Análise de Dados Categóricos
Introdução à Disciplina

Enrico A. Colosimo/UFMG

Depto. Estatística - ICEx - UFMG

Análise de Dados Categóricos

- Resposta vs Covariáveis
- Resposta: categórica
- Nesta disciplina vamos estender para Resposta Discreta.

Resposta

1. Categóricas
   - Nominais: gênero, raça, religião, status (doente/saudável), etc
   - Ordinais: IMC (eutrófico, sobrepeso, obeso); Infecção (sem, mono ou poli), etc.
2. Discreta (contagem)
   - número de cáries por paciente;
   - número de chamadas telefônicas por hora;
   - número de ovos por volume de fezes, etc;

Estimação - Prevalência

Um Estudo sobre Hipertensão Arterial Sistêmica na Cidade de Campo Grande, MS

A Study on Systemic Arterial Hypertension in Campo Grande, MS, Brazil

Ana Rita Araújo de Souza, Anselmo Costa, Diogo Nakamura, Leandro Nascimento Mocheti, Paulo Roberto Stevanato Filho, Luiz Alberto Ovando

Universidade para o Desenvolvimento do Estado e da Região do Pantanal, Hospital Beneficente Santa Casa de Campo Grande, Sociedade Brasileira de Cardiologia – MS - Campo Grande, MS

Resumo

Objetivos: Detectar a real prevalência de hipertensão arterial sistêmica em Campo Grande, MS, e fatores freqüentes.

Métodos: Estudo transversal com amostra randomizada da população adulta da cidade de Campo Grande, MS, num total de 892 pessoas. Foi aplicado questionário sobre idade, sexo, escolaridade, tabagismo, etilismo, aspectos sobre o tratamento. Foram colhidos dados antropométricos (peso e altura). Segundo a OMS, foi considerado peso normal: IMC<25 kg/m²; sobrepeso: 25≥IMC<30; obeso: IMC≥ 30. Os critérios para hipertensão foram baseados no VII Joint, com valores de corte de Pressão Arterial de 140 x 90 mmHg.

Resultados: A prevalência de hipertensão foi de 41,4%, variando conforme idade (até 29 anos: 11,8%; 30-39: 24,8%; 40-49: 43,3%; 50-59: 42,4%; 60-69: 48,6% e ≥ 70: 62,3%). Houve maior prevalência nos homens (51,8%), enquanto nas mulheres foi de 33,1%. As pessoas com formação escolar de 1º grau primário tendem a apresentar maiores índices pressóricos. Nos indivíduos com sobrepeso e obesidade, observou-se maior prevalência de pressão elevada: IMC normal (27,9%), sobrepeso (45,6%) e obesidade (58,6%). A partir dos 60 anos existe um maior porcentual de hipertensão sistólica isolada, representado por 16,4% (60-69 anos) e de 24,6% (>70 anos). Etilismo diário ou semanal também está relacionado a maior incidência, respectivamente, de 63,2% e 47,2%. Apenas 59,7% eram sabidamente hipertensos. Das pessoas que apresentaram hipertensão, 57,3% fazem algum tratamento. Dos que fazem tratamento regularmente, 60,5% apresentaram hipertensão.

Conclusão: A prevalência de hipertensão foi de 41,4%, ultrapassando a média detectada em alguns trabalhos, alertando para piora epidemiológica e repercussões cardiovasculares, o que evidencia necessidade de maior investimento público no que tange ao esclarecimento e instrução desses grupos populacionais quanto à prevenção.

Palavras-chave: Hipertensão/prevalência, hipertensão/epidemiologia, fatores de risco.

Arq Bras Cardiol 2007; 88(4)

Comparação - Incidência

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 DECEMBER 28, 2006 VOL. 355 NO. 26

An Intervention to Decrease Catheter-Related Bloodstream Infections in the ICU

Peter Pronovost, M.D., Ph.D., Dale Needham, M.D., Ph.D., Sean Berenholtz, M.D., David Sinopoli, M.P.H., M.B.A., Haitao Chu, M.D., Ph.D., Sara Cosgrove, M.D., Bryan Sexton, Ph.D., Robert Hyzy, M.D., Robert Welsh, M.D., Gary Roth, M.D., Joseph Bander, M.D., John Kepros, M.D., and Christine Goeschel, R.N., M.P.A.

ABSTRACT

BACKGROUND

Catheter-related bloodstream infections occurring in the intensive care unit (ICU) are common, costly, and potentially lethal.

