THE UNIVERSITY OF KANSAS Medical Center
Department of Biostatistics

PALESTRA 3

19 de outubro de 2011

Jonathan D. Mahnken, Ph.D., PStat®

Parte I - Esboço

- Comparando dois grupos
  - Inferência para comparar médias
  - Inferência para comparar medianas
  - Inferência para comparar proporções
  - Estimativa de tamanho de amostra

Inferência sobre Dois Grupos

- Métodos que usamos para tirar conclusões sobre duas populações de amostras
  - Intervalos de confiança
  - Teste de hipótese
- Suposições
  - Amostra aleatória
  - Amostra Representativa
  - Amostras são independentes
    - "…sabendo que as observações para um grupo não fornecem nenhuma informação sobre as observações no segundo grupo." (p 133)
    - Diferente de um grupo observado duas vezes

Inferência sobre Dois Grupos

- Dados numéricos
  - Teste-T
    - Teste-T de duas amostras
  - Teste da soma dos postos de Wilcoxon
- Dados categóricos
  - Métodos de distribuição-z
  - Teste de qui-quadrado
  - Teste exato de Fisher

Inferência sobre as Médias

- Intervalos de confiança
  - *CI = ESTATÍSTICO* $\pm (z_{\alpha/2})(SE_{ESTATÍSTICO})$
  - Estima a diferença entre as médias do grupo

$$\overline{X}_1 - \overline{X}_2$$

  - Suposições
    - Cada amostra ~ $N(\mu_i,\sigma^2)$
    - Homogeneidade de variância
      - Não é necessária quando os tamanhos da amostra são iguais
    - Grupos independentes

Inferência sobre as Médias

- Intervalos de confiança
  - Desvio Padrão combinado

$$DP_p = \sqrt{\frac{(n_1 - 1)DP_1^2 + (n_2 - 1)DP_2^2}{n_1 + n_2 - 2}}$$

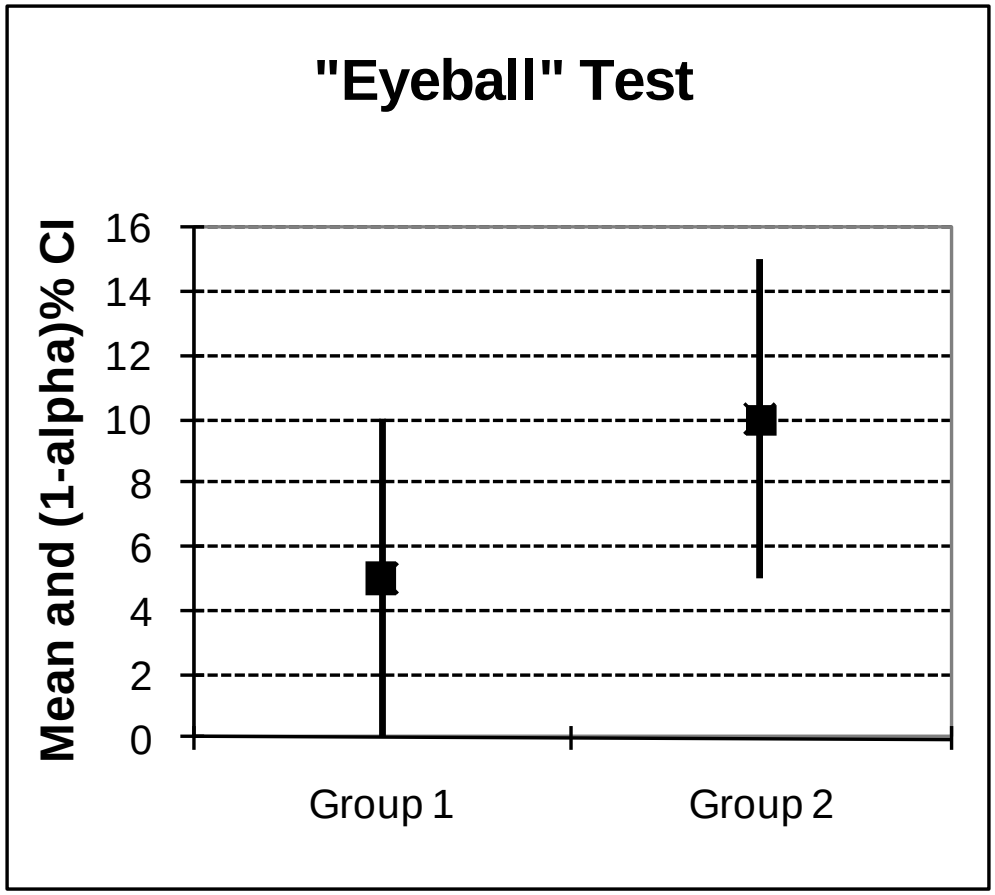
  - Erro padrão da diferença

$$SE_{\bar{X}_1 - \bar{X}_2} = DP_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Inferência sobre as Médias

