) complexo do que trabalhar com toda a população (caso seja possível).

Distribuição Amostral de $\bar{X}_n$

- Nas aulas anteriores vimos que a média amostral $\bar{X}$ é um estimador não-viesado para a quantidade alvo θ. No entanto, o teorema abaixo estabelece um resultado importante.

Teorema 1

Seja $X_1, \cdots, X_n$ uma a.a. de tamanho n extraída de uma população representada pela v.a. X. Tal que, $\mu = \mathbb{E}(X)$ e $\sigma^2 = Var(X)$. Então, segue que,

1. $\mathbb{E}_{\mu}\left(\bar{X}_n\right) = \mu$,
2. $Var_{\sigma^2}\left(\bar{X}_n\right) = \sigma^2/n$.

Prova: A parte (i) da prova já foi apresentada anteriormente. A prova da parte (ii) é trivial, e será apresentada na sala de aula!!!

Distribuição Amostral de $\bar{X}_n$

- É importante esclarecer que o resultado do Teorema 1 vale para qualquer população associada a v.a. X.
- Numa primeira fase, vamos considerar o caso em que a a.a. é proveniente de uma população normal com média $\mu \in \mathbb{R}$ e variância $\sigma^2 > 0$, sendo os parâmetros de localização e escala, respetivamente. O teorema a seguir estabelece uma distribuição amostrar para $\bar{X}$.

Teorema 2

Seja $X_1, \cdots, X_n$ uma a.a. de tamanho n proveniente de uma população normal, isto é, $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$. Então, segue imediatamente que $\bar{X}_n \sim \mathcal{N}\left(\mu, \sigma^2/n\right)$.

Distribuição Amostral de $\bar{X}_n$

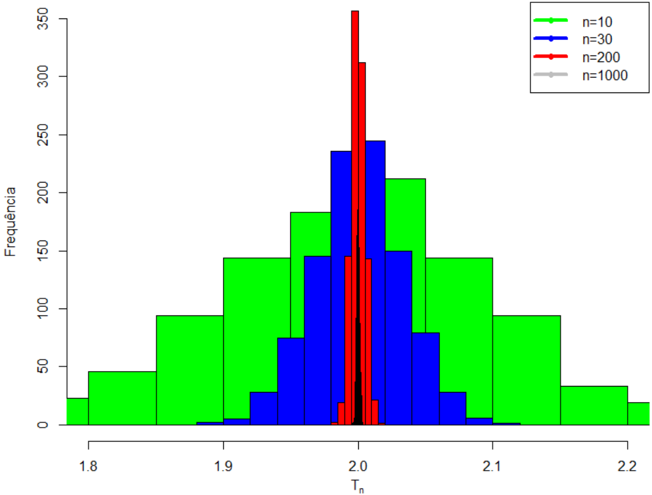

Figura: Distribuição amostral da média $T_n = \bar{X}_n$, para a.a.'s extraídas de uma população $X \sim \mathcal{N}(2, 1)$.

Distribuição Amostral de $\bar{X}_n$

- Os histogramas apresentados na figura anterior, mostram claramente que a variância de $\bar{X}_n$ decresce com o aumento do tamanho de amostra (consequência do fato de que a variância de $\bar{X}_n$ depende do tamanho de amostra).
- Assim, fica evidente que a variância da v.a. X será sempre maior do que a variância de $\bar{X}_n$, $\forall_{n>1}$.
- Encontrada a distribuição de probabilidades de $T_n = \bar{X}_n$, podemos calcular probabilidades do estimador ser maior, menor, ou estar compreendido em um determinado intervalo.

Exemplo 7: A capacidade máxima do um elevador é de $500kg$. Se a distribuição dos pesos dos usuários segue o modelo normal com média $\mu = 70kg$ e variância $\sigma^2 = 100kg^2$, qual é a probabilidade de que 7 pessoas ultrapassem esse limite? E de 6 Pessoas?

