Teste de Significância da Hipótese Nula, Nível Alfa, Valor *p* e Resultado Significativo

Premissas

**1. Estado do mundo pode ser dividido em duas grandes categorias: um efeito/relação existe *ou* não existe;**

**2. Não sabemos qual é o verdadeiro estado do mundo. É por isso que conduzimos pesquisas científicas;**

**3. Resultados são probabilísticos;**

**O Que é Teste de Significância da Hipótese Nula (NHST)?**

**- é um procedimento que visa tomar uma determinada decisão sobre uma hipótese, dentro do *framework* frequentista;**

**O Que é o “Nulo” da Hipótese Nula ()?**

**- o nulo do NHST quer dizer a hipótese que tentaremos nulificar (falsear);**

**- pode ou não significar hipótese “nil” (e.g., hipótese de diferença zero ou de correlação zero);**

**- hipótese “nil” (hipótese “zero”): Não há diferença média nas alturas de homens e de mulheres. Formalmente, ;**

**- hipótese nula não-“nil”: A diferença média entre alturas de homens e de mulheres é inferior a 10 cm. Formalmente, ;**

**- Cohen (1994): psicólogos entendem que testar a hipótese nula é sinônimo de testar a hipótese “nil” (hipótese “zero”);**

**- no entanto, como vimos anteriormente, nulo no NHST não quer dizer hipótese zero;**

O Que é Alfa?

***-* alfa() é o critério de decisão do pesquisador. […] Indica a probabilidade teórica de que a hipótese nula será rejeitada *se ela for verdadeira*; o alfa não garante que a probabilidade de rejeitar a hipótese nula será mantida constante, a menos que se assuma que essa hipótese nula é sempre verdadeira (Nickerson, 2000);**

**- Neyman–Pearson: o alfa deve ser especificado *antes* de conduzirmos um experimento, de modo a podermos interpretá-lo como a frequência de erros no longo prazo (i.e., erro do Tipo I; Gigerenzer, 1993).**

**- Fisher: em sua primeira fase, sugeriu o nível de significância de 0,05 (Gigerenzer, 1993);**

O Que é Valor *p*?

**- é a probabilidade de observar um conjunto dados igual ou mais extremo que os observados no meu estudo, dado que *H*0 é verdadeira. Formalmente, . O valor *p*:**

**a. NÃO DIZ ALGO sobre a hipótese nula;**

**b. NÃO DIZ ALGO sobre a hipótese alternativa;**

**c. NÃO DIZ ALGO sobre uma teoria ou outra;**

**d. NÃO DIZ ALGO sobre a probabilidade não condicional dos dados;**

**- os valores *p*s nos informam o quão surpreendentes os dados são, assumindo que não existe efeito (Lakens, 2022);**

**- *p* = 0,25 e *p* = 0,001: o segundo valor é mais surpreendente porque é *mais improvável* que uma pessoa que diz ter poderes parapsicológicos acerte cor do cabelo, cor da camiseta, que está com um livro debaixo do braço, considerando a hipótese nula de que a pessoa não possui poderes parapsicológicos (i.e., *menor* valor *p* indica resultado *mais* surpreendente, assumindo que a hipótese nula é verdadeira);**

Duas Analogias

****

O Que Significa um Resultado Estatisticamente Significativo?

**- a interpretação epistêmica de Fisher se refere *a um experimento em particular*: o valor *p* indica o grau de falsidade da (i.e., uma visão quase-Bayesiana);**

**- a interpretação comportamental de Neyman se refere *a um conjunto repetido de experimentos*: um valor *p* menor que o alfa pré-definido leva o pesquisador a tomar uma decisão de ação (i.e., indica que é falsa; caso contrário, haja como se fosse verdadeira);**

**- alfa = 0,05, *p* = 0,10; não vou afirmar que o medicamento e o placebo são iguais para reduzir sintomas de dores de cabeça, mas vou agir como se fosse;**

**- observação: significativo (*significant*) é diferente de *meaningful* (importante, relevante);**

Exemplo Didático

**- simulações contribuem para exercitar o raciocínio probabilístico, além de serem úteis para ilustrar conceitos frequentistas;**

**- Jacob assume que uma moeda é honesta. Ele lança essa moeda para cima 100 vezes e anota seus resultados. Se o resultado for provável de ocorrer em 5% ou menos das repetições hipotéticas desse experimento, Jacob rejeitará a hipótese de que a moeda é honesta. Com base nesse cenário, responda:**

