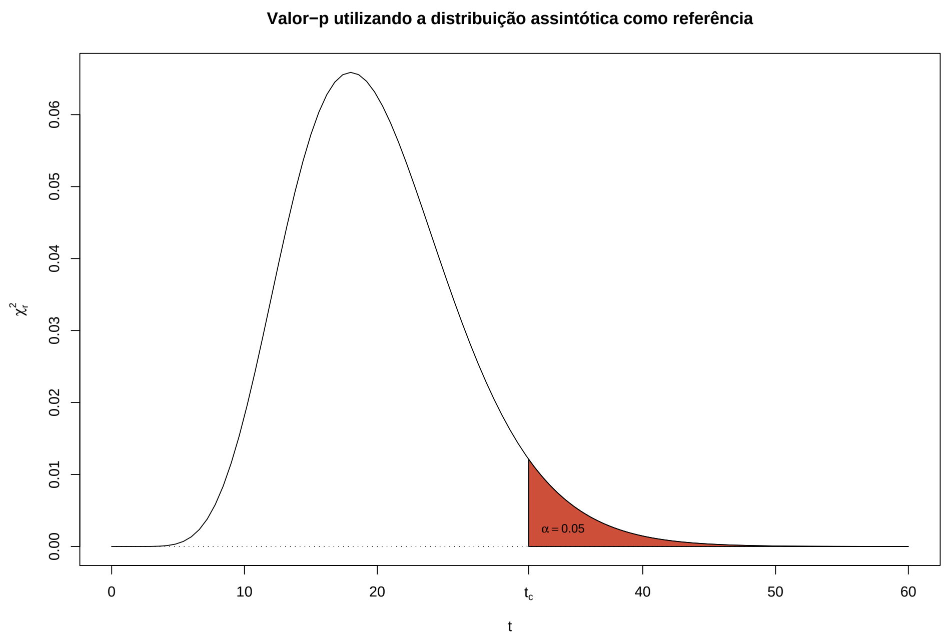
# Verificando se o tamanho amostral é suficiente para aproximações assintóticas

Alexandre Galvão Patriota

Todos os modelos estatísticos possuem suposições que devem ser satisfeitas. Nos últimos anos, um grande esforço vem sendo empregado a fim de flexibilizar tais suposições para atender com maior precisão a situações reais. Entretanto, tais flexibilizações tornam os modelos estatísticos mais complexos ficando inviável resolver o problema apenas com papel e caneta. Com o advento dos computadores com processadores rápidos tornou-se, de certa forma, simples (usando algoritmos iterativos) estimar os parâmetros de um modelo estatístico complexo.

Geralmente o pesquisador tem uma hipótese de interesse e o estatístico precisa bolar uma estatística para testar esta hipótese. (Neste post será considerado apenas modelos frequentistas.). Em um modelo muito complexo, não é possível encontrar a distribuição exata desta estatística proposta e portanto métodos assintóticos são utilizados. Este método consiste basicamente em encontrar a distribuição desta estatística proposta quando o tamanho da amostra tende ao infinito. Assim teremos uma distribuição assintótica que pode ser utilizada como aproximação em situações reais, ou seja, quando o tamanho amostral é finito. Mas qual o tamanho amostral razoável para que a aproximação seja coerente, n=10, 50, 1000?

