Professor: Fábio Kravetz

Leia atentamente as instruções:

- Responder a todas as questões apresentadas abaixo. O objetivo da atividade em questão é a consolidação da compreensão sobre o conteúdo.
- As soluções devem ser realizadas de maneira e escrita e com a linguagem R.

1. Uma **organização de pesquisa de mercado** está planejando dois estudos para avaliar a aceitação de um novo conceito de aplicativo em diferentes bases de usuários. Para garantir a validade estatística e comparabilidade, ambos os projetos devem seguir os mesmos critérios:

   I. Nível de confiança: 95%;
   II. Erro máximo: 4%;

População finita:

N: Tamanho da população;

Z: Valor crítico da distribuição normal;

p: Probabilidade estimada do evento. É definida como 0,5 caso esta não seja conhecida antes de realizar a amostragem;

e: Erro amostral desejado;

n: Tamanho da amostra.

O Valor Crítico (Z) é fundamental porque ele **define as fronteiras** para decisões estatísticas, estabelecendo os limites do intervalo de confiança. Ele separa a área central da curva (correspondente ao nível de confiança) das áreas extremas (as "caudas").

Z = 1,64 – nível de confiança igual a 90%

Z = 1,96 – nível de confiança igual a 95%

Z = 2,58 – nível de confiança igual a 99%

$$n = \frac{N \cdot Z^2 \cdot p \cdot (1-p)}{\left(N \cdot e^2 + Z^2 \cdot p \cdot (1-p)\right)}$$

$$n = \frac{2500 \cdot 1{,}96^2 \cdot 0{,}5 \cdot (1-0{,}5)}{\left(2500 \cdot 0{,}04^2 + 1{,}96^2 \cdot 0{,}5 \cdot (1-0{,}5)\right)}$$

$$n = 484$$

População infinita:

$$n = \frac{Z^2 \cdot p \cdot (1-p)}{e^e}$$
$$n = \frac{1{,}96^2 \cdot 0{,}5 \cdot (1-0{,}5)}{0{,}04^2}$$
$$n = 600{,}25$$

Utilizando a linguagem R, implemente um script que:

a) Calcule e exiba o tamanho da amostra necessário para o estudo com um grupo de usuários de testes que corresponde a uma população finita de 2.500 indivíduos.
b) Calcue e exiba o tamanho da amostra necessário para o estudo com o público-alvo geral, que é considerado uma população infinita.

2. Explique o código abaixo de forma detalhada. O que tal código está implementando?

```
pop_total <- 5000
nivel_significancia <- 1 - 0.90
erro_maximo <- 0.03
prob_p <- 0.5
prob_q <- 1 - prob_p
valor_critico_z <- qnorm(1 - nivel_significancia/2)

num_amostra <- pop_total * (valor_critico_z^2) * prob_p * prob_q
den_amostra <- (erro_maximo^2) * (pop_total - 1) + (valor_critico_z^2) * prob_p * prob_q


n_calculado <- num_amostra / den_amostra

n_selecao <- ceiling(n_calculado)
print(n_selecao)

ids_unidades <- 1:pop_total

intervalo_k <- floor(pop_total / n_selecao)
print(intervalo_k)

set.seed(456)

ponto_partida_r <- sample(1:intervalo_k, 1)
print(ponto_partida_r)

amostra_final <- seq(from = ponto_partida_r,
              to = pop_total,
              by = intervalo_k)

amostra_final <- amostra_final[1:n_selecao]
print(amostra_final)
print(length(amostra_final))
```

3. Uma rede de postos de gasolina afirma que, em seus estabelecimentos não se vende gasolina adulterada. Sabe-se que, de acordo com os padrões de qualidade, a gasolina não pode conter mais de 240 ml de álcool por litro. O órgão de fiscalização colheu 25 medições do produto nos postos dessa rede, obtendo a partir delas um a média de 240,75 ml de álcool/litro. Admitindo-se que a quantidade de álcool presente na gasolina tem uma distribuição normal com desvio-padrão de 2,5 ml/litro. Calcule o valor do teste t.

Resolução

1. Hipóteses:

H_0: $\mu \leq 240$ (A gasolina não está adulterada).
H_1: $\mu > 240$ (A gasolina está adulterada).