**a. Qual é a hipótese nula ()?**

**A hipótese nula é de que a moeda é honesta. Formalmente, .**

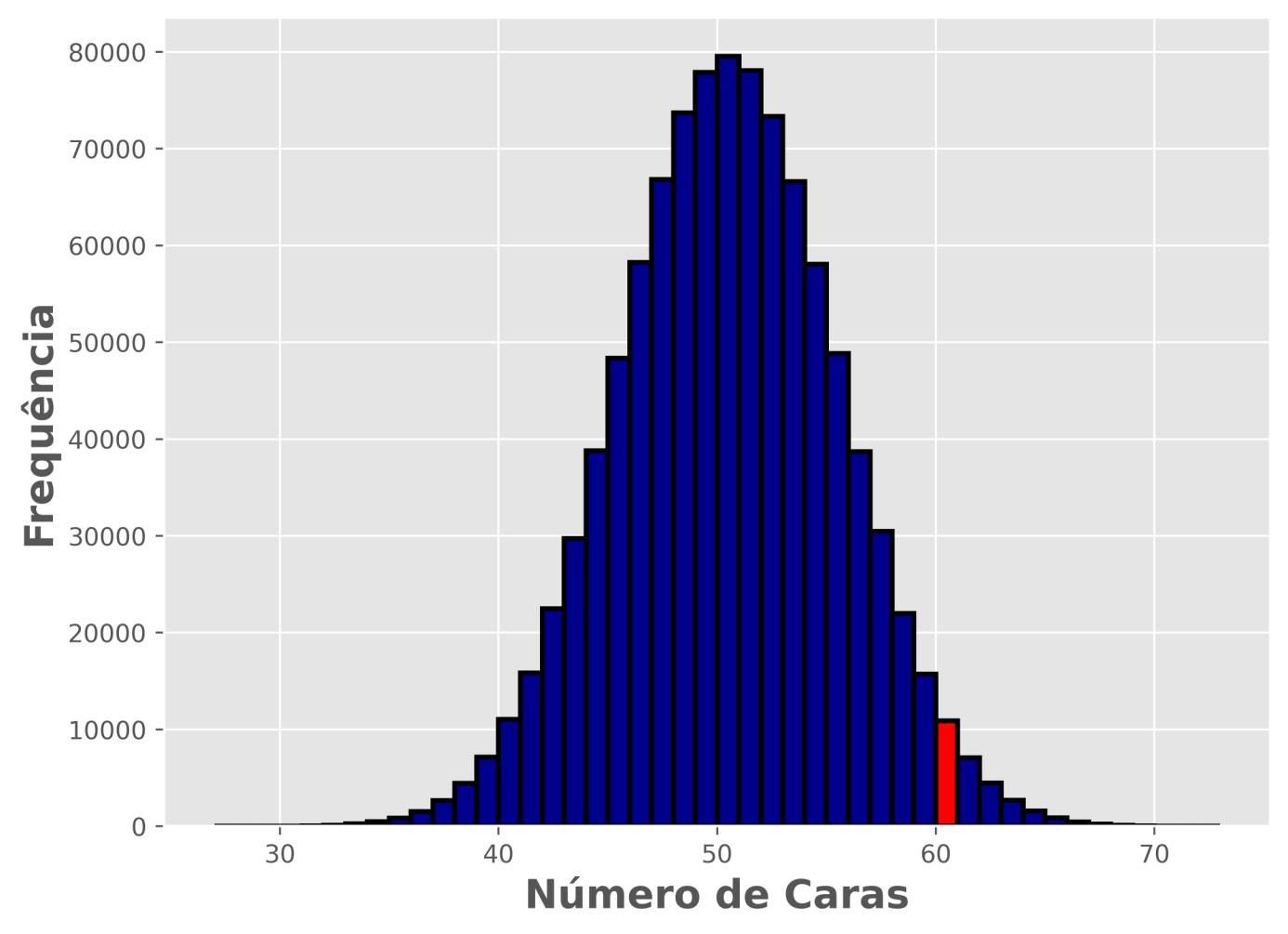
**b. Qual é o alfa?**

**Jacob decide rejeitar a hipótese de que a moeda é honesta se o resultado de seu experimento de 100 lançamentos for provável de ocorrer em 5% ou menos das repetições hipotéticas desse experimento. Logo, o nível alfa adotado por Jacob é de 0,05. Em outras palavras, isso significa que Jacob considerará os resultados surpreendentes o suficiente para “duvidar” da veracidade de se resultados iguais ou mais extremos que esse forem esperados em, no máximo, 5% das repetições do experimento “jogar a moeda para cima 100 vezes e observar seus resultados.”**

**c. Jacob observa 60 caras e 40 coroas. O quão provável é que ocorram resultados iguais ou mais extremos que esse no longo prazo, assumindo-se verdadeira?**

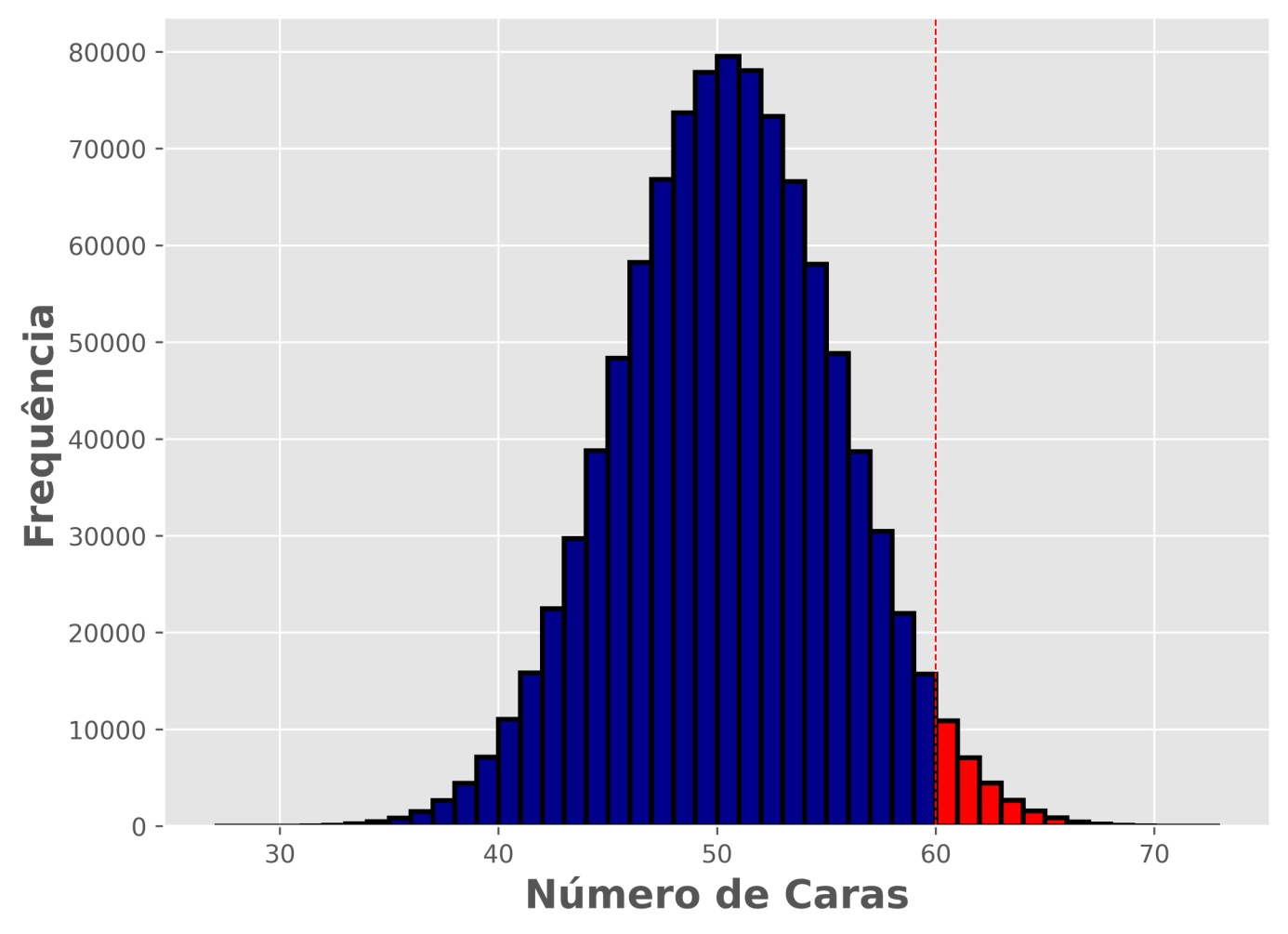
**A probabilidade de se obter 60 caras em 100 lançamentos de moedas, assumindo verdadeira, pode ser obtida pelo cálculo da probabilidade de *x* sucessos em *n* eventos dicotômicos com probabilidade *p* de sucesso:**

**A probabilidade de se obter exatamente 60 caras em 100 lançamentos de moedas, assumindo que a moeda é honesta, é dada por:**

**Esse valor é expresso no gráfico a seguir: **

**No entanto, não queremos saber a probabilidade de obtermos 60 caras, mas sim a probabilidade de obtermos 60 *ou mais* caras!**

**Em outras palavras, . Essa probabilidade é ilustrada no gráfico a seguir:**

****

Resumo da Lógica do NHST

**1. Especificamos de antemão nossa e nosso nível alfa. Assumimos como verdadeira. O alfa é a nossa “barra,” o nosso critério de decisão:**

**a. Alfas maiores indicam que estamos dispostos a nos surpreender com menores desvios do que seria altamente esperado em um mundo em que é verdadeira;**

**b. Alfas menores indicam que estamos dispostos a nos surpreender apenas diante de evidências substanciais de desvios do que seria esperado em um mundo em que é verdadeira;**

**2. Coletamos um conjunto de dados (i.e., jogamos uma moeda 100 vezes, conduzimos um experimento, realizamos uma pesquisa de levantamento);**

**3. Consideramos todos os resultados possíveis da nossa coleta de dados por meio de algum modelo estatístico. Quantificamos quantos desses cenários são iguais ou mais extremos que o cenário empírico que observamos em no passo 2. O valor de probabilidade atribuído nesse processo de quantificação é o nosso valor *p*;**

**4. Se o valor *p* obtido no passo 3 for menor que o alfa especificado no passo 1, temos duas possibilidades:**

**a. O resultado que observamos, apesar de improvável, foi decorrente apenas de variação amostral (i.e., a famosa cagada). Ainda podemos confiar que é verdadeira e;**

**b. O resultado que observamos é tão improvável em um mundo em que é verdadeira que, possivelmente, isso indica que não é uma boa descrição do estado do mundo. Por isso, rejeitamos (como uma boa descrição do mundo). Em outras palavras, a diferença observada provavelmente é real;**