Distribuição Amostral de $\bar{X}_n$

- Observe que, a normalidade da distribuição amostral de $\bar{X}_n$ apresentada anteriormente, é exata, no sentido de que $\bar{X}_n$ resulta de uma combinação linear de v.a.'s normais.
- No entanto, na teoria estatística, é comum nos depararmos com a v.a. X segue uma distribuição qualquer (podendo ser próxima ou distante da normal). Assim, como podemos proceder para encontrar a distribuição amostral de $\bar{X}_n$?
- A resposta para essa pergunta está associada ao famoso Teorema Central do Limite (TCL)

Distribuição Amostral de $\bar{X}_n$

Teorema 3 (TCL)

Seja $X_1, \cdots, X_n$ uma a.a. proveniente de uma população F, caracterizada pela média $\mu < \infty$ e variância $\sigma^2 > 0$. Então, a distribuição da média amostral $\bar{X}_n$ converge para uma distribuição normal de média μ e variância σ^2/n, i.e., $\bar{X}_n \overset{a}{\sim} \mathcal{N}\left(\mu, \sigma^2/n\right)$.

- Como consequência do TCL, segue imediatamente que,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \overset{a}{\sim} \mathcal{N}\left(0, 1\right). \tag{1}$$

Prova: Uma das formas de provar a veracidade do TCL, seria por meio da função geradora de momentos (fgm), ou função característica. Porém, os detalhes não serão apresentados aqui.

Distribuição Amostral de $\bar{X}_n$

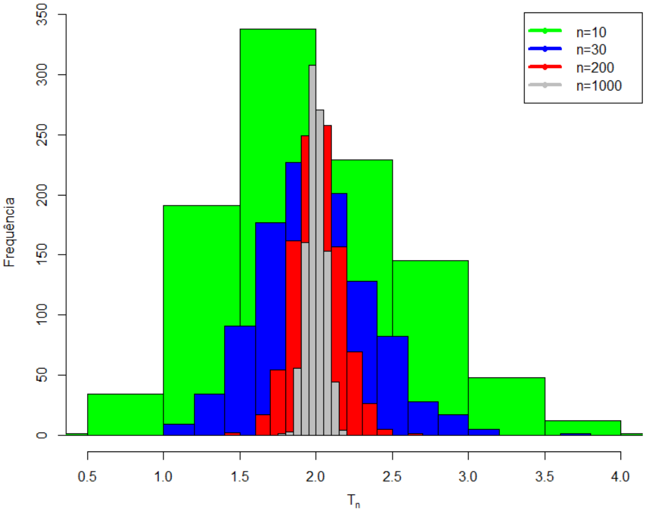

Figura: Aproximação da distribuição exponencial para normal, para $T_n = \bar{X}_n$.

Distribuição Amostral de $\bar{X}_n$

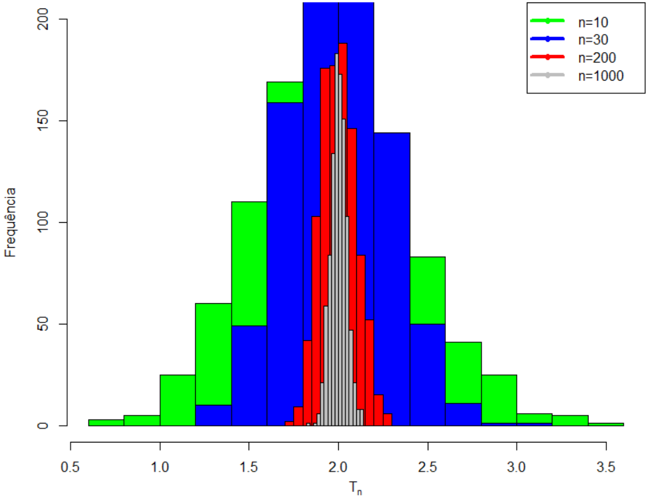

Figura: Aproximação da distribuição Poison para normal, para $T_n = \bar{X}_n$.

Distribuição Amostral de S^2

- Diferentemente do que foi apresentado sobre a distribuição da média amostral $\bar{X}_n$, a obtenção da distribuição da variância amostral S^2 requer que a a.a. seja proveniente de uma população normal.
- Como é de conhecimento, existem dois estimadores para a variância populacional, $\hat{\sigma}_n^2$ e S^2, sendo que, o primeiro é um estimador viesado para σ^2.

Teorema 4

Seja $X_1, \cdots, X_n$ uma a.a. extraída de uma população normal com média $\mu \in \mathbb{R}$ e variância $\sigma^2 > 0$. Assim, a distribuição da variância amostral S^2 é Qui-Quadrado com $n-1$ graus de liberdade.