2. Dados:

Média da amostra ($\bar{x}$): 240,75;
Média de referência (μ_0): 240
Desvio padrão da amostra (s): 2,5 ml/litro
Tamanho da amostra (n): 25
Nível de significância (α): 0,05

3. Cálculo da estatística T: **(UMA AMOSTRA)**

A partir da fórmula abaixo:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{240,75 - 240}{2,5/\sqrt{25}} = \frac{0,75}{2,5/5} = \frac{0,75}{0,5} = 1,5$$

4. Em relação a questão abaixo, identifique se o teste de hipótese a ser conduzido será Unicaudal à Esquerda, Unicaudal à Direita ou Bicaudal. Apresente às figuras relativas aos testes descritos.
   I. Uma linha de produção de pneus é considerada eficiente se a duração média de seus pneus não for inferior a 45.000 km.
   II. O departamento de controle de qualidade deseja saber se o peso médio de um lote de barras de cereal é diferente de 100 gramas, conforme especificado na embalagem.
   III. Uma nova estratégia de *marketing* só será aprovada se demonstrar que a taxa de conversão do *e-commerce* é superior a 5%.

   Assinale a alternativa que corresponde, na ordem, à classificação correta dos testes:

a) **Unicaudal à Esquerda, Bicaudal e Unicaudal à Direita.**
b) Unicaudal à Direita, Bicaudal e Unicaudal à Esquerda.
c) Bicaudal, Unicaudal à Esquerda e Unicaudal à Direita.
d) Unicaudal à Esquerda, Unicaudal à Direita e Bicaudal.
e) Bicaudal, Unicaudal à Direita e Unicaudal à Esquerda.

Uma linha de produção de pneus é considerada eficiente se a duração média de seus pneus não for inferior a 45.000 km. (Unicaudal à Esquerda).

A frase que define o teste é "...não for inferior a 45.000 km". O teste foca em verificar se a média cai abaixo do valor de referência (45.000 km).

H_0: $\mu \geq$ 45.000 km (Não inferior)
H_1: $\mu <$ 45.000 km (Inferior)

Portanto, o sinal de menor que (<) na hipótese alternativa concentra a região crítica na cauda inferior (esquerda) da distribuição.

O departamento de controle de qualidade deseja saber se o peso médio de um lote de barras de cereal é diferente de 100 gramas, conforme especificado na embalagem. (Bicaudal).

A frase que define o teste é "...seja diferente de 100 gramas". O teste foca em verificar qualquer desvio em relação ao valor de referência (100 gramas), seja para mais ou para menos.

H_0: $\mu =$ 100 gramas
H_1: $\mu \neq$ 100 gramas

Portanto, o sinal de diferente (≠) na hipótese alternativa divide a região crítica em ambas as caudas (superior e inferior) da distribuição.

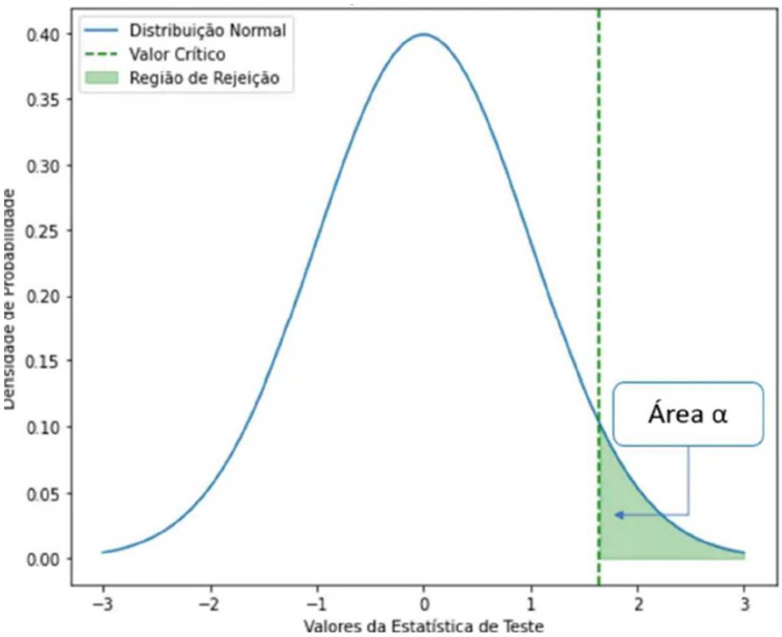
Uma nova estratégia de *marketing* só será aprovada se demonstrar que a taxa de conversão do *e-commerce* é superior a 5%. (Unicaudal à Direita).