Erros dos Tipos I e II e Poder Estatístico

Premissas

**1. Estado do mundo pode ser dividido em duas grandes categorias: um efeito/relação existe *ou* não existe;**

**2. Não sabemos qual é o verdadeiro estado do mundo. É por isso que conduzimos pesquisas científicas;**

**3. Resultados são probabilísticos;**

Erros dos Tipos I e II

***-* Se você antecipadamente afirma que um efeito/relação não existe () e essa afirmação está correta (i.e., o efeito no mundo real de fato não existe), um erro do Tipo I consiste em rejeitar ; *p* < alfa;**

**- Se você antecipadamente afirma que um efeito/relação não existe () e essa afirmação está incorreta (i.e., o efeito no mundo real na verdade existe), um erro do Tipo II consiste em falhar em rejeitar ;**

**- o erro que você pode cometer depende do estado real do mundo. O problema é que você não sabe qual é esse estado real. Afinal de contas, se você soubesse, por que estaria conduzindo sua pesquisa?;**

**- exemplo:**



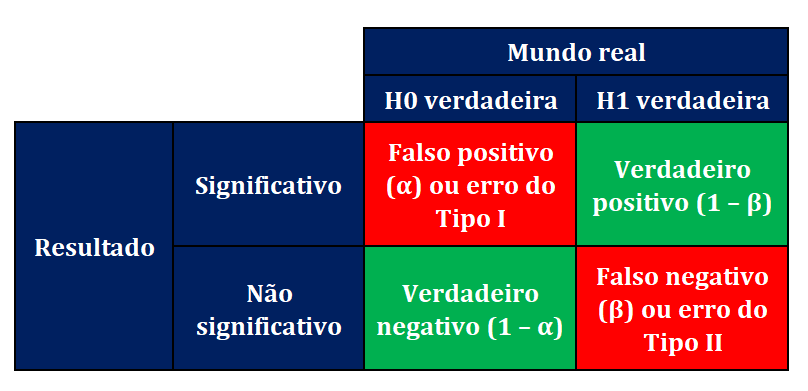
Poder Estatístico

***-* esse conceito de Neyman–Pearson aplica-se a casos em que um efeito/relação existe no mundo real. Desse modo, ele só faz sentido se considerarmos uma hipótese rival à , chamada de hipótese alternativa, . Ela pode ser pensada e definida em função de , referindo-se ao grau pelo qual é falsa. Se é “pouco” falsa, é uma hipótese rival não tão distante de . Se é “muito” falsa, é uma hipótese rival bem distante de ;**

**- consiste na probabilidade de detectar um efeito, caso ele exista;**

**- o conceito de poder estatístico deve ser considerado em relação a um tamanho de efeito pré-especificado. Por exemplo, podemos ter poder estatístico para detectar um(a) efeito/correlação grande, mas não para detectar um(a) efeito/correlação pequeno(a);**

**- analogia: um telescópio pode ser potente o suficiente para detectar a presença de determinado corpo celeste maior, mas não para detectar um corpo celeste menor. Para detectar o corpo celeste menor, pode ser necessário um telescópio mais potente;**

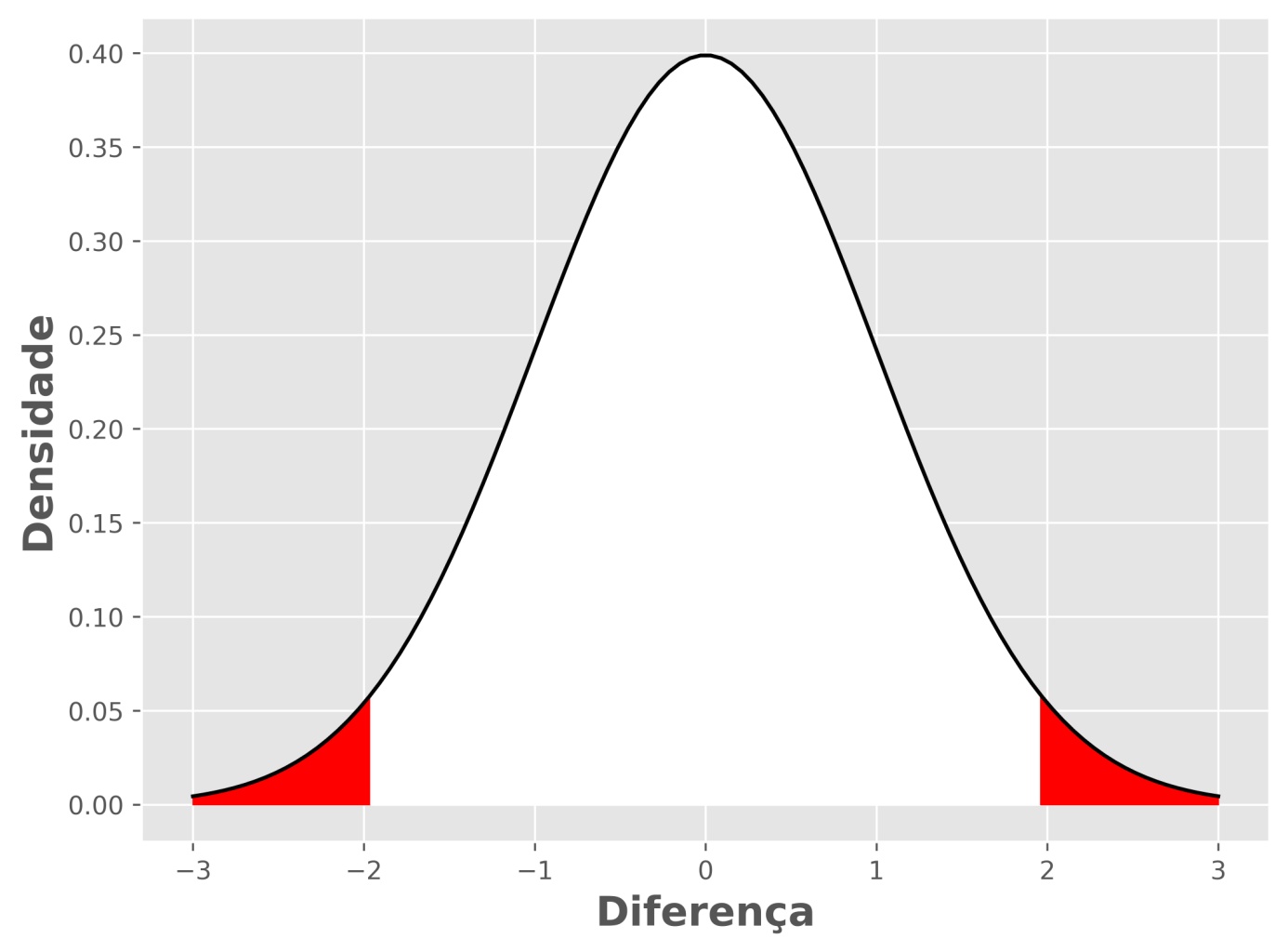


Fatores Que Afetam o Poder Estatístico e Outras Simulações

Introdução ao Exemplo Usado nas Simulações Subsequentes: Diferenças em Escores de Inteligência Entre Homens e Mulheres

**- existem diferenças na inteligência de homens e de mulheres?**

**- simulações contribuem para exercitar o raciocínio probabilístico, além de contribuírem para ilustrar conceitos frequentistas;**