METHODS

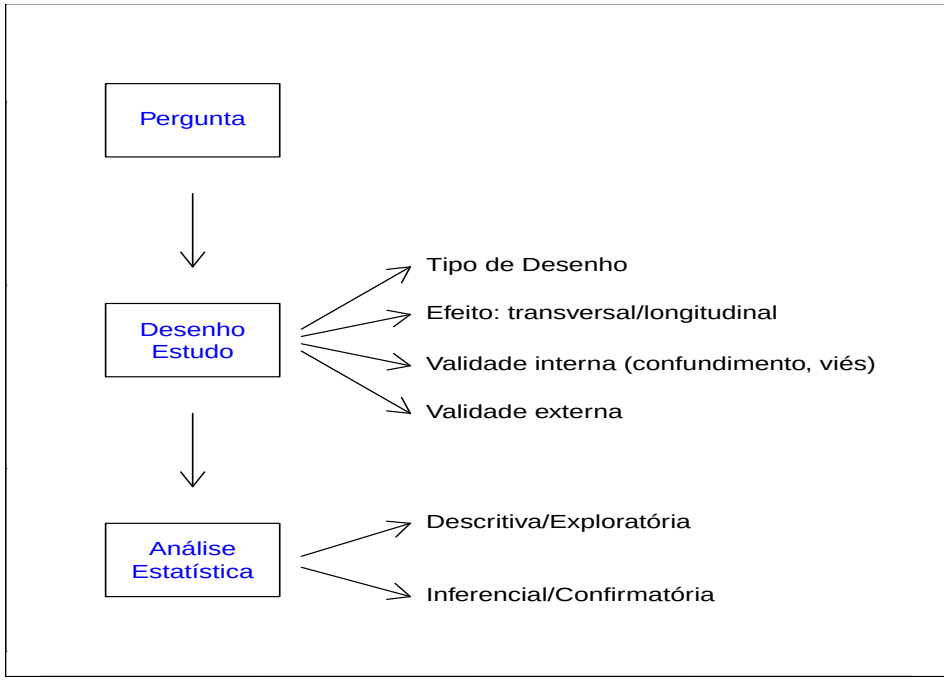
We conducted a collaborative cohort study predominantly in ICUs in Michigan. An evidence-based intervention was used to reduce the incidence of catheter-related bloodstream infections. Multilevel Poisson regression modeling was used to compare infection rates before, during, and up to 18 months after implementation of the study intervention. Rates of infection per 1000 catheter-days were measured at 3-month intervals, according to the guidelines of the National Nosocomial Infections Surveillance System.

RESULTS

From the School of Medicine (P.P., D.N., S.B., S.C., B.S.), the School of Professional Studies in Business and Education (D.S.), and the Bloomberg School of Public Health (H.C.), Johns Hopkins University, Baltimore; and the University of Michigan, Ann Arbor (R.H.); William Beaumont Hospital, Royal Oak (R.W.); Ingham Regional Medical Center, Lansing (G.R.); Harper University Hospital, Detroit (J.B.); Sparrow Health System, Lansing (J.K.); and the Michigan Health and Hospital Association Keystone Center for Patient Safety and Quality, Lansing (C.G.) — all in Michigan.

N Engl J Med 2006;355:2725-32.

Pesquisa Científica



Pesquisa Científica

1. Pergunta de Interesse;
2. Desenho do Estudo/Coleta dos Dados/Observar;
3. Análise Estatística: Modelar/Predizer;
    - Conhecer o Banco de Dados;
    - Análise Descritiva (cada variável separadamente);
    - Análise Bivariada (resposta vs cada covariável);
    - Modelo de Regressão (paramétrico ou não-paramétrico);
    - Inferência: Clássica ou Bayesiana;
    - Resposta da Pergunta/Interpretação dos Resultados.

Pergunta de Interesse

- Comparação de Grupos.
- Identificação de Fatores de Risco ou Prognóstico.
- Estimação/Predição.

Desenho do Estudo

1. Tipos de Desenho de Estudo.
2. Efeito Transversal vs Longitudinal.
3. Tipos de Viés.
4. Validade do estudo.

Perguntas Relevantes

- Os grupos são comparáveis?
- As variáveis de confusão foram medidas/controladas?
- É possível alocar tratamento às unidades amostrais de forma aleatória?
- Os erros de medição podem ser medidos e controlados?
- As perdas (dados perdidos) podem viciar os resultados?
- Podemos estender os resultados para outros estudos?