A frase que define o teste é "...seja superior a 5%". O teste foca em verificar se o valor da taxa supera o limite de 5%.

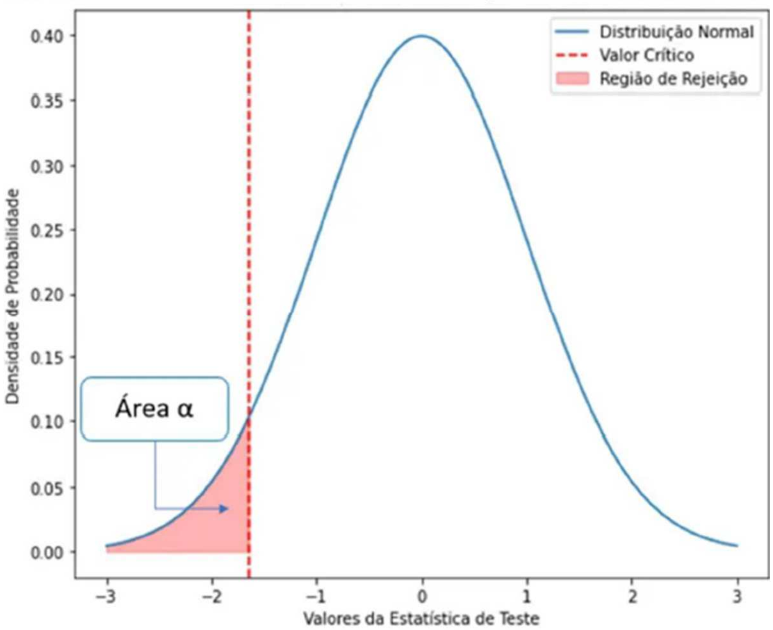
H_0: taxa $\leq$ 5%
H_1: taxa $>$ 5%

Portanto, o sinal de maior que (>) na hipótese alternativa concentra a região crítica na cauda superior (direita) da distribuição.

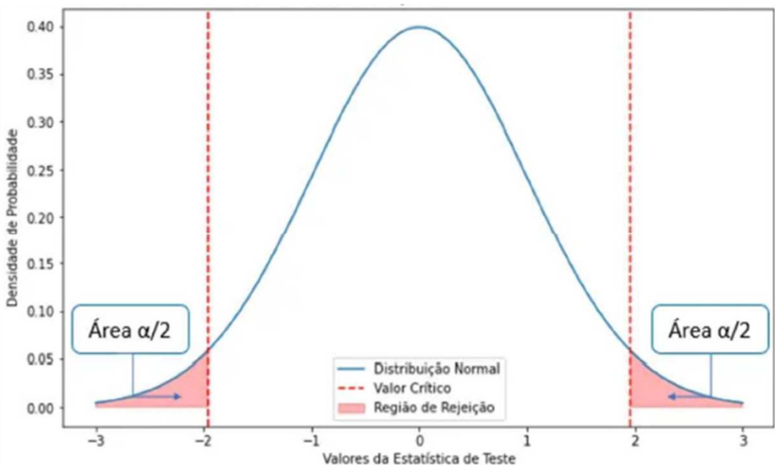
Teste Unicaudal à Direita



Teste Unicaudal à Esquerda



Teste Bicaudal



5. Uma rede supermercados realizou um estudo para comparar o tempo que os clientes levam na fila em caixas operados por humanos (Grupo 1) e caixas de autoatendimento (Grupo 2). Os analistas assumiram que a variabilidade no tempo de espera é a mesma para ambos os tipos de caixa.
   A amostra dos caixas operados por humanos (Grupo 1) consistiu em 30 observações, com um tempo médio de espera de 4.5 minutos. Para os caixas de autoatendimento (Grupo 2), a amostra teve 35 observações, com um tempo médio de 4.9 minutos. A análise combinada resultou em uma variância agrupada de 0.8. Deste modo, informe as hipóteses nula e alternativa e calcule a estatística de teste para determinar se existe uma diferença estatisticamente significativa no tempo médio de espera entre os dois tipos de caixas.