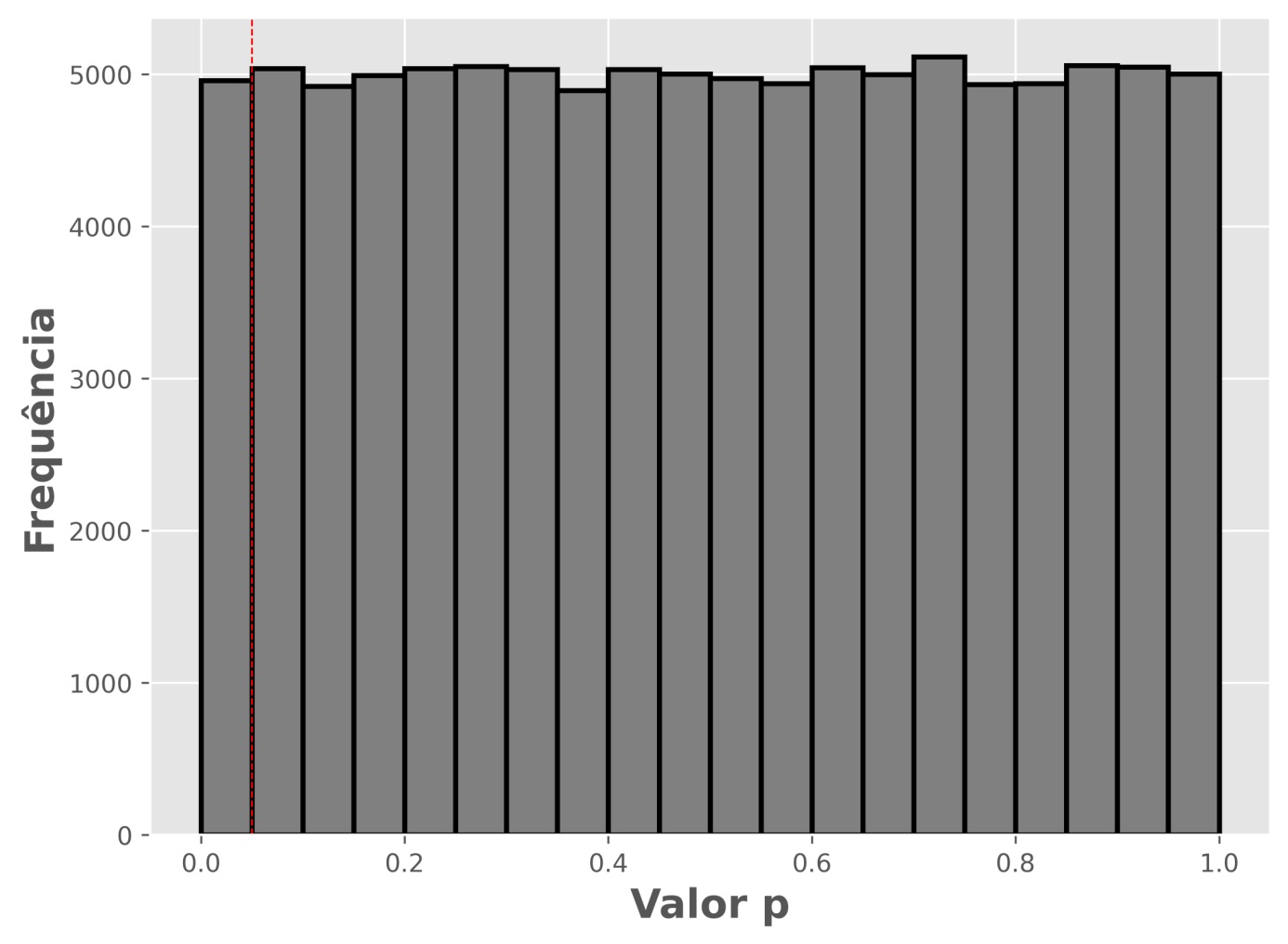


Simulação 1: A Inteligência Média de Homens e Mulheres é Igual (Hipótese Nula Verdadeira)

**Questão 1: Após um elevado número de estudos (100.000) com tamanho amostral fixo (*n* = 20 por grupo), qual é o percentual de valores *p*s que será menor que o alfa?**

**Questão 2: Como é a distribuição de valores *p*s quando a hipótese nula é verdadeira?**

**Questão 1: Considerando verdadeira, a frequência relativa no longo prazo de *ps* < α tenderá a (α × 100)%. Esse valor representa justamente a taxa de erro do Tipo I que o pesquisador pré-especificou. Se o pesquisador deslocar seu critério (nível alfa) para a direita no gráfico a seguir, o percentual de valores *p*s < α aumentará.**



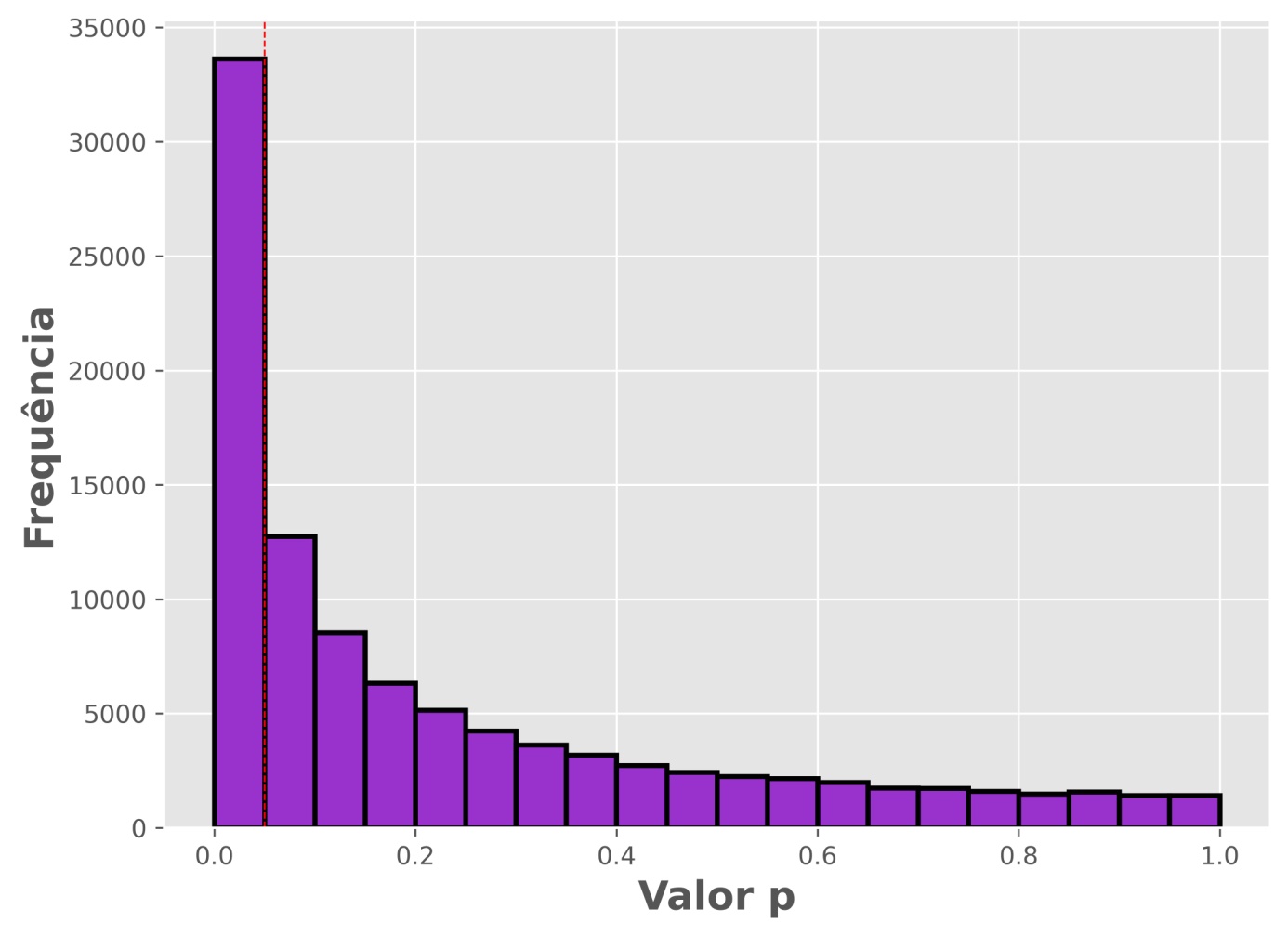
**Questão 2: Considerando verdadeira, a distribuição de valores *p*s de estudos sucessivos será uniforme ou retangular.**