Tipos de Estudos

1. Estudos Transversais
2. Estudos Longitudinais
   - Observacionais;
     - Coorte (prospectivo ou histórico);
     - Caso-controle (retrospectivo);
   - Experimentais: Ensaio Clínico/"Cross-over".

Estudo Transversal ou de Prevalência

- Características Básicas
  - Amostra tomada em um tempo pré-determinado;
  - Causalidade reversa (impossível determinar causa e efeito).
  - Não é apropriado para estudar doenças raras e nem de curta duração.

Estudo de Coorte

- Características Básicas
  - Estudos observacionais;
  - Grupos de comparação (braços da coorte): usualmente definidos pela presença ou não de uma exposição de interesse;
  - Podem ser prospectivos (forma mais comum) ou retrospectivo/histórico.

Estudo Caso-Controle

- Características Básicas
  - Estudos observacionais e retrospectivos;
  - Grupos de comparação: definidos pela presença ou não de uma doença de interesse.

Estudo Clínico Aleatorizado

- Características Básicas
  - Presença de grupos de comparação.
  - Estudos experimentais. Isto é, a intervenção do investigador consiste em aleatorizar indivíduo ao grupo;
  - Vantagem: controla por fatores de confusão medidos e não medidos.

Viés

1. Desvio da verdade por defeito no delineamento ou na condução de um estudo.

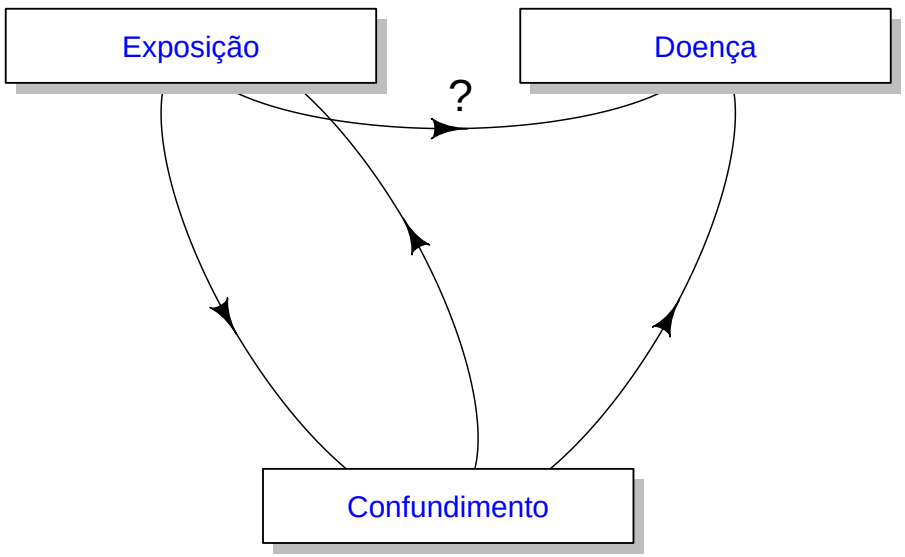
2. Erro sistemático no delineamento, condução e análise de um estudo resultando em erro na estimativa da magnitude da associação entre variável explicativa e a resposta de interesse.

Fontes de Viés

1. Fatores de confusão.
2. Viés de Seleção: alocação das unidades de análise privilegia subgrupos com probabilidade diferenciada de apresentar a resposta. Exemplo: Perda de acompanhamento em estudos longitudinais.
3. Viés de Informação: erro sistemático na classificação das variáveis sob estudo.
4. Outros: viés de publicação, etc.

Fator de Confusão

Definição: Um terceiro fator que está associado tanto com a exposição/covariável quanto com a resposta/doença, mas não se encontra no elo causal entre eles.



Fator de Confusão

- Duas condições para caracterizar um fator de confusão:
  - Ser associado com a covariável/exposição sem ser sua consequência.
  - Estar associado com o resposta/desfecho independente da exposição.

Confundimento: Exemplos

- Idade na associção entre fumo e câncer de estômago.
- Fumo na associação entre consumo de café e câncer de pulmão.
- (contra-exemplo: no elo causal?) Colesterol na associação entre dieta e infarto.

Validação do Estudo

- Validade Interna: sujeito a confundimento e viés;
- Validade Externa: representatividade da amostra.

Sujeito a critérios de inclusão e exclusão do Estudo.