Resolução

1. Hipóteses:

H0: $\overline{x1} = \overline{x2}$ 240 (O tempo médio de espera é igual nos dois grupos).
H1: $\overline{x1} \neq \overline{x2}$ (O tempo médio de espera é diferente nos dois grupos).

2. Dados:

Média da amostra – Grupo 1 ($\overline{x1}$): 4,5;

Média da amostra – Grupo 2 ($\overline{x2}$): 4,9;

Tamanho da amostra (n_1): 30;

Tamanho da amostra (n_2): 35;

Variância agrupada: 0,08

3. Cálculo da estatística T: **(DUAS AMOSTRAS INDEPENDENTES – VARIÂNCIAS HOMOGÊNEAS)**

A partir da fórmula abaixo:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{4,5 - 4,9}{\sqrt{\frac{0,8}{30} + \frac{0,8}{35}}} = -1,7978$$

Quando a variância é considerada homogênea, o teste estatístico apropriado é o Teste t de Student Clássico (Apresentar aos alunos).

Graus de liberdade – Variação Homogênea

$$df = (n_1 - 1) + (n_2 - 1)$$
$$df = n_1 + n_2 - 2$$

6. Um laboratório farmacêutico comparou a eficácia de dois componentes (Composto Alpha e Composto Beta) na redução de uma substância no sangue. O pesquisador assumiu que a variabilidade da resposta aos dois compostos é diferente.
   Para o Composto Alpha (Grupo 1), uma amostra de 15 pacientes mostrou uma redução média de 1,8%, com uma variância de 0,45. Para o Composto Beta (Grupo 2), uma amostra de 18 pacientes mostrou uma redução média de 1,3%, com uma variância de 0,15. Determine a estatística de teste apropriada para verificar se a redução média de substância é diferente entre o Composto Alpha e o Composto Beta.

Resolução

1. Hipóteses:

H_0: $\overline{x1} = \overline{x2}$ (A eficácia na redução média de uma substância no sangue é igual para os dois compostos).
H_1: $\overline{x1} \neq \overline{x2}$ (A eficácia na redução média de uma substância no sangue é diferente para os dois compostos).

2. Dados:

Média da amostra – Grupo 1 ($\overline{x1}$): 1,8;
Média da amostra – Grupo 2 ($\overline{x2}$): 1,3;
Tamanho da amostra (n_1): 15;
Tamanho da amostra (n_2): 18;
Variância – Grupo 1: 0,45;
Variância – Grupo 2: 0,15.

3. Cálculo da estatística T: **(DUAS AMOSTRAS INDEPENDENTES – VARIÂNCIAS HETEROGÊNEAS)**

A partir da fórmula abaixo:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{1,8 - 1,3}{\sqrt{\frac{0,45}{15} + \frac{0,15}{18}}} = 2,553$$

A questão 6 aborda a premissa de homogeneidade de variâncias necessária antes de aplicar um teste t para comparar as médias de dois grupos (Composto Alpha e Composto Beta).

O teste utilizado para verificar essa premissa é o Teste de Bartlett. O Teste de Bartlett verifica se as variâncias dos grupos são estatisticamente iguais:

- Hipótese Nula (H_0): As variâncias são iguais em todos os grupos;
- Hipótese Alternativa (H_1): Pelo menos um grupo tem variância diferente dos demais.

Se p-valor < nível de significância (0,05) a hipótese nula é rejeitada, ou seja, existe evidência estatística de que as variâncias dos compostos são diferentes (Heterogêneas).

Portanto, a rejeição de H_0 no Teste de Bartlett indica que a premissa de variâncias iguais para o Teste t padrã foi violada. O teste correto para este caso é o Teste de Welch que não assume variâncias iguais (variâncias heterogêneas).

O Teste t é a estatística principal para comparar a média entre dois grupos. O teste de Welch corrige o erro dos graus de liberdade quando as variâncias são diferentes. Deste modo, tem-se que o grau de liberdade ajustado, geralmente, não é um número inteiro.