Simulação 2: A Inteligência Média de Mulheres é 0,5 DP (7,5 pontos) Maior que a de Homens (Hipótese Nula Falsa)

**Questão 3: Após um elevado número de estudos (100.000) com tamanho amostral fixo (*n* = 20 por grupo), qual é o percentual de valores *p*s que serão menores que o alfa?**

**Questão 4: O que esse percentual representa em termos de distribuição e de teoria estatística**?

**Questão 3: A resposta a essa questão dependerá do poder estatístico (probabilidade de observar um resultado significativo quando é falsa). O percentual de valores *p*s menores que o alfa tenderá ao poder estatístico. Por exemplo, se o poder estatístico é de 33%, espera-se que aproximadamente 33% dos *p*s sejam menores que o alfa; se o poder é de 80%, espera-se que aproxiadamente 80% dos *p*s sejam menores que o alfa, e assim por diante.**

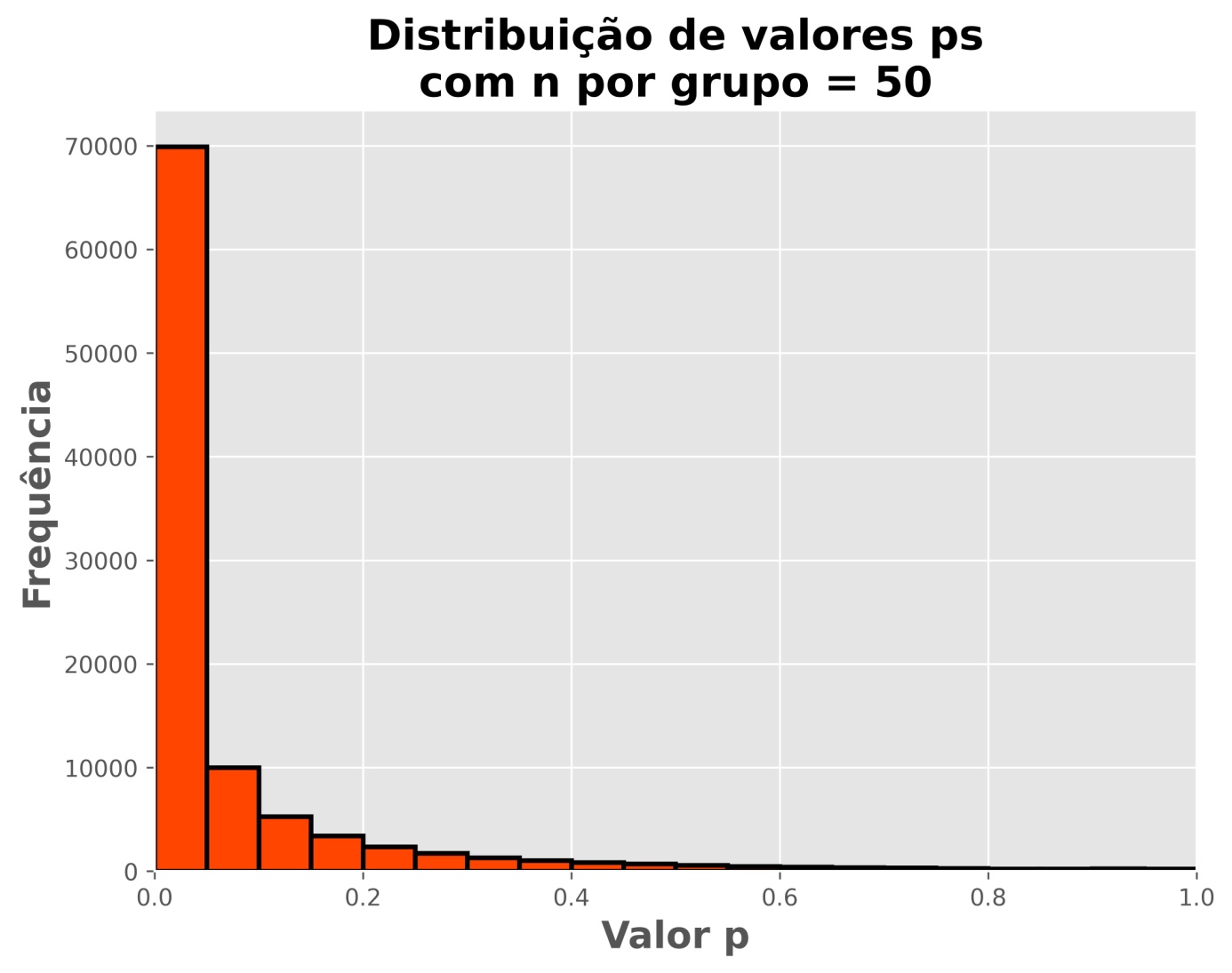


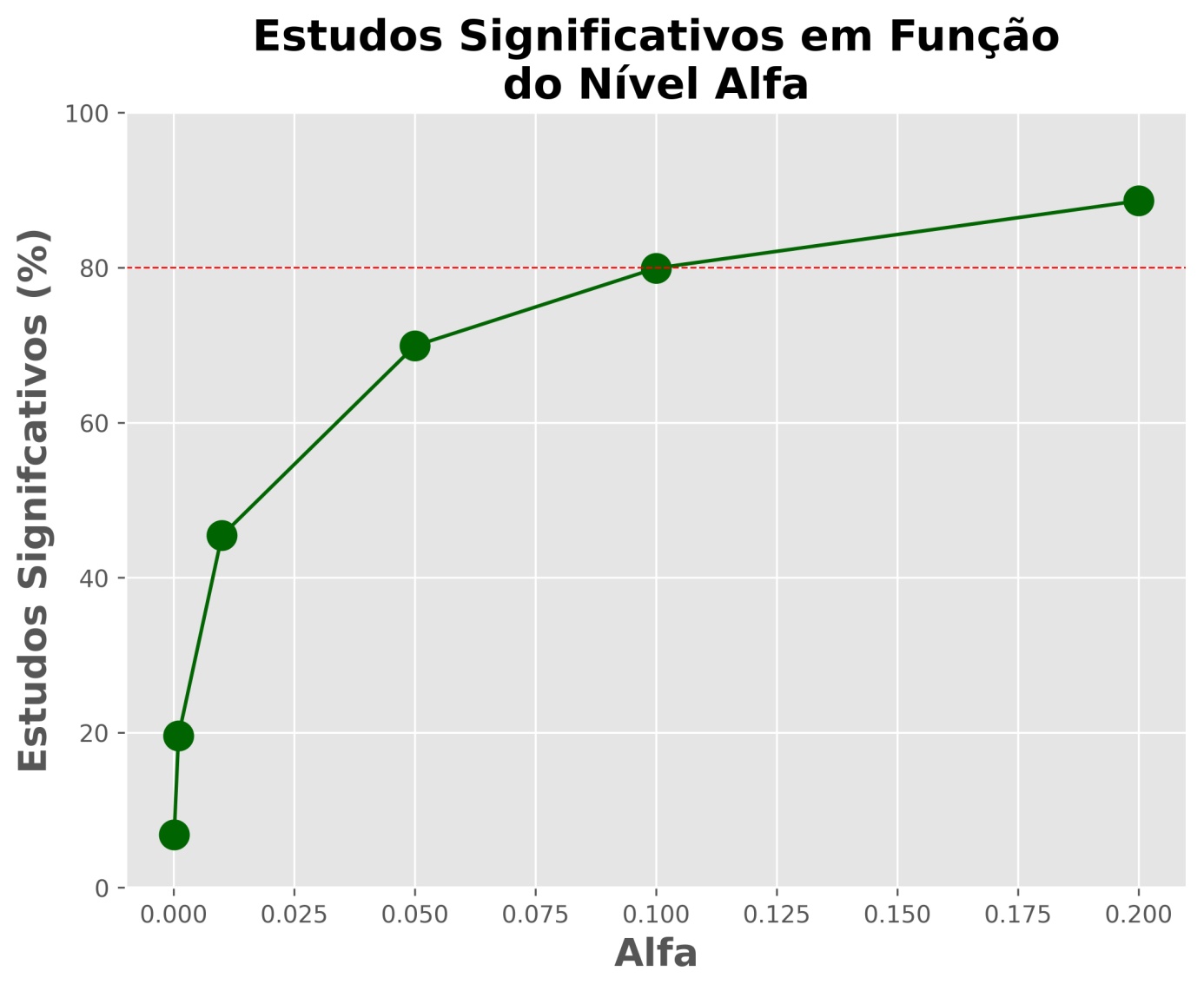
**Questão 4: Esse percentual representa justamente o poder estatístico. Em termos de distribuição, quanto maior for o poder estatístico, mais assimétrica será a distribuição (i.e., mais os valores *p*s se aglutinarão abaixo do alfa, com poucos valores tendendo na direção de *p* = 1). Além disso, desse raciocínio se depreende a ideia de que, mantidos os demais fatores constantes, se dois estudos obtêm resultados não significativos, aquele com tamanho amostral maior é mais informativo, pois sugere que o efeito pode ser pequeno ou irrelevante do ponto de vista teórico ou prático.**