- Teste do “Olho” (Eyeball Test)
  - Lote
    - Média de cada grupo
    - (1-α)% CI
      - Barras/margens de erro
  - *É <u>sempre</u> bom olhar para seus dados!*

Teste do “Olho” (Eyeball Test)



Inferência sobre as Médias

- Teste-T de duas amostras
  - H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 \neq \mu_2$
  - Teste no qual as médias do grupo são as mesmas (sem diferença)

$$\overline{X}_1 - \overline{X}_2 = 0$$

  - Suposições
    - Cada amostra ~ $N(\mu_i, \sigma^2)$
    - Homogeneidade de variância
      - Não é necessária quando os tamanhos da amostra são iguais
    - Grupos independentes

Inferência sobre as Médias

- Teste-T de duas amostras

$$t_{(n_1+n_2-2)} = \frac{\bar{X}_1 - \bar{X}_2}{SE_{\bar{X}_1-\bar{X}_2}}$$

onde

$$SE_{\bar{X}_1-\bar{X}_2} = DP_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

e

$$DP_p = \sqrt{\frac{(n_1-1)DP_1^2 + (n_2-1)DP_2^2}{n_1+n_2-2}}$$

Inferência sobre as Médias

- Suposição de Homogeneidade de variâncias
  - Não necessária para grupos de mesmo (ou quase mesmo) tamanho
  - Testes para Homogeneidade de variâncias
    - Test-F
    - Teste de Levene
  - Se há heterogeneidade presente
    - Correção de Satterthwaite
      - Faz com o que teste-T fique mais conservativo
    - Método não paramétrico
    - Transformação

Inferência sobre as Medianas

- Teste da soma dos postos de Wilcoxon
  - a.k.a. Mann-Whitney Teste-U
  - Procedimento não paramétrico
  - H_0: $\text{mediana}_1 = \text{mediana}_2$ versus H_1: $\text{mediana}_1 \neq \text{mediana}_2$
  - Valores observados substituídos por postos
    - Probabilidades exatas
      - Computacionalmente cara
    - Probabilidades aproximadas
      - Teste-T nos postos

Inferência sobre as Proporções

- Intervalos de confiança
  - $CI = ESTATÍSTICO \pm (z_{\alpha/2})(SE_{ESTATÍSTICO})$
  - Estima a diferença entre as médias do grupo
    - $\pi_1 - \pi_2$
  - Observações Binomiais
  - Aproximação normal à binomial
    - $np > 5$
  - Suposições
    - $n_i p_i > 5$ para $i = \{1,2\}$
    - Grupos independentes

Inferência sobre as Proporções

- Intervalos de confiança
  - "A estimativa combinada de p fornece uma estimativa melhor para se usar no erro padrão…" (p 145)

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

  - Erro padrão para a diferença em proporções

$$SE_{p_1 - p_2} = \sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}$$

Inferência sobre as Proporções

- Teste de hipótese
  - H_0: $\pi_1 = \pi_2$ versus H_1: $\pi_1 \neq \pi_2$
    - Testa que as proporções dos grupos são a mesma (sem diferença)
  - Observações binomiais
  - Aproximação normal à binomial
    - $np > 5$
  - Suposições
    - $n_i p_i > 5$ para i = {1,2}
    - Grupos Independentes

Inferência sobre as Proporções

- Teste de hipótese
  - Teste-Z

$$z = \frac{p_1 - p_2}{SE_{p_1 - p_2}}$$

onde

$$SE_{p_1 - p_2} = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Inferência sobre as Proporções

- Teste do qui-quadrado (χ^2)
  - Qui-quadrado de Pearson
  - Mesma conclusão como o CI anterior e teste de hipótese usando distribuição-Z
    - Teste aproximado
  - As perguntas do teste qui-quadrado respondem:
    - Há uma diferença nas proporções?
    - Há uma associação entre as variáveis?
    - As variáveis são independentes?
  - Mesmo resultado, independentemente da questão