Os graus de liberdade para o Teste de Welch é dado por:

$$df = \frac{\left(\frac{s_1}{n_1} + \frac{s_2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)}{n_2 - 1}}$$

7. Uma universidade implementou um programa de reforço de interpretação de texto e deseja saber se ele realmente aumenta o desempenho dos alunos. Eles aplicaram um teste de interpretação de texto antes do programa e depois do programa em um grupo de 10 alunos.
   A análise das diferenças de pontuação (Pós-Programa menos Pré-Programa) para cada aluno indicou que a média dessas diferenças foi de 2,5 pontos, e o desvio padrão das diferenças foi de 1,2 pontos. Calcule a estatística de teste t.

Resolução

1. Hipóteses:

H_0: $\bar{d} = 0$ (A produtividade média não mudou após o treinamento).
H_1: $\bar{d} \neq 0$ (A produtividade média mudou após o treinamento).

2. Dados:

Média das diferenças ($\bar{d}$): 2,5;

Desvio padrão: 1,2;

Tamanho da amostra: 10;

3. Cálculo da estatística T: **(TESTE T PAREADO)**

A partir da fórmula abaixo:

$$t = \frac{\bar{d}}{s/\sqrt{n}} = \frac{2,5}{1,2/10} = 6,588$$

8. Considere, o código em linguagem R abaixo, referentes as quantidades de mercúrio no sangue de grupos expostos em garimpos na Amazônia legal.

```
gar = c(24, 19, 25, 23, 13) # Garimpeiros
rib = c(16, 8, 10, 7, 15) # Ribeirinhos
ind = c(28, 30, 19, 23, 22) # Índios
cont = c(12, 6, 8, 7, 9) # Controle

 # dados
 y_dosi = c(gar,rib,ind,cont) # Variável resposta
 x_dosi = rep(c("Garimpeiros","Ribeirinhos","Índios","Controle"),each=5) # Variável categórica
 dados_dosi = data.frame(y_dosi,x_dosi)
```

Considere o nível de significância igual a 5% e, deste modo verifique a diferença entre as respostas médias dos grupos. Os dados obedecem a condição de normalidade? Os grupos possuem homogeneidade de variâncias entre si? Quais são as hipóteses para o caso descrito? É possível implementar o gráfico box plot?

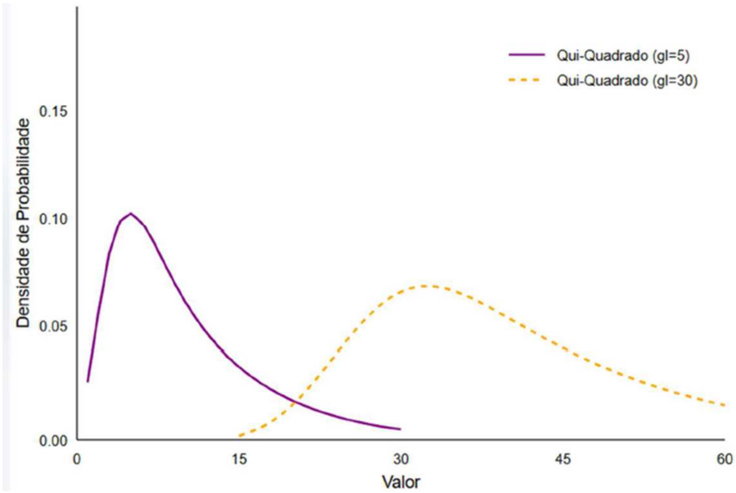
9. Verifique as afirmações abaixo como Verdadeiro (V) ou Falso (F) com base nas características das Distribuições de Probabilidade.

I. **(V)** A Distribuição Normal é a mais importante da estatística, modelando variáveis contínuas, sendo definida pela Média (que centraliza a curva) e pelo Desvio Padrão (que define sua largura);

R: A Distribuição Normal é a mais importante da estatística, descreve dados que se agrupam em torno de uma média, possui formato de curva de sino, modela variáveis contínuas, e seus parâmetros são a Média (centro) e o Desvio Padrão (largura).