Simulação 3: A Inteligência Média de mulheres é 0,5 DP (7,5 pontos) Maior que a de Homens (Hipótese Nula Falsa)

**No entanto, repetiremos as simulações para vários cenários, com *n* = 50 por grupo. Suponha que 7 diferentes pesquisadores adotem 7 critérios alfas distintos (0,0001, 0,001, 0,01, 0,05, 0,10 e 0,20) nesses estudos.**

**Questão 5: Qual pesquisador afirmará ter encontrado mais estudos com diferenças significativas e qual afirmará ter encontrado mais estudos sem diferenças significativas? Por quê?**





**Questão 5: O pesquisador que adotar o nível alfa mais liberal () é aquele que afirmará ter observado mais resultados estatisticamente significativos, enquanto aquele que adotar o nível alfa mais conservador () é aquele que afirmará ter observado mais estudos sem diferenças estatisticamente significativas. A explicação é simples: ao adotar diferentes alfas, diferentes pesquisadores estão erguendo suas “barras” a diferentes patamares. Um resultado que é significativo para um pesquidador que adota um alfa de 0,0001 também será significativo para um pesquisador que adota um alfa de 0,20, mas o contrário não é verdadeiro. Se o pesquisador “mais incrédulo” acerca da falsidade de (aquele cujo ) considera um achado surpreendente, certamente o pesquisador “mais crédulo” (aquele cujo ) também achará o resultado surpreendente. No entanto, em muitas ocasiões, o resultado surpreenderá o pesquisador mais crédulo, mas não será convincente para o mais incrédulo para que ele rejeite .**

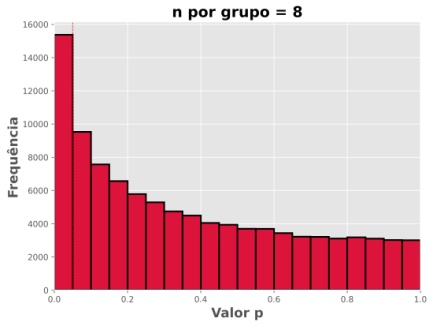
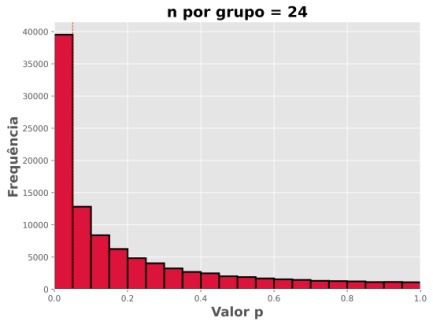
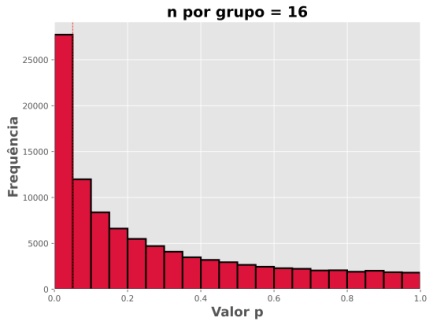
**Adicionalmente, esse resultado nos mostra que adotar um nível alfa maior *aumenta* o poder estatístico, enquanto adotar um nível alfa menor o *diminui*.**

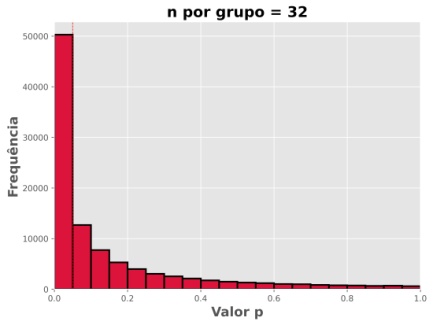
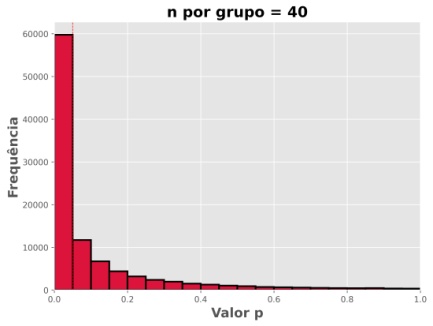
Simulação 4: A Inteligência Média de mulheres é 0,5 DP (7,5 pontos) Maior que a de Homens (Hipótese Nula Falsa)

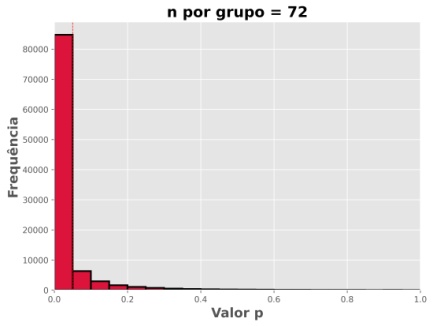
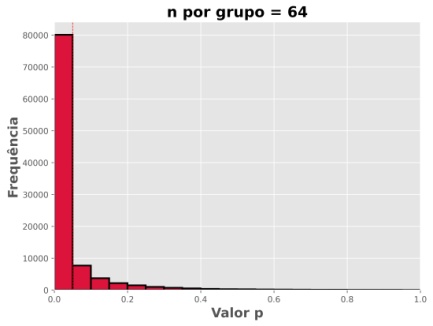
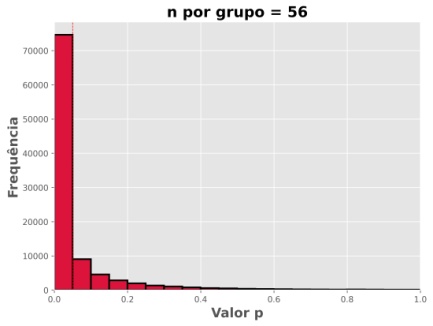
**No entanto, repetiremos as simulações várias vezes, com cenários considerando diferentes tamanhos amostrais, de 8 a 96, em incrementos de 8 participantes por grupo.**