Inferência sobre as Proporções

- Teste qui-quadrado
  - Compara o observado versus esperado
  - H_0: A,B independentes (sem associação) vs. H_1: A,B não independentes (associadas)
    - Abaixo H_0: $P\{A \cap B\} = P\{A\} \times P\{B\}$
      - Regra de multiplicação
  - Frequências esperadas geradas baseado no produto de frequências marginais
  - Se a variação entre as frequências observadas e esperadas forem maiores que a esperada por acaso, rejeite H_0

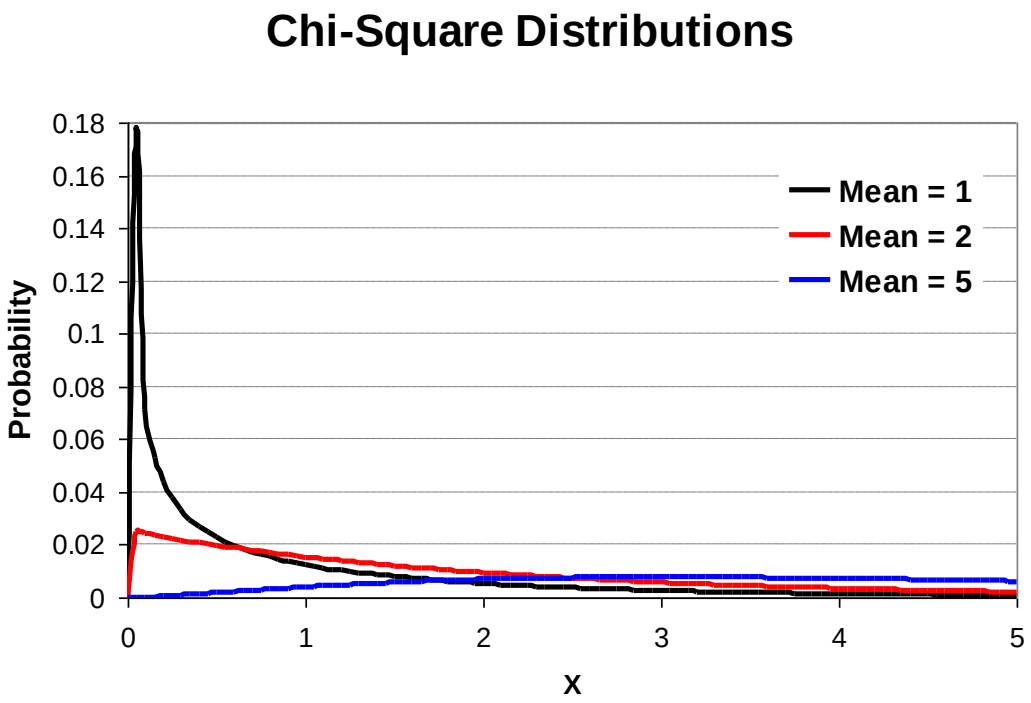
Inferência sobre as Proporções

Variável A	Variável B		
	Sim	Não	Total
Sim	n_{11}	n_{12}	n_{1+}
Não	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Abaixo H_0: $P\{n_{jk}\} = P\{n_{+k}\} \times P\{n_{j+}\}$ por regra de multiplicação!

Inferência sobre as Proporções

- Teste qui-quadrado
  - A forma depende dos graus de liberdade (gl)



Inferência sobre as Proporções

- Teste qui-quadrado
  - Graus de liberdade (gl)
    - Dados os totais marginais, quantas células podem variar
  - $$\chi^2_{(df)} = \sum \frac{(O-E)^2}{E}$$
  - *Valores mais extremos que χ^2 são evidência contra a hipótese nula de independência!*

Inferência sobre as Proporções

- Teste exato de Fisher
  - Não paramétrico
  - Contagem de células pequenas *esperadas*
  - Muito entediante e computacionalmente caro
    - Rapidamente se transforma difícil demais para computadores calcularem