Distribuição Amostral de S^2

Prova: para provar o teorema 4, defina $U = (n-1)\,S^2/\sigma^2 \sim \chi^2_{n-1}$. De seguida, vamos decompor a quantidade,

$$\underbrace{\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2}_{\sim\chi^2_n} = \sum_{i=1}^{n}\left(\frac{X_i-\bar{X}_n}{\sigma}\right)^2 + \underbrace{\left(\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}\right)^2}_{\sim\chi^2_1}.$$

Consequentemente, $U = (n-1)\,S^2/\sigma^2 = \sum_{i=1}^{n} Z_i^2 - Z_n^2 \sim \chi^2_{n-1}$. O resultado anterior é facilmente obtido usando a *fgm*.

Distribuição Amostral de S^2

- Portanto, em decorrência do resultado anterior, podemos concluir que, a fdp da v.a. U é dada por,

$$f(u|n) = \frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)} u^{\frac{n-1}{2}-1} \exp\left(-\frac{u}{2}\right). \tag{2}$$

- Nisso, segue imediatamente que, $\mathbb{E}(U) = n-1$ e $Var(U) = 2(n-1)$. Ademais, podemos escrever S^2 como, $S^2 = \frac{U\sigma^2}{(n-1)}$. Assim,

$$\begin{aligned}
(i)\ \ \mathbb{E}\left(S^2\right) &= \mathbb{E}\left\{\frac{U\sigma^2}{(n-1)}\right\} = \frac{\sigma^2}{(n-1)}\mathbb{E}(U) = \sigma^2 \\
(ii)\ \ Var\left(S^2\right) &= Var\left\{\frac{U\sigma^2}{(n-1)}\right\} = \frac{\sigma^4}{(n-1)^2}Var(U) = \frac{2\sigma^4}{(n-1)}
\end{aligned}$$

Distribuição Amostral de S^2

Algumas Características da Distribuição χ^2_{n-1}

- A distribuição χ^2_{n-1} é um caso particular da distribuição $\Gamma\left(\frac{n-1}{2}, \frac{1}{2}\right)$.
- É uma distribuição assimétrica positiva.
- Seu aspecto visual depende do tamanho de amostra, que é o parâmetro caracterizador da distribuição.
- É uma distribuição aditiva, no sentido de que, se as v.a.'s são iids $Y_i \sim \chi^2_{n_i}$ então, $\sum_{i=1}^{n} Y_i \sim \chi^2_m$, com $m = \sum_{i=1}^{n} n_i$.
- Se $Z \sim \mathcal{N}(0,1)$, segue imediatamente que, $Z^2 \sim \chi^2_1$.
- A distribuição χ^2_{n-1} se aproxima da distribuição normal que n cresce. Esse é um resultado direto do TCL. Isto é,

$$\frac{Y - (n-1)}{\sqrt{2(n-1)}} \overset{a}{\sim} \mathcal{N}(0,1).$$

Distribuição Amostral da Proporção $\hat{p}_n$

- O TCL anunciado anteriormente afirma que, se X é uma v.a. proveniente de uma população qualquer F, tal que, $\mu = \mathbb{E}(X)$ e $\sigma^2 = Var(X)$, então, a média amostral $\bar{X}_n$ segue uma distribuição normal com média μ e variância σ^2/n.

Teorema 5

Seja $X_1, \cdots, X_n$ uma a.a. proveniente de uma população com fda F. Se $\mu = \mathbb{E}(X)$ e $\sigma^2 = Var(X)$ denotam a média e a variância de X, então, para a estatística $S_n = \sum_{i=1}^{n} X_i = n\bar{X}_n$, podemos estabelecer o seguinte resultado:

i. $\mathbb{E}(S_n) = n\mu$,
ii. $Var(S_n) = n\sigma^2$.

Distribuição Amostral da Proporção $\hat{p}_n$

- Supondo que X_i assume apenas dois resultados possíveis, isto é,

$$X_i = \begin{cases} 1 & \text{com probilidade } p \\ 0 & \text{com probilidade } q \end{cases}$$

então, a estatística S_n denota a contagem do número de sucessos em n repetições independentes do experimento. Assim, $\mathbb{E}\left(X_i\right) = p$ e $Var\left(X_i\right) = p\left(1-p\right)$, segue imediatamente que,

  1. $\mathbb{E}\left(S_n\right) = np$,
  2. $Var\left(S_n\right) = np\left(1-p\right)$.