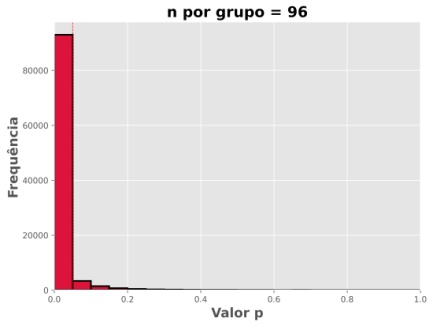
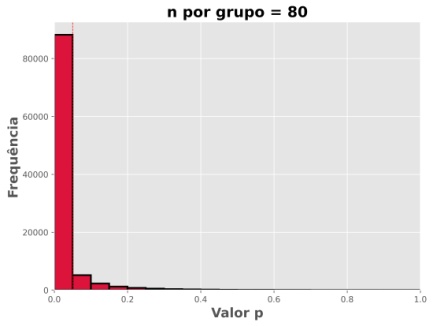
**Questão 6: O percentual de valores *p*s que serão menores que o alfa variará em função dos tamanhos dos grupos? Por quê?**

**Questão 7: Mantido o tamanho de efeito constante, qual é o significado da relação entre percentual de estudos significativos e tamanho amostral?**

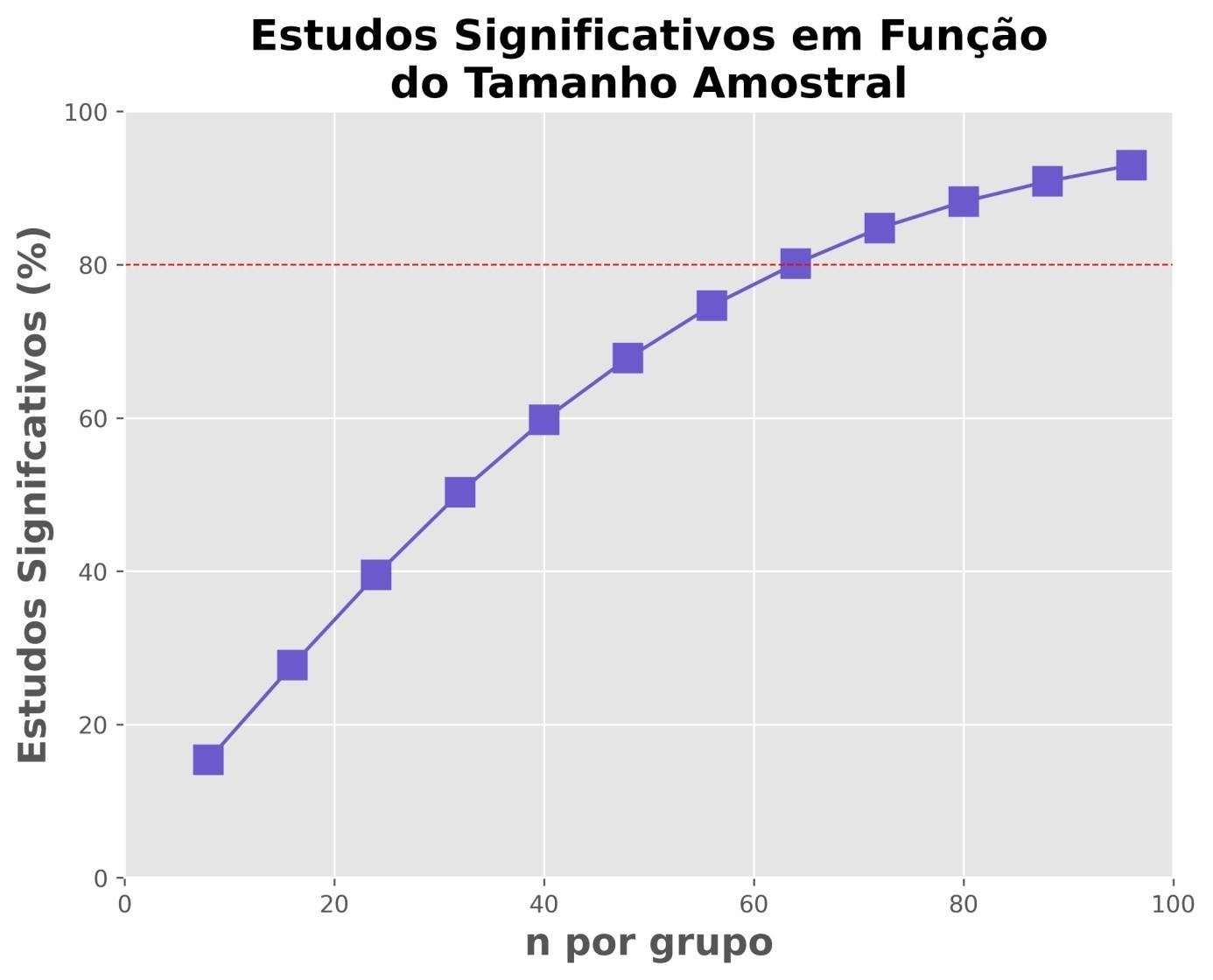
 





**Dica: olhe para os gráficos acima antes de verificar a resposta na página a seguir.**

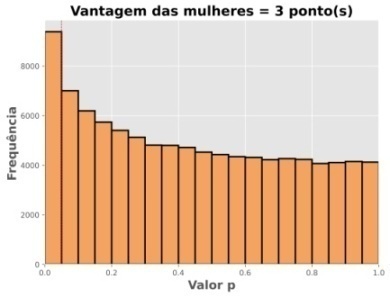
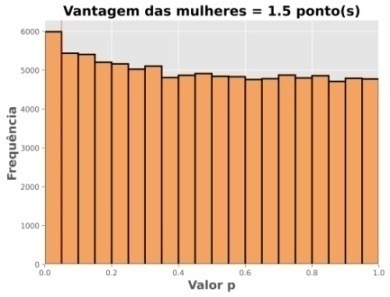
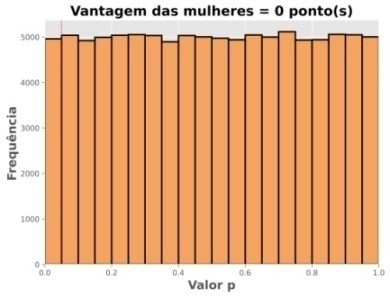
**Questão 6: Assumindo um alfa fixo de 0,05 em todas as simulações, a resposta à questão 6 é positiva, isto é, o percentual de valores *p*s aumentará com o aumento do tamanho dos grupos. O motivo disso é que amostras maiores produzem estimativas mais precisas e, consequentemente, possuem maior poder estatístico para detectar diferenças entre grupos.**