Estimativa de tamanho de amostra

- Compara as <u>médias</u> de dois grupos
  - Qual é α?
  - Qual é β?
    - Potência = 1-β
  - O que é uma diferença clinicamente importante?
    - $\mu_1 - \mu_2 = \Delta$
  - Quais são boas estimativas de σ_1 e σ_2?
    - Considere $\sigma_1 = \sigma_2$
- $n = 2(z_\alpha - z_\beta)^2(\sigma/\Delta)^2$ (n = tamanho de ***<u>cada</u>*** grupo)
  - *Sempre arredonde para o maior número inteiro!*

Estimativa de tamanho de amostra

- Compara as <u>proporções</u> de dois grupos
  - Qual é α?
  - Qual é β?
    - Potência = 1-β
  - O que é uma diferença clinicamente importante?
    - $\pi_1 - \pi_2 = \Delta$
  - σ é função de π
    - determinando π também determina σ
- $n = \{z_\alpha[2\pi_1(1-\pi_1)]^{1/2} - z_\beta[\pi_1(1-\pi_1)+\pi_2(1-\pi_2)]^{1/2}\}^2(1/\Delta)^2$
  (n = tamanho de *<u>cada</u>* grupo)
  - *Sempre arredonde para o maior número inteiro!*

Parte II - Esboço

- Relações entre variáveis/medidas
  - Comparando variáveis contínuas
    - Correlação
      - Abordgem paramétrica
      - Abordgem não paramétrica
    - Regressão linear
    - Regressão múltipla
  - Comparando variáveis dicotômicas
    - Razão das chances
    - Razão do risco

Inferência sobre Relações Entre Variáveis

- Resultado Numérico e preditor
  - Correlação
    - Coeficiente de correlação de Pearson
    - Coeficiente de correlação de Spearman
  - Regressão Linear
    - Regressão simples
    - Regressão dos mínimos quadrados ordinários (OLS)
- Resultado Categórico e preditor
  - Razão de Risco
  - Razão das chances

Coeficiente de correlação de Pearson

- Parâmetro de população = ρ
- Estatística da amostra= r
  - Medida da relação *linear*
  - Sem unidade, [-1,1]
- *Pode deixar passar uma relação não-linear forte, portanto* ***SEMPRE PLOT SEUS DADOS!***

Coeficiente de correlação de Pearson

- Suposições
  - $(X,Y) \sim BVN(\boldsymbol{\mu},\boldsymbol{\Sigma})$
    - $X \sim N(\mu_X,\sigma_X^2)$ e $Y \sim N(\mu_Y,\sigma_Y^2)$ ***NEM*** sempre implicam que $(X,Y) \sim BVN(\boldsymbol{\mu},\boldsymbol{\Sigma})$
  - (X_i,Y_i) são independentes
- Quando as suposições não são satisfeitas, os dados podem ser transformados ou métodos não paramétricos podem ser utilizados

Coeficiente de correlação de Pearson

- Teste de hipótese
  - H_0: $\rho = 0$
    - Distribuição simétrica
    - Distribuição-T
  - H_0: $\rho = \rho_0$ ($\neq 0$)
    - Distribuição assimétrica pois ponto de corte além [-1,1]
    - Transformação-Z de Fisher
- r significante $\Rightarrow$ regressão linear significante
- Intervalos de confiança (o mesmo que acima)

Coeficiente de Determinação

- Coeficiente de Determinação= r^2
  - [0,1]
  - [Coeficiente de Determinação]% da variação em uma medida pode ser explicada por se conhecer o valor da outra medida

Coeficiente de Correlação Ordinal de Spearman

- Parâmetro de população= ρ_S
- Estatística da amostra= r_S
- Útil quando as observações são assimétricas ou contém valores atípicos (outliers)
- Observações de posto
- Coeficiente de correlação de Pearson quando nenhum dos postos estiverem empatados
- Medida de relação *linear* de *postos*

Coeficiente de Correlação Ordinal de Spearman

- Teste de hipótese e intervalos de confiança
  - Ordena os dados
  - Utiliza métodos para o coeficiente de correlação de Pearson
- Mais complexo onde há empates

Regressão Linear

- "…o objetivo é predizer o valor de uma característica a partir do conhecimento de outra…" (p 194)
- Relação Linear
- Regressão simples
  - Um preditor
- Regressão Múltipla
  - Mais de um preditor

Regressão Linear

- Equação de regressão da população
  - $Y = \beta_0 + \beta_1 X + \varepsilon$
- Equação de regressão da amostra
  - $Y' = a + bX$
    - Y' é o valor previsto de Y
      - Frequentemente chamado de Y-hat
    - a é uma estimativa de β_0 (y-intercepção)
      - Valor previsto de Y quando X = 0
    - b é uma estimativa de β_1 (declive)
      - Mudança prevista em Y para cada mudança de 1-unidade em X
  - $e = Y - Y'$
    - residual