- Portanto, segue do TCL que, $S_n \sim \mathcal{N}\left(np, np(1-p)\right)$ quando $n \to \infty$. Consequentemente,

$$\frac{S_n - \mathbb{E}\left(S_n\right)}{\sqrt{Var\left(S_n\right)}} = \frac{S_n - np}{\sqrt{np\left(1-p\right)}} \overset{a}{\sim} \mathcal{N}\left(0,1\right).$$

Distribuição Amostral da Proporção $\hat{p}_n$

Atenção: a aproximação da distribuição binomial para a normal será considerada válida quando as seguintes condições forem satisfeitas:

i. $np \geq 5$
ii. $n(1-p) \geq 5$.

- Usando o resultado da aproximação binomial à normal, e usando o fato de que $\hat{p}_n = S_n/n$, segue imediatamente que, $\hat{p}_n \sim \mathcal{N}\left(p, p(1-p)/n\right)$, quando $n \to \infty$.

- Alternativamente, temos que:

$$\frac{S_n/n - \mathbb{E}\left(S_n/n\right)}{\sqrt{Var\left(S_n/n\right)}} = \frac{\hat{p}_n - p}{\sqrt{p\left(1-p\right)/n}} \overset{a}{\sim} \mathcal{N}\left(0,1\right).$$

Com $\mathbb{E}\left(\hat{p}_n\right) = p$ e $Var\left(\hat{p}_n\right) = p\left(1-p\right)/n$.

Distribuição Amostral da Proporção $\hat{p}_n$

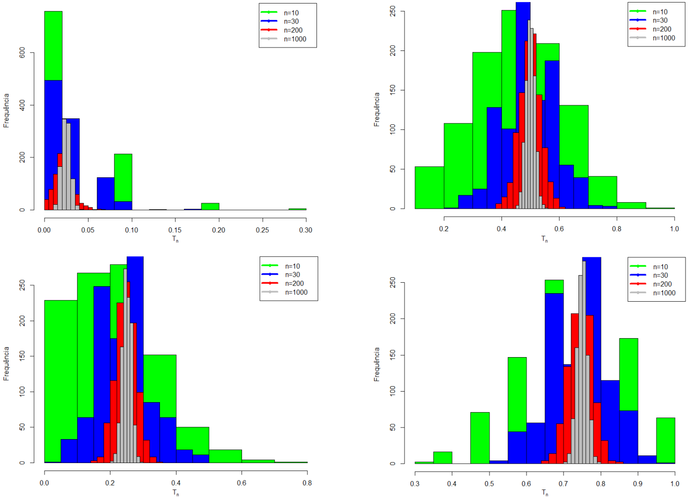

Figura: Distribuição amostral da proporção $T_n = \hat{p}_n$. Painel-1 ($p = 0,025$), Painel-2 ($p = 0,25$), Painel-3 ($p = 0,50$), Painel-4 ($p = 0,75$).

Distribuição Amostral da Proporção $\hat{p}_n$

Exemplo 8: De um grande lote de produtos manufaturados, extrai-se uma a.a. simples de 200 itens. Se 10% dos itens são defeituosos, qual é a probabilidade de serem sorteados no máximo, 24 itens defeituosos?

Exemplo 9: A confiabilidade de um componente é a probabilidade de que ele funcione sob as condições desejadas. Uma amostra a.a. simples de 2500 desses componentes é extraída, e cada componente testado. Pede-se para calcular a probabilidade de se obter pelo menos 75 itens defeituosos, supondo que a confiabilidade seja de:

a) 0,995
b) 0,850.

Distribuição F de Snedecor

- Outra distribuição que tem sido frequentemente considerada na literatura, é a distribuição F de Snedecor.
- Tal distribuição é obtida por meio do quociente entre duas variáveis com distribuição Qui-Quadrado, dividido pelos respectivos graus de liberdade.