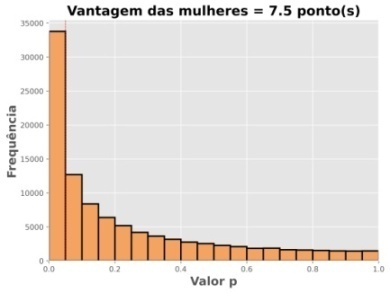
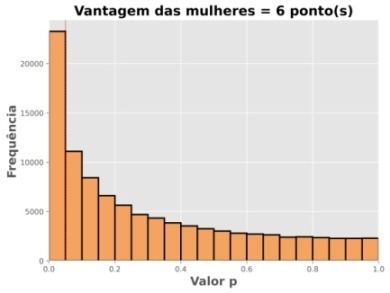
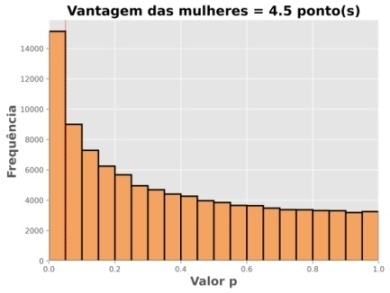
**Questão 7: A interpretação desse aumento é que, mantidos outros fatores constantes, o poder estatístico aumenta com o aumento do tamanho amostral. Amostras com tamanhos grupais maiores ou iguais a 64 atingem um poder estatístico superior a 80%, considerando-se um tamanho de efeito de *d* = 0,50 (referente aos 7,5 pontos a mais assumidos que mulheres têm nos escores de inteligência quando comparados a homens). Estudos com tamanhos grupais iguais a 8, por outro lado, têm um poder estatístico menor que 20%. Isto é, mesmo que mulheres sejam 0,5 *DP* mais inteligentes que homens, estudos conduzidos com apenas 8 homens e 8 mulheres, adotando um alfa de 0,05, vão *falhar* em detectar esse efeito em 4 de cada 5 estudos! Em outras palavras, tais estudos não são informativos de se um resultado não significativo se deve a uma ausência real de diferenças em inteligência entre homens e mulheres, ou se a um baixo poder estatístico para detectar tais efeitos existentes.**

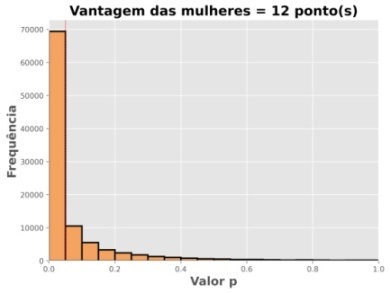
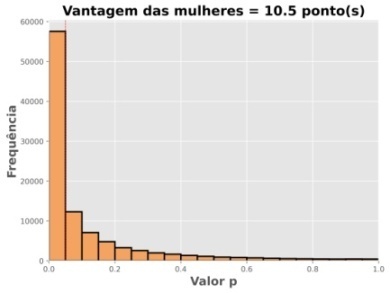
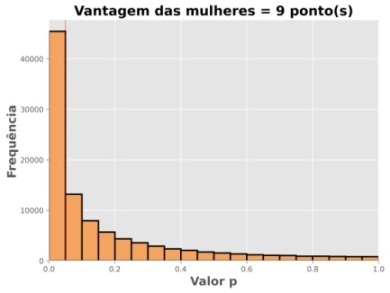
Simulação 5: Mulheres Terão Diferentes Graus de Vantagem em Escores de Inteligência em Relação a Homens, Mas Tamanho Amostral (20 por Grupo) e Alfa Serão Fixos (0,05)

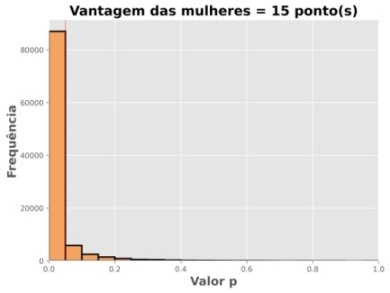
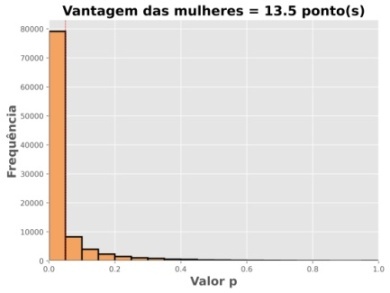
**Esses cenários incluem homens sempre com inteligência média igual a 100. Mulheres terão inteligências médias variando em 100 e 115, com incrementos de 1,5, correspondendo, respectivamente, a *d*s de Cohen de 0 a 1, em incrementos de 0,1. Questão 8: O percentual de valores *p*s que serão menores que o alfa variará em função dos escores médio de inteligência das mulheres? Por quê?**

**Questão 9: Mantido o tamanho amostral e o nível de significância constante, qual é o significado da relação entre percentual de estudos significativos e tamanho de efeito?**

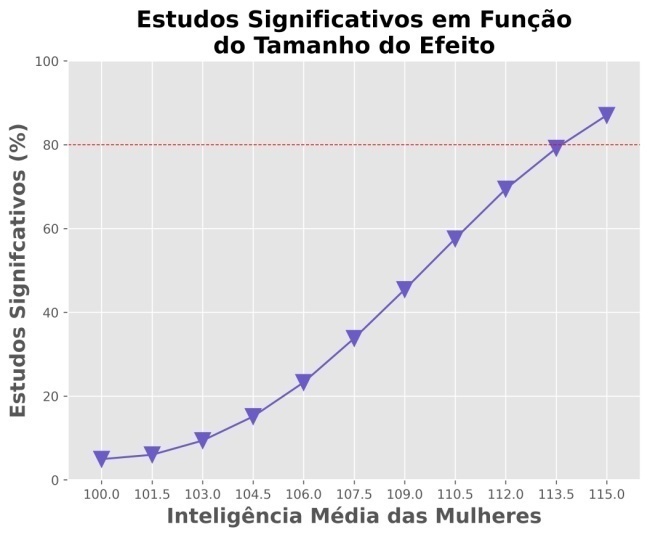
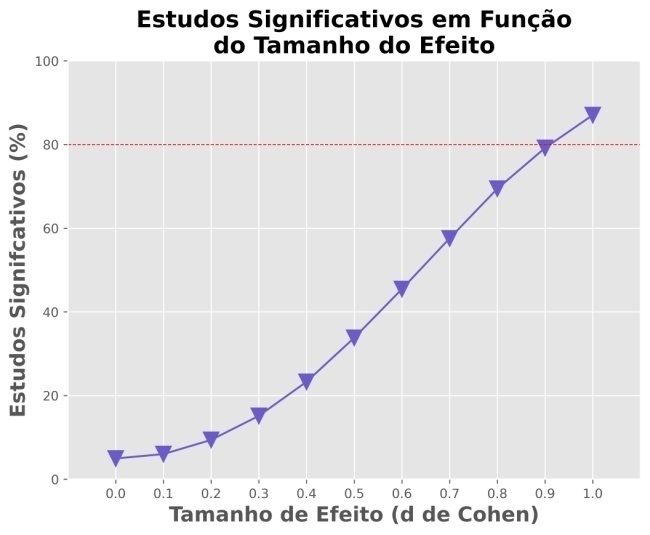








**Dica: olhe para os gráficos acima antes de verificar a resposta a seguir.**

**** ****

**Questão 8: Sim. Um teste estatístico tem maior sensibilidade para detectar um efeito maior do que para detectar um efeito menor, assumindo-se alfa e tamanho amostral constantes.**

**Questão 9: Mantidos outros fatores constantes, na medida em que o tamanho de efeito aumenta, torna-se mais fácil detectá-lo (i.e., o poder estatístico para detectá-lo é maior com um *N* fixo do que para detectar outro tamanho de efeito menor).**

Referências

**Cohen, J. (1994). The earth is round (*p* < .05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997**

**Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Lawrence Erlbaum Associates, Inc.**

**Lakens, D. (2022). *Improving your statistical inferences*. https://doi.org/10.5281/zenodo.6409077**

**Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301. https://doi.org/10.1037//1082-989X.5.2.241**

Sugestões de Materiais