Regressão Linear

- Mínimos Quadrados Ordinários (OLS)
  - Minimiza a soma do quadrado da distância vertical entre Y e Y' (observado e previsto)

$$\min\left[\sum_i (Y_i - \hat{Y}_i)^2\right]$$

- Relação entre OLS e o coeficiente de correlação de Pearson

Regressão Linear

- Suposições
  - Equação de regressão da população subjacente
    - $Y = \beta_0 + \beta_1 X + \varepsilon$
  - $a + bX$ é uma estimativa de $\beta_0 + \beta_1 X$
  - $\varepsilon \sim N(0,\sigma^2)$
    - Homocedasticidade
  - Relação Linear
  - Y_i independente de $Y_j \; \forall \; i \neq j$

Regressão Linear

- “Regressão é um procedimento robusto e pode ser utilizado em muitas situações nas quais as suposições não são atendidas, desde que as medições sejam bastante confiáveis e que o modelo de regressão correto seja utilizado.” (p 197)

Regressão Linear

- Teste de hipótese
  - Testa ***ambos*** β_0 e β_1 separadamente
    - Ambos irão variar
  - As Fórmulas são diferentes para cada um
    - Utilize SE da regressão
  - H_0: $\beta_0 = 0$
  - H_0: $\beta_1 = 0$
  - Teste-T
- Intervalos de confiança utilizam as mesmas estimativas do SE como exibido acima

Regressão Linear

- Média prevista = observação prevista
- CI para a média $\neq$ PI para a observação
  - Maior variação quando se prevê uma única observação (vs. uma média)
  - Menor variação (para ambas médias e observações previstas) quando X próximo de diagrama de X-barra

Regressão Linear

- Comparando-se duas linhas de regressão
  - "… a abordagem preferida é a utilização de modelos de regressão para mais de uma variável independente – um procedimento chamado de regressão múltipla – para responder a essas perguntas." (p 201)
    - "O modelo mais simples – utilizando o menor número possível de variáveis explicativas para explicar os fenômenos adequadamente – é então selecionado." (p 201)

Regressão Linear

- Residuais
  - e = Y – Y’
  - Plot deve ser espalhado aleatoriamente ao redor de e = 0
    - r_{ex} deve igualar a 0
    - Sem padrões

Regressão Múltipla

- Regressão Linear com mais de uma variável preditora
  - R^2 é a quantidade da variação em Y descrita conhecendo X_1, X_2,…
  - Quando apenas um preditor, $R^2 = r^2$
    - r é a medida da relação ***independente*** da escala
    - $r_{YY'} = r_{YX}$ pois Y' = a + bX é apenas um reescalonamento de X

Razão de Chances

Exposição	Doença	
	Sim	Não
Sim	a	b
Não	c	d

$$RC = \frac{ad}{bc}$$

Razão de Chances

- Intervalo de confiança para RC
  - Encontrar SE de ln[OR]

$$\ln\left(SE\left[RC\right]\right)=\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

  - Então

$$CI_{(1-\alpha)\%}=\exp\left(\ln\left[RC\right]\pm z_{\alpha/2}\cdot\ln\left(SE\left[RC\right]\right)\right)$$

- Teste de hipótese feito em ln[RC] e utiliza SE de ln[RC] para teste-t

Razão de Risco

Exposição	Doença	
	Sim	Não
Sim	a	b
Não	c	d

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

Razão de Risco

- **Intervalo de confiança para RR**
  - **Encontre SE de ln[RR]**

$$\ln\left(SE\left[RR\right]\right) = \sqrt{\frac{1-\left[a/(a+b)\right]}{a} + \frac{1-\left[c/(c+d)\right]}{d}}$$

  - **Então**

$$CI_{(1-\alpha)\%} = \exp\left(\ln\left[RR\right] \pm z_{\alpha/2} \cdot \ln\left(SE\left[RR\right]\right)\right)$$

- **Teste de hipótese feito em ln[RR] e usa SE de ln[RR] para teste-t**

Referência

- Dawson B and Trapp RG (2001). *Basic & Clinical Biostatistics*, 3rd ed., McGraw Hill: New York