Definição 1 (Distribuição F)

Se $X_{11}, \cdots, X_{1n}$ e $X_{21}, \cdots, X_{2m}$ denotam duas a.a.'s retiradas de uma mesma população normal, com média μ e variância σ^2, defina as quantidades, $U = \sum_{i=1}^{n} \left(\frac{X_{1i} - \bar{X}_1}{\sigma}\right)^2 \sim \chi^2_{n-1}$ e $V = \sum_{i=1}^{m} \left(\frac{X_{2i} - \bar{X}_2}{\sigma}\right)^2 \sim \chi^2_{m-1}$. Então, pelo teorema de Snedecor, segue que,

$$W = \frac{U/(n-1)}{V/(m-1)} \sim F_{(n-1,m-1)}.$$

Distribuição F de Snedecor

- Assim, a fdp da v.a. W com distribuição $F_{(n-1,m-1)}$ de Snedecor é dada por:

$$f\left(w|n,m\right)=\frac{\Gamma\left(\frac{n+m-2}{2}\right)\left(\frac{n-1}{m-1}\right)^{\frac{n-1}{2}}w^{\frac{n-1}{2}}}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)\left[\left(\frac{n-1}{m-1}\right)w+1\right]^{(n+m-2)/2}},\ 0<w<\infty.$$

- Segue da definição 9 temos igualmente que,

$$W=\frac{U/(n-1)}{V/(m-1)}=\frac{\frac{(n-1)S_1^2/\sigma^2}{n-1}}{\frac{(m-1)S_2^2/\sigma^2}{m-1}}=\frac{S_1^2}{S_2^2}\sim F_{(n-1,m-1)}.$$

Distribuição F de Snedecor

- A média e variância de uma v.a. com distribuição $F_{(n-1,m-1)}$ são dadas por (os detalhes não serão apresentados aqui):

$$\mathbb{E}(W) = \frac{m-1}{m-3}, \text{ com } m > 3$$

$$Var(W) = \frac{2(m-1)^2(n+m-4)}{(n-1)(m-3)^2(m-5)}, \text{ para todo } n > 1 \; m > 5.$$

- A distribuição F é assimétrica positiva, onde seu aspecto gráfico depende dos valores de n e m.
- Se a v.a. $W \sim F_{(n-1,m-1)}$ então, $\frac{1}{W} \sim F_{(n-1,m-1)}$.

Distribuição t-Student

- Entre as distribuições de probabilidades apresentadas até o momento, a distribuição t-Student é que está mais próxima da distribuição normal. Ou seja, ela é um caso particular da distribuição normal.

Definição 2

Seja Z uma v.a. com distribuição normal padrão, e U uma v.a. com distribuição χ^2_{n-1}. Então, a v.a.

$$Y = \frac{Z}{\sqrt{\frac{U}{n-1}}} = \frac{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{U}{n-1}}} \sim t_{(n-1)}. \tag{3}$$

Distribuição t-Student

- A fdp de uma v.a. com distribuição t-Student pode ser derivada usando o método de Jacobiano, ou mesmo com base na fgm (os detalhes não serão apresentados aqui). Assim,

$$f_Y\left(y|n\right) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi\left(n-1\right)}}\left(1+\frac{y^2}{n-1}\right)^{-n/2},$$

  com $n-1$ denotando o número de graus de liberdade da distribuição t-Student.

- Os principais momentos da distribuição t-Student são:

$$\begin{aligned} \mathbb{E}\left(Y\right) &= 0. \\ Var\left(Y\right) &= \frac{n-1}{n-3}, \text{ com } n > 3. \end{aligned}$$

  Note que, $Var\left(Y\right) \to 1$, quando $n \to \infty$.

Distribuição t-Student

Definição 3

Se $X_1 \cdots, X_n$ denota uma a.a. proveniente de uma distribuição normal de média μ e variância σ^2. Ademais, vimos que $U \sim \chi^2_{n-1}$, portanto, usando o fato de que $\bar{X}_n$ e $(n-1)\,S^2$ são independentes (fato que não será provado aqui), então, segue que:

$$Y = \frac{Z}{\sqrt{\frac{U}{n-1}}} = \frac{\left(\bar{X}_n - \mu\right)}{S/\sqrt{n}} \sim t_{n-1}. \tag{4}$$

- Ou seja, nos casos em que a amostra é proveniente de uma população normal, porém, com σ^2 desconhecido a quantidade acima segue uma distribuição t-Student com $n-1$ graus de liberdade.

Distribuição t-Student

Algumas Característica da Distribuição t-Student

a. A distribuição t-Student é simétrica em de zero, isto é, $\mathbb{E}\left(T\right) = 0$.
b. Valores menores de graus de liberdade levam a distribuição t-Student com caudas mais pesadas e a intervalos de confiança maiores.
c. Quando o número de graus de liberdade aumenta a distribuição t-Student se aproxima da distribuição $\mathcal{N}\left(0, 1\right)$.
d. O quadrado de uma v.a. Y com distribuição t de Student com $n - 1$ graus de liberdade tem uma distribuição F de Snedecor com 1 e $n - 1$ graus de liberdade. Isto é, se $Y \sim t_{n-1}$, então, $Y^2 \sim F_{(1,n-1)}$.