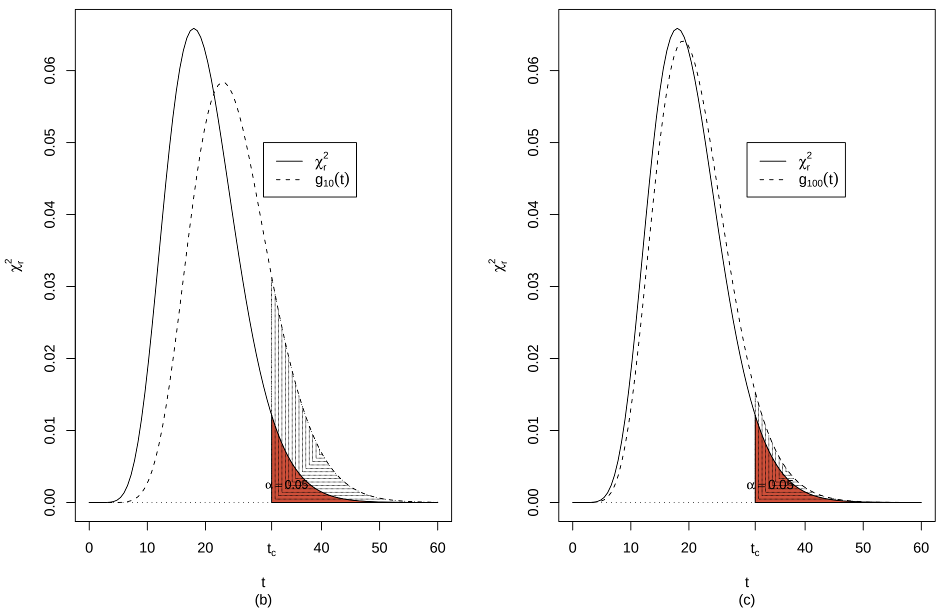
Suponha que $\theta$ seja a quantidade de interesse. Assuma que a hipótese nula é $H_0 : \theta = \theta_0$ contra $H_0 : \theta \neq \theta_0$. Seja $t_n$ a estatística proposta para testar a hipótese nula. Suponha que, sob $H_0$, a distribuição assintótica de $t_n$ seja uma qui-quadrado com $r$ graus de liberdade, $\chi^2(r)$. Note que $\chi^2(r)$ não é a verdadeira distribuição de $t_n$, mas ela pode ser utilizada como aproximação. Dessa forma, podemos calcular os quantis da distribuição $\chi^2(r)$ e verificar se o valor da estatística observada $t_n$ está numa região crítica. A Figura (a) nos mostra a área referente a $\alpha = 0,05$ utilizando a distribuição assintótica como referência, o valor $t_c$ é o valor crítico. Dessa forma, se $t_n < t_c$ aceitamos a hipótese $H_0$ (ou seja, hão há evidências para rejeitar a hipótese nula). Se $t_n \geq t_c$ rejeitamos a hipótese nula a 5% de significância.




(a)

Perceba que a distribuição $\chi^2(r)$ é apenas uma aproximação para a verdadeira distribuição de $t_n$. Considere que $g_n(t)$ seja a verdadeira distribuição de $t_n$, obviamente $g_n(t)$ converge para $\chi^2(r)$ quando $n$ tende ao infinito. As figuras (b) e (c) mostram uma possível situação quando a amostra é $n = 10$ e $n = 100$, respectivamente. Note que se $n = 10$ a distribuição verdadeira, $g_n(t)$, e a assintótica, $\chi^2(r)$, estão bem distantes, enquanto que quando $n = 100$ estas distribuições estão mais próximas. Uma forma de verificar se estas distribuições estão próximas é simulando a taxa de rejeição sob a hipótese nula. Ou seja, simulamos os dados sob a hipótese nula e calculamos a estatística $t_n$, depois verificamos se rejeitamos ou aceitamos a hipótese nula usando o valor crítico calculado da distribuição assintótica, $t_c$.

A área alaranjada das figuras (a), (b) e (c) indicam a probabilidade do Erro tipo I que o pesquisador tolera em um teste de hipótese. Este erro consiste em rejeitar a hipótese nula quando esta é verdadeira. Note que na figura (b) pensamos que a probabilidade do erro tipo I é 5%, contudo a verdadeira distribuição, $g_{10}(t)$, da estatística proposta para n=10 nos dá uma área muito maior do que 5% (veja a área hachurada abaixo da curva $g_{10}(t)$). Casos como este podem invalidar conclusões de estudos impor-




tantes, portanto é necessário muita cautela. Se a taxa de rejeição simulada estiver muito distante do nível nominal de significância, há forte evidência de que o tamanho amostral não é adequado e outras metodologias devem ser adotadas (correção de barttlet, testes utilizando técnicas de bootstrapping, etc.). Pode-se olhar também para o nível descritivo simulado sob a hipótese nula, que deve ter uma distribuição uniforme.

*Pseudo-código para verificar se o tamanho amostral é suficiente para aproximações assintóticas:*

Estou assumindo que você tenha um conjunto de dados de tamanho n.

1. Gere dados considerando todas as suposições do seu modelo sob a hipótese nula. Os parâmetros livres (que não estão definidos na hipótese nula) devem ser substituídos pela estimativa de máxima verossimilhança ou outra que seja consistente.

2. Calcule a sua estatística $t_n$, compare com o nível crítico calculado usando a distribuição assintótica (calculado para um $\alpha$ fixado) e armazene o número 1 se houve rejeição e 0 (zero) caso contrário.

3. Repita os passos 1. e 2. até atingir um número N de vezes (normalmente admite-se $N = 15000$).

No final você terá um vetor de zeros e uns, basta tirar a média e verificar se ela está próxima do nível nominal adotado $\alpha$. Pode-se armazenar o nível descritivo do teste em cada simulação, sob a hipótese nula a distribuição do nível descritivo deve ser uma uniforme (no final da simulação basta fazer um histograma do vetor de níveis descritivos para se ter uma idéia de quão longe você está da distribuição limite). Outra forma seria armazenar o valor da própria estatística e compara-la com a distribuição limite.