Validade do Estudo/Amostra

1. Critério de inclusão e exclusão restritivo $==>$ população pequena
   - Validade Interna: aumenta;
   - Validade Externa: diminue.
2. Critério de inclusão e exclusão flexível $==>$ população grande
   - Validade Interna: dimunue.
   - Validade Externa: aumenta.

Resposta e Covariáveis

1. Resposta/Desfecho
   - Contínua;
   - Discreta/Contagem;
   - Categórica.
2. Covariáveis
   - Interesse (se existir): Exposição;
   - Contínua ou categórica.

Como identificar o modelo adequado?

1. Modelos para Resposta Contínua.
   - Regressão linear múltipla.
   - Modelos de Análise de Sobrevivência: Paramétrico ou de Cox.
   - Modelos beta, gama, etc.
2. **Modelos para Resposta Categórica ou Contagem.**
   - Modelo Log-Linear (Tabela de Contingência).
   - Modelo de Regressão de Poisson: contagem.
   - Modelo de Regressão Logística (binária ou politômica).

Programa da Disciplina

1. **Conceitos Básicos**: escalas de medida, distribuição binomial, inferência no modelo binomial.
2. **Tabelas de Contingência**: tabelas 2×2, esquemas amostrais, tipos de estudos, testes qui-quadrado e da razão de verossimilhanças, medidas de associação, amostras pareadas, tabelas $r \times c$.
3. **Modelo Log-linear**: análise estratificada, teste de Mantel-Haenszel, tabelas multidimensionais, modelos de independência mútua, marginal e condicional, inferência para modelos log-lineares.
4. **Modelo de Regressão de Poisson**: forma do modelo, inferência e aplicações, superdispersão, modelo binomial negativo.
5. **Modelo de Regressão Logística**: forma do modelo, estimador de máxima verossimilhança, verificando a adequação do modelo, modelo politômico, aplicações.
6. **Medidas repetidas**: modelos marginal e de efeitos aleatórios.

Exemplo: Ecologia

- Descrição: ecologistas desejam estudar o equilíbrio entre machos e fêmeas de uma certa espécie, em risco de extinção, em uma certa localidade.
- Desenho Amostral: uma amostra aleatória de tamanho 20 foi retirada desta localidade, obtendo 8 machos e 12 fêmeas.
- O que podemos concluir a partir desta amostra?

Formulação Estatística

Amostra: $Y_1, \ldots, Y_{20}$; $\quad Y : 0/1$ (macho/fêmea)

Considere a estatística:

$$X = \sum_{i=1}^{20} y_i : \text{número de fêmeas na amostra de tamanho 20}$$

$$X \sim bin(n = 20, \pi)$$

$\pi = P(Y = 1)$: probabilidade de ocorrer uma fêmea.

Pergunta de Interesse (teste de hipóteses):

$$H_0 : \pi = 0,5 \quad \text{vs} \quad H_1 : \pi \neq 0,5$$

Propostas de Solução

1. Enfoque Clássico
   - Teste Exato
   - **Testes Aproximados**: qui-quadrado, Wald, razão de verossimilhanças, etc
2. Enfoque Computacional: Monte Carlo e bootstrap
3. Enfoque Bayesiano

Teste de Hipóteses

1. Enfoque Clássico: encontrar uma estatística teste cuja distribuição, sob H_0, é conhecida.
   - Teste Exato: distribuição exata (difícil de ser encontrada).
   - Testes Aproximados/assintótico: distribuição aproximada/limite.
2. Enfoque Computacional (Monte Carlo e bootstrap): desenhar (histograma) a distribuição. Usualmente, utilizamos a mesma estatística teste do enfoque clássico.
3. Enfoque Bayesiano: baseado na distribuição a posteriori da quantidade desconhecida.

1- Teste Exato

$$X = \sum_{i=1}^{20} y_i$$

é a estatística teste.