Distribuição t-Student

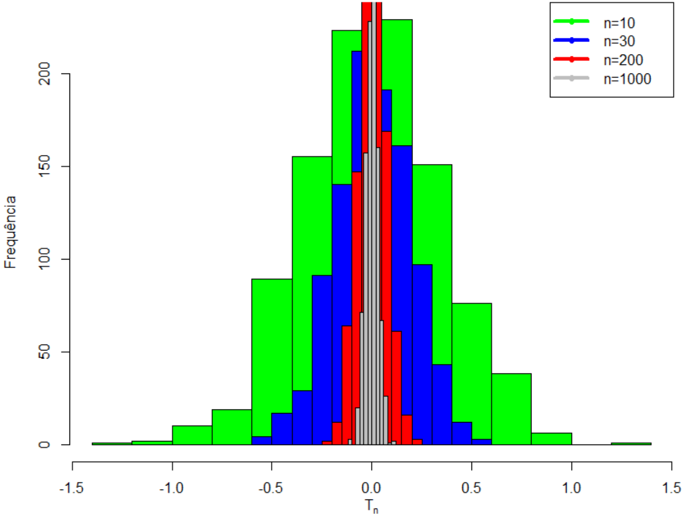

Figura: Formas da distribuição t-Student quando o número de graus de liberdade aumenta.

Distribuição Amostral de $\bar{X}_{1n} - \bar{X}_{2m}$

- Nas aulas anteriores abordamos a distribuição da média amostral obtida de uma única população.
- Em alguns casos, é de suma importância para o pesquisador comparar (ou fazer inferência) com o intuito de obter algumas conclusões sobre os parâmetros de duas ou mais populações.
- Neste tópico vamos derivar as distribuições amostrais da diferença de médias baseadas em a.a.'s provenientes de populações diferentes. Para tal, vamos considerar *três casos principais*.

Distribuição Amostral de $\bar{X}_{1n} - \bar{X}_{2m}$

Caso 1

Assuma que as duas a.a. $X_{11}, \cdots, X_{1n}$ e $X_{21}, \cdots, X_{2n}$ são provenientes de duas populações normais caracterizadas pelos parâmetros de média μ_1 e μ_2, e de variância σ_1^2 e σ_2^2, isto é, $X_{1i} \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right)$ e $X_{2i} \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right)$, respectivamente.

- Seja $\bar{X}_{1n}$ e $\bar{X}_{2m}$ as médias obtidas por meio das duas a.a.'se selecionadas nas duas populações. Supondo que as variâncias populacionais são conhecidas, mas com $\sigma_1^2 \neq \sigma_2^2$, segue que, $\bar{X}_{1n} - \bar{X}_{2m} \sim \mathcal{N}\left(\mu_1 + \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$.

Distribuição Amostral de $\bar{X}_{1n} - \bar{X}_{2m}$

- Ou seja, pelo teorema de aditividade da distribuição normal, segue que,

$$\frac{\left(\bar{X}_{1n} - \bar{X}_{2m}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim \mathcal{N}\left(0, 1\right),$$

  com n e m denotando os tamanhos amostrais, que podem ser iguais ou diferentes.

- Observe que, se as a.a.'s forem provenientes de duas populações quaisquer, com $n, m \geq 30$, o TCL pode ser aplicado para obter a normalidade assintótica da quantidade acima.

Distribuição Amostral de $\bar{X}_{1n} - \bar{X}_{2m}$

Caso 2

Neste caso, vamos assumir que as a.a.'s $X_{11}, \cdots, X_{1n}$ e $X_{21}, \cdots, X_{2n}$ são provenientes de duas populações normais, $X_{1i} \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right)$ e $X_{2i} \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right)$ com $\sigma_1^2 \neq \sigma_2^2$, sendo ambas desconhecidas. Neste caso, vamos usar os estimadores não-viesados para σ_1^2 e σ_2^2, respectivamente. Assim,

$$\frac{\left(\bar{X}_{1n} - \bar{X}_{2m}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}} \sim t_{(v)},$$

com $v = \frac{\left(\frac{S_1^2}{n} + \frac{S_2^2}{m}\right)^2}{\left(\frac{S_1^2}{n}\right)^2/(n-1) + \left(\frac{S_2^2}{m}\right)^2/(m-1)}$, denotando o número de graus de liberdade.