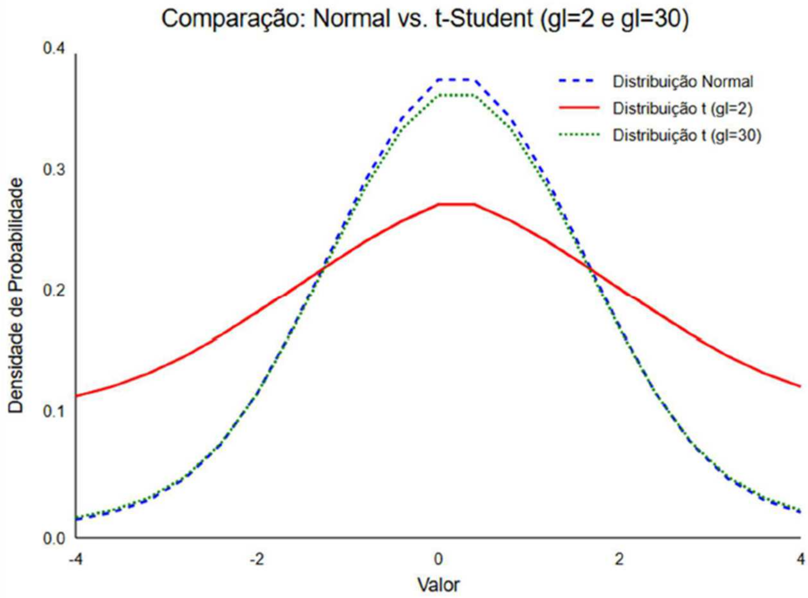
II. **(V)** Para a Distribuição Qui-Quadrado, uma característica fundamental é ser positivamente assimétrica e possuir valores que são sempre positivos;

R: A Distribuição Qui-Quadrado possui características de ser assimétrica positiva (puxada para a direita) e seus valores são sempre positivos.



III. **(V)** A Distribuição T de Student é recomendada para analisar a média de uma população quando o tamanho da amostra é pequeno e o desvio padrão populacional é desconhecido;

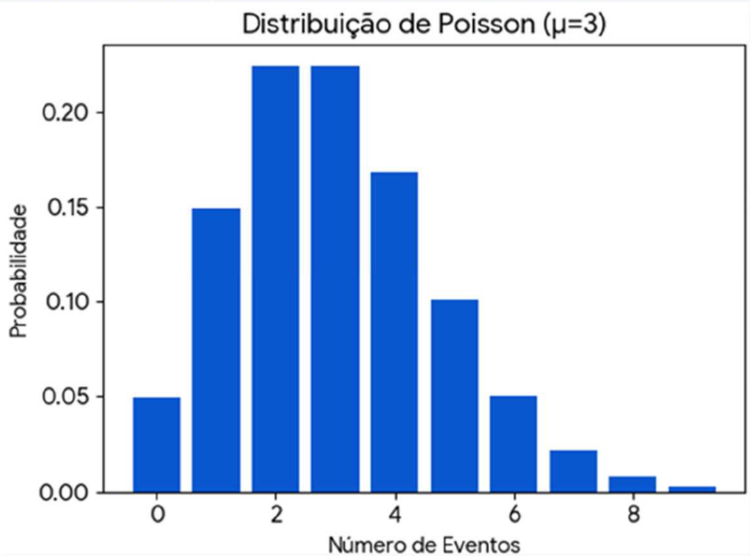
R: A Distribuição T de Student é adequada para analisar a média de uma população quando o tamanho da amostra é pequeno ($n < 30$) e o desvio padrão é desconhecido.
Para $n \geq 30$ mesmo que o desvio padrão seja desconhecido, a distribuição *t* se torna quase idêntica à distribuição normal.
Exemplo:



IV. **(V)** Na Distribuição de Poisson, que modela o número de eventos em um intervalo fixo, o parâmetro crucial é o (Lambda), representando a taxa média de ocorrência, e esta distribuição não utiliza o conceito de graus de liberdade;

R: A Distribuição de Poisson modela o número de eventos em um intervalo fixo (com taxa média constante), sendo o (Lambda) a taxa média de ocorrência. Ela se caracteriza por ser uma variável discreta e, diferentemente de outras distribuições, não trabalha com graus de liberdade.