- $X \sim bin(n, \pi), \quad n = 20.$
- sob $H_0 : \pi = \pi^0$ temos que:

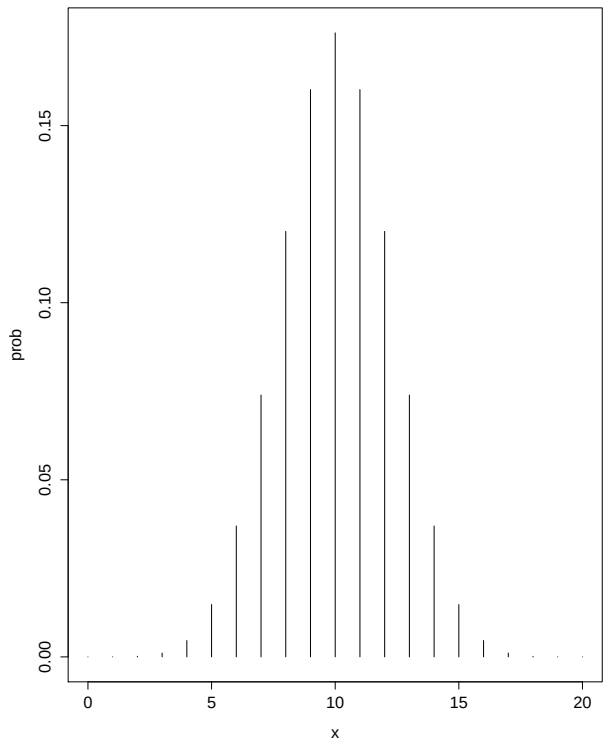
$$X \sim bin(n, \pi^o)$$

Exemplo: $\pi^0 = 0,5$

valor-p $= 2\ P(X$ ser igual ou mais desfavorável que$12 | \pi = 0,5]$

.

Exemplo - sob H_0



valor-p $= 2P[X \geq 12|\pi = 1/2] = P[X \geq 12|\pi = 1/2] + P[X \leq 8|\pi = 1/2] = 1 - [P[(X = 9|\pi = 1/2) + P(X = 10) + P(X = 11)] = 0,5034.$

Intervalo de confiança

Baseado na inversão da Região de Rejeição (Bickel e Doksum, p.180).

1. Limite inferior (π_I) é a solução de:

$$\sum_{r=12}^{20} \binom{20}{r} \pi_I^r (1-\pi_I)^{20-r} = 0,025 \Rightarrow \pi_I = 0,361$$

2. Limite superior (π_S) é a solução de

$$\sum_{r=0}^{12} \binom{20}{r} \pi_S^r (1-\pi_S)^{20-r} = 0,025 \Rightarrow \pi_S = 0,809$$

Um intervalo de 95% de confiança para π é $(0,36; 0,81)$.

2.Teste Qui-Quadrado (mais utilizado)

É o teste assintótico mais utilizado na prática, com a seguinte expressão:

$$X^2 = \sum_{i=1}^{2} \frac{(n_i - E_i)^2}{E_i}$$

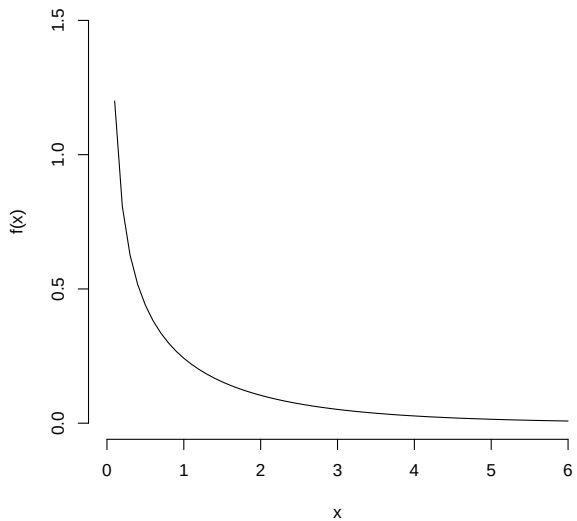
No nosso caso:

$$X^2 = \frac{(x - n\pi^0)^2}{n\pi^0} + \frac{((n-x) - n(1-\pi^0))^2}{n(1-\pi^0)} = \frac{(12-10)^2}{10} + \frac{(8-10)^2}{10} = 0,8$$

Rejeitar H_0 para valores grandes de χ^2

2- Teste Qui-Quadrado

Sob H_0, X^2 tem uma dist. limite qui-quadrado com 1 g.l.



$$\text{valor-p} = P[X^2 > 0,8] = 0,371$$

Intervalo de Confiança

Baseado na inversão da Região de Rejeição

$$RR(x/X^2(x) > \chi^2_{1-\alpha}, H_0)$$

Invertendo RR, temos o Intervalo de $(1-\alpha)100\%$ de confiança para π:

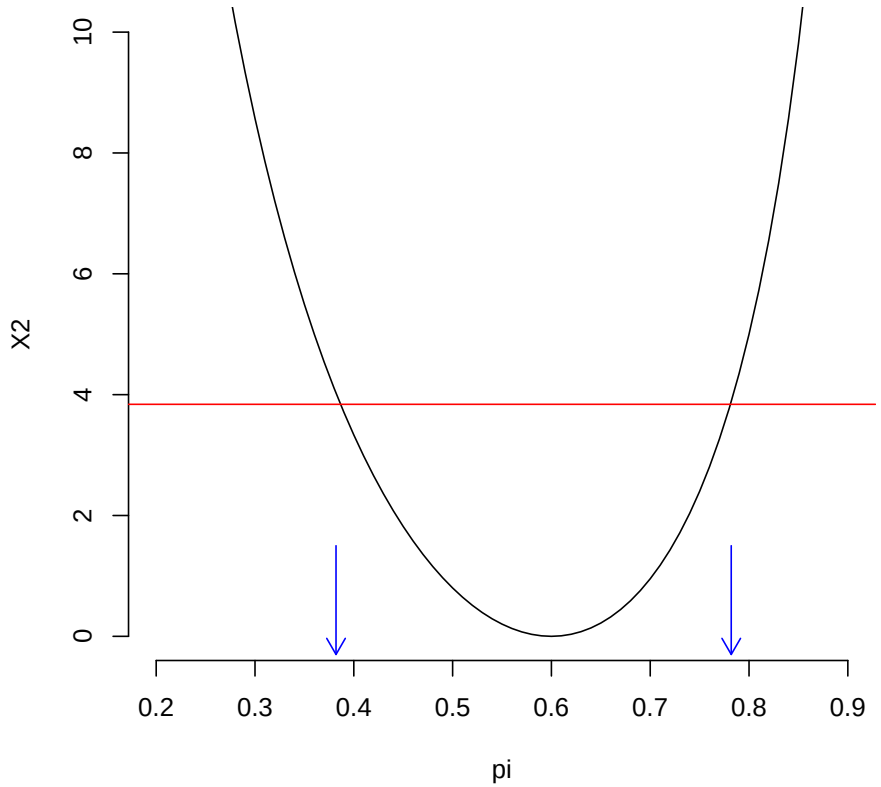
$$\{\pi^0 \in (0,1)/X^2 < \chi^2_{1-\alpha}, X = x\}$$

e para o valor observado $X = 12$ e $\alpha = 0,05$,

$$X^2 = \frac{(12 - 20\pi^0)^2}{20\pi^0} + \frac{((20-12) - 20(1-\pi^0))^2}{20(1-\pi^0)} < 3,84$$

Intervalo de Confiança

$$IC(\pi^0/X = 12, \alpha = 0,05) \Rightarrow^2 < 3,84$$



Um intervalo de 95% de confiança para π é $(0,385; 0,782)$

3- Aproximação Binomial pela Normal

$$X \sim bin(n, \pi)$$

pode ser aproximado por

$$X \sim N(n\pi, n\pi(1-\pi))$$

$$\text{valor-p} = 2P[X \geq 12|\pi = 1/2] = 2P[Z \geq 2/\sqrt{5}] = 0,371$$

Pode-se mostrar que, neste caso, é exatamente igual ao teste qui-quadrado.

3- Aproximação Normal: Intervalo de Confiança

$$\hat{\pi} \pm 1,96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{20}}$$

$$0,60 \pm 1,96\sqrt{\frac{0,24}{20}} \Rightarrow (0,39; 0,82)$$

4 - Simulação de Monte Carlo

- $X \sim bin(20, \pi)$
- Sob $H_0 : X \sim bin(20, 1/2)$
- Gerar 2000 $bin(20, 1/2)$