- Para $n, m \geq 30$, a quantidade acima segue uma distribuição normal padrão.

Distribuição Amostral de $\bar{X}_{1n} - \bar{X}_{2m}$

Caso 3

Supondo agora que as variâncias populacionais σ_1^2 e σ_2^2 são ambas desconhecidas, mas que as duas populações apresentam a mesma variabilidade, isto é, $\sigma_1^2 = \sigma_2^2$, segue imediatamente que, se $n, m < 30$ então,

$$\frac{\left(\bar{X}_{1n} - \bar{X}_{2m}\right) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim t_{(n+m-2)},$$

onde S_p^2 denota a variância ponderada, cuja sua expressão é dada por:

$$S_p^2 = \frac{(n-1)\,S_1^2 + (m-1)\,S_2^2}{n+m-2}.$$

Distribuição Amostral de $\hat{p}_{1n} - \hat{p}_{2m}$

- Considere duas populações Bernoulli, caracterizadas pelos parâmetros P_1 e P_2, que denotam as verdadeiras proporções de ocorrência do sucessos nas duas populações.
- Em muitas situações práticas, é de suma importância fazer inferência em duas ou mais populações independentes. Como por exemplo:
  - *i* Podemos estar interessados em comparar a proporção de consumidores interessados em um novo produto na área rural, com a da área urbana.
  - *ii* Podemos estar interessados em inferir as proporções de respostas favoráveis a uma campanha publicitária feita em dois seminários diferentes, entre outros exemplos.
- Nos exemplos supracitados, pretende-se concluir algo sobre a quantidade $p_1 - p_2$ que se desconhece.

Distribuição Amostral de $\hat{p}_{1n} - \hat{p}_{2m}$

Definição 4

Considere $X_{11}, \cdots, X_{1n}$ *e* $X_{21}, \cdots, X_{2m}$ *duas a.a. provenientes de duas populações de Bernoulli,isto é,* $\mathcal{B}er\,(p_1)$ *e* $\mathcal{B}er\,(p_2)$*, respectivamente. Se* $S_n = \sum_{i=1}^{n} X_{1i}$ *e* $S_m = \sum_{i=1}^{m} X_{2i}$ *denotam o número de sucessos em experimentos de Bernoulli independentes, as proporções amostrais são dadas por:*

$$\hat{p}_{1n} = \frac{S_n}{n} \ \ e \ \ \hat{p}_{2m} = \frac{S_m}{m}$$

Com base nos resultados da Definição 12, e do TCL, segue imediatamente que, $\hat{p}_{1n} \sim \mathcal{N}\left(p_1, p_1\left(1 - p_1\right)/n\right)$, e $\hat{p}_{2m} \sim \mathcal{N}\left(p_2, p_2\left(1 - p_2\right)/m\right)$, quando $n, m \to \infty$.

Distribuição Amostral de $\hat{p}_{1n} - \hat{p}_{2m}$

- A consequência direta do resultado anterior, é:

$$\hat{p}_{1n} - \hat{p}_{2m} \sim \mathcal{N}\left(p_1 - p_2, \frac{p_1\left(1 - p_1\right)}{n} + \frac{p_2\left(1 - p_2\right)}{m}\right), \; n, m \to \infty.$$

- De igual forma, a normalidade acima implica que,

$$\frac{\left(\hat{p}_{1n} - \hat{p}_{2m}\right) - \left(p_1 - p_2\right)}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}} \sim \mathcal{N}\left(0, 1\right) \text{ quando } n, m \to \infty. \tag{5}$$

- Assim, o resultado em (5) estabelece a distribuição amostrar para a diferença de proporções amostrais.

Distribuições Amostrais dos Extremos

- Nesta parte da disciplina vamos apresentar a noção (ou conceito) de estatísticas de ordem. Como por exemplo, a menor e maior observação de um conjunto de dados, respectivamente.
- O processo que consiste em obter as estatísticas estremais é muito fácil. Porém, para obter a distribuição amostral associada a tais estatísticas exige um trabalho considerável.
- Portanto, a obtenção das distribuições associadas às estatísticas extremais é por vezes importante na literatura estatística.