Exemplos:
- Número de chamadas recebidas em uma central por minuto;
- Número de clientes que chegam a uma loja por hora;
- Número de acidentes em um cruzamento por mês;
- Número de mutações em um trecho de DNA.

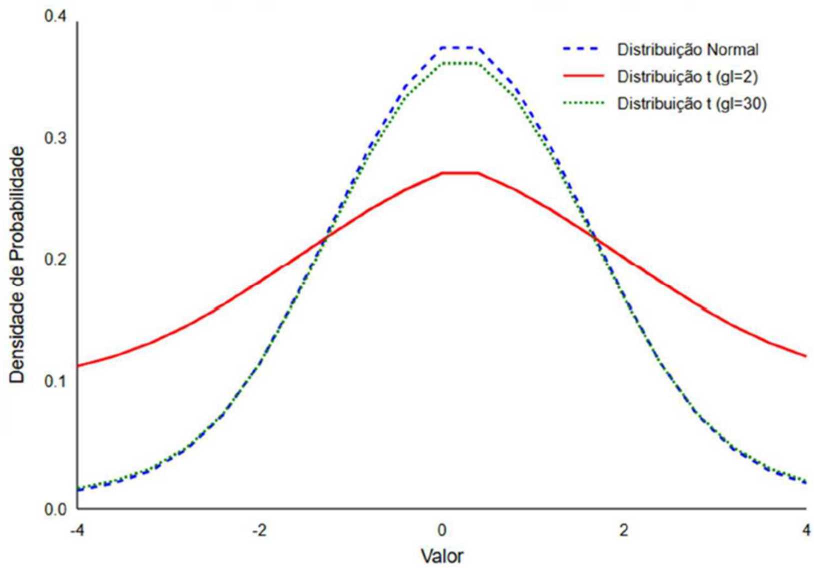


V. **(F)** A Distribuição de Bernoulli é um caso particular da Distribuição Binomial, diferenciando-se por modelar o resultado de múltiplos experimentos independentes, contando o número de sucessos em um número fixo de tentativas;

R: Distribuição de Bernoulli **é** um caso especial da Binomial onde há apenas uma tentativa (n = 1), modelando um único experimento com dois resultados possíveis (sucesso ou falha). O Binomial, por sua vez, é que modela o número de sucessos em um número *fixo* de tentativas.

VI. **(V)** Baixos Graus de Liberdade na Distribuição T de Student resultam em uma curva mais "achatada" e com caudas mais "pesadas" que a curva Normal, indicando uma probabilidade maior de encontrar valores extremos, devido à alta incerteza sobre a população.

R: Quando os graus de liberdade são baixos (amostra pequena), a curva t é mais "achatada" e com caudas mais "pesadas" que a curva Normal para compensar a alta incerteza sobre os parâmetros populacionais. Isso resulta em uma probabilidade maior de encontrar valores extremos.



10. A composição química do petróleo deve apresentar 85% de carbono. Um Engenheiro Químico controla esse processo industrial retirando uma amostra aleatória e realizando um teste de hipóteses, a fim de avaliar se, em média, a composição do petróleo produzido está confor me especificação técnica. Para um determinado teste, o Engenheiro obteve p-valor igual a 0.04. Dessa forma, ao nível de significância de 7%, deve-se concluir que:

a) Não rejeita-se a hipótese nula. Não há evidência de que, em média, a composição química está incorreta.
b) Não rejeita-se a hipótese nula. Em média, a composição química está definitivamente incorreta.
c) Rejeita-se a hipótese nula. Há evidência de que, em média, a composição química está correta.
d) Rejeita-se a hipótese nula. Há evidência de que, em média, a composição química está incorreta.
e) Não rejeita-se a hipótese nula. Há evidência de que, em média, a composição química está incorreta.

Solução:

Primeiramente deve-se observar que as hipóteses do teste são:

H_0: p = 85% (Composição química do petróleo apresenta 85% de carbono);
H_1: p ≠ 85% (Composição química do petróleo não apresenta 85% de carbono).

Sabendo que p-valor é igual a 0,04 e o nível de significância é igual a 0,07.

p-valor < α, logo rejeita-se a hipótese nula.