Vamos utilizar a estatística Qui-Quadrado

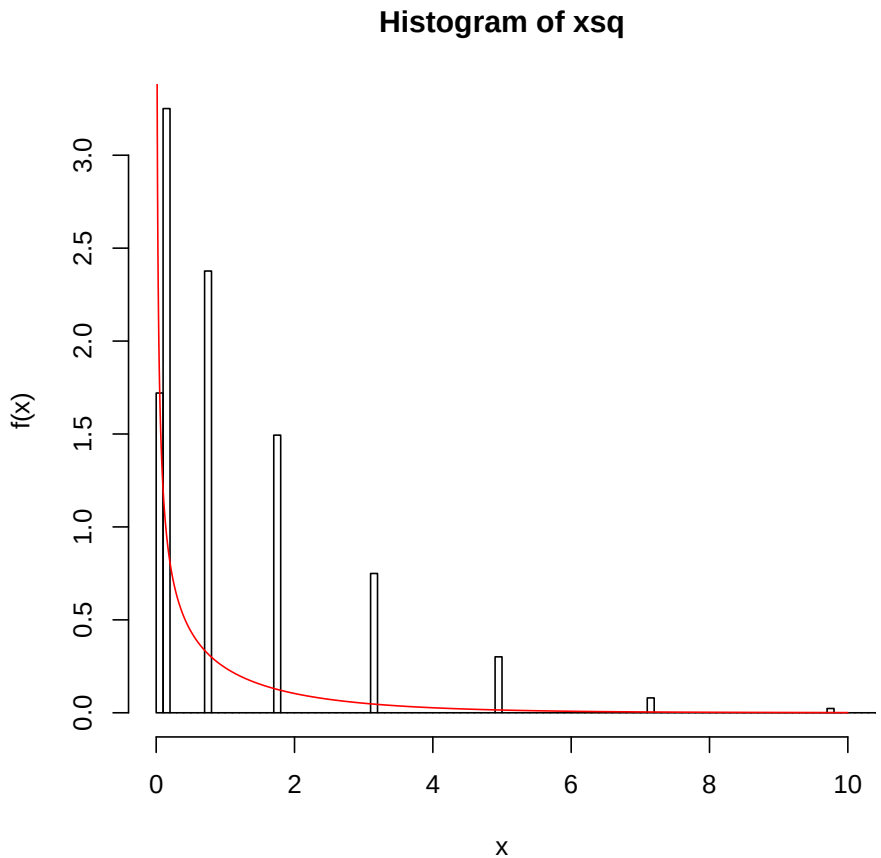
$$X_j^2 = \sum_{i=1}^{2} \frac{(n_{ij} - E_i)^2}{E_i}; \quad j = 1, \ldots, 2000$$

$$\text{valor-p} = \frac{\#X_j^2 \geq X_0^2 = 0,8}{2000} \approx 0,500$$

Podemos comparar a distribuição de X^2 com χ_1^2

4- Simulação de Monte Carlo

É possível desenhar a verdadeira distribuição de χ^2 fazendo um histograma dos 2000 valores observados de $\chi^2_1, ..., \chi^2_{2000}$



Intervalo de 95% de confiança para π é $(0,4; 0,8)$ (gerando $\pi = 0,6$).

5- Bootstrap (Efron e Tibshirani, 1993)

Amostragem com reposição sob H_0

- Efron e Tibshirani (1993, p.224-27).
- Gerar b=2000 amostras com reposição e
- $$X_j^2 = \sum_{i=1}^{2} \frac{(n_{ij} - E_i)^2}{E_i} = \frac{2[(n_{ij} - 2) - 10]^2}{10}$$

  j= 1,...,b
- $$\text{Valor-p} = \frac{\#X_j^2 \geq \chi_0^2 = 0,8}{2000} \approx 0,500$$

Bootstrap (Efron e Tibshirani, 1993)

Intervalo de confiança Bootstrap percentílico

$$\underline{P}[50] = 0,40$$

$$\bar{P}[1950] = 0,80$$

Inferência Clássica

Inferência clássica foi a estudada até o momento

Frequentista: frequência relativa de diferentes resultados em um número grande de experimentos repetidos;

1. Toda a informação sobre π está contida na amostra;
2. π é fixo e desconhecido;
3. Uso de aproximações (resultados assintóticos);
4. Interpretados por repetição/frequência;
5. Dificuldade numérica para encontrar máximo de funções.

Inferência Bayesiana

A probabilidade é uma medida subjetiva da ignorância sobre π.

1. A informação está contida na amostra e na informação subjetiva a priori;
2. π é aleatório (representando a incerteza sobre a quantidade desconhecida);
3. Interpretações são probabilísticas;
4. Dificuldade de quantificação da distribuição a priori;
5. Dificuldade numérica: integrar funções.

Enfoque Bayesiano

Uso de probabilidade para quantificar incerteza.

- $\pi \sim p(\pi)$: distribuição a priori
- $x|\pi \sim bin(n, \pi)$: dados (verossimilhança)

Regra de Bayes: encontrar a distribuição a posteriori de π

$$p(\pi|x) = \frac{p(\pi,x)}{p(x)} = \frac{p(x|\pi)p(\pi)}{p(x)} \propto p(x|\pi)p(\pi)$$

p(x): constante proporcionalizadora (preditiva).