Definição 5

Seja $X_1, \cdots, X_n$ uma a.a. de X, extraída de uma população que é caracterizada pela fdp $f_X(x)$.

Distribuições Amostrais dos Extremos

- Se $X_1, \cdots, X_n$ denota a a.a. da definição anterior, considere $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ os valores ordenados da a.a. (ou estatísticas de ordem). Tal que, $X_{(j)}$ para todo $1 \leq j \leq n$ a j-ésima estatística de ordem.
- Desta forma, $X_{(1)} = \min\{X_1, \cdots, X_n\}$ denota a menor observação, e $X_{(n)} = \max\{X_1, \cdots, X_n\}$ a maior observação da a.a.
- A distribuição amostral para qualquer estatística de ordem $X_{(j)}$ pode ser obtida. No entanto, nessa disciplina nos restringiremos apenas nas distribuições das estatísticas extremais, $X_{(1)}$ e $X_{(n)}$.

Distribuições Amostrais dos Extremos

Distribuição do Máximo: antes de procedemos com derivação da fda de uma estatística de ordem, considere $F(x)$, tal que,

$$F_X(x) = P(X \leq x) = \int_0^x f_X(y)dy$$

- Assim, nosso interesse reside em obter a fda e fdp associada às estatísticas extremais. Assim, a distribuição do máximo é dada por:

$$\begin{aligned} F_{X_{(n)}}(x) &= P\left(X_{(n)} \leq x\right) = P\left(\max\{X_1, \cdots, X_n\} \leq x\right) \\ &= P\left(X_1 \leq x, \cdots, X_n \leq x\right) = \left[P\left(X_1 \leq x\right)\right]^n \\ &= \left[F_X(x)\right]^n. \end{aligned}$$

- Consequentemente, a fdp do máximo é dada por:

$$f_{X_{(n)}}(x) = \frac{d}{dx}F_{X_{(n)}}(x) = nf_X(x)\left[F_X(x)\right]^{n-1}.$$

Distribuição do Mínimo: usando a mesma abordagem apresentada anteriormente, podemos derivar facilmente a fda e fdp do mínimo. Para tal,

$$F_{X_{(1)}}(x) = P\left(X_{(1)} \leq x\right) = 1 - P\left(X_{(1)} > x\right) \tag{6}$$

Observe agora que,

$$\begin{aligned} P\left(X_{(1)} > x\right) &= P\left(\min\left\{X_1, \cdots, X_n\right\} > x\right) \\ &= P\left(X_1 > x, \cdots, X_n > x\right) = \left[P\left(X_1 > x\right)\right]^n \\ &= \left[1 - F_X(x)\right]^n. \end{aligned}$$

Distribuições Amostrais dos Extremos

- Com base na expressão em (6), obtemos a fda de $X_{(1)}$ representada por:

$$F_{X_{(1)}}(x) = P\left(X_{(1)} \leq x\right) = 1 - P\left(X_{(1)} > x\right) = 1 - [1 - F_X(x)]^n$$

- Por fim, a fdp do mínimo, $X_{(1)}$ tem o seguinte formato:

$$f_{X_{(1)}}(x) = \frac{d}{dx} F_{X_{(1)}}(x) = n f_X(x) \left[1 - F_X(x)\right]^{n-1}.$$

- É importante salientar que, uma expressão geral para computar a distribuição amostral da j-ésima estatística de ordem é estabelecida pelo teorema abaixo.

Distribuições Amostrais dos Extremos

Teorema 6

Sejam $X_{(1)} < \cdots < X_{(n)}$ estatísticas de ordem obtidas por meio de uma a.a. extraída de uma população caracterizada pela fdp $f_X(x)$ e fda $F_X(x)$. Então, a fdp da j-ésima estatística de ordem, $X_{(j)}$ é dada por:

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!\,(n-j)!} f_X(x) \left\{F_X(x)\right\}^{j-1} \left\{1 - F_X(x)\right\}^{n-j},$$

com $1 \leq j \leq n$.

Exemplo 10: Seja X uma v.a. extraída de uma população com distribuição uniforme, isto é, $X \sim \mathcal{U}\,(10; 20)$. Assuma que X_1, X_2, X_3, X_4, X_5 denote uma a.a. selecionada dessa mesma população. Assim, pede-se para:

a. Calcular a probabilidade de que o mínimo da amostra seja maior que 12.
b. Calcular a probabilidade de que o máximo da amostra seja menor que 15.