6- Retornando ao exemplo de Ecologia

$p(\pi)$: Beta (α, β) conjugada da binomial.

- $p(\pi) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}$; $0 < \pi < 1$
- $p(\pi|x) \propto [\pi^{\alpha-1}(1-\pi)^{\beta-1}]\pi^{x}(1-\pi)^{n-x}$
- $p(\pi|x) \propto \pi^{\alpha+x-1}(1-\pi)^{n+\beta-x-1}$
- $beta(\alpha + x, n + \beta - x)$

Inferência Bayesiana

- Como especificar α e β? (Informação a priori)

  Por exemplo, utilizar $\alpha = \beta = 1$, distribuição uniforme.

  Então temos:

  $\pi|x \sim beta(1 + x, n - x + 1)$

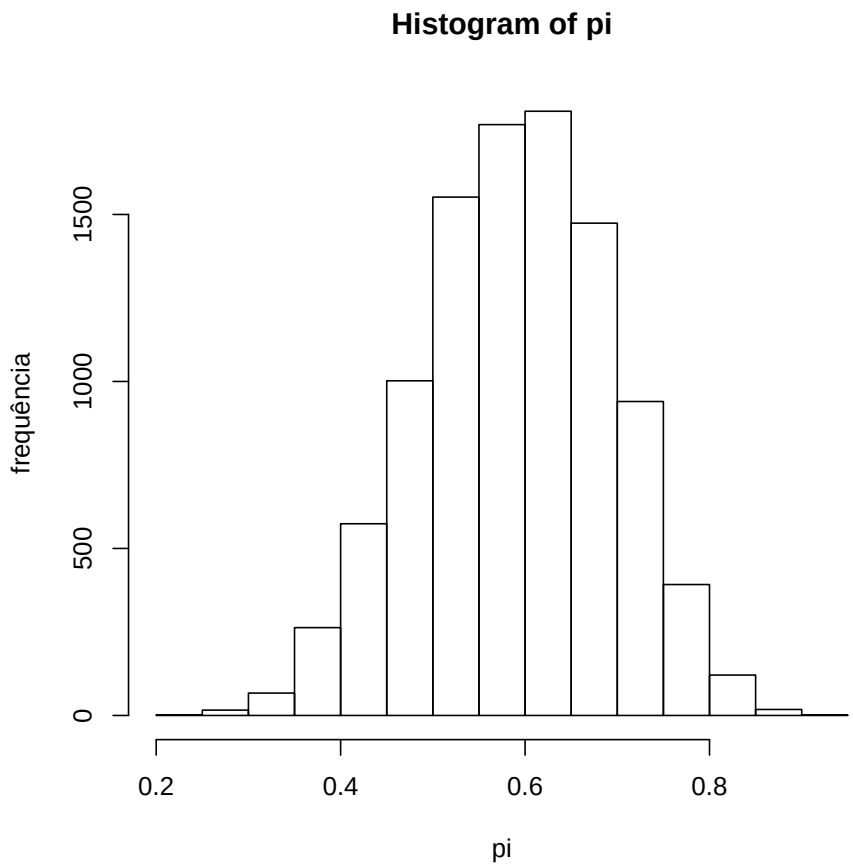
- Nosso caso: $\pi|x \sim beta(13, 9)$

Resumindo Inferência a posteriori

- Média = $E(\pi) = \frac{\alpha}{\alpha+\beta} = \frac{13}{13+9} = \frac{13}{22}$
- *Moda* $= \frac{\alpha-1}{\alpha+\beta-2} = \frac{12}{20}$
- Variância$= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{13\times9}{22^2\times23}$
- I.C Assintótico $\frac{13}{22} \pm 1,96\sqrt{Var(\pi)} = (0,39; 0,79)$

Int. Credibilidade a posteriori

Usando MC a partir da *beta* $(13, 9)$



(0,38;0,78)

Resumo dos Resultados

Teste	Valor-p	I.C.
Exato	0,503	(0,361;0,809)
Qui-Quadrado	0,371	(0,385;0,782)
Aprox. Normal	0,371	(0,385;0,815)
Monte Carlo	$\approx$0,50	(0,4;0,8)
Bootstrap	$\approx$0,50	(0,4;0,8)
Bayesiano		(0,38;0,78)