Estimação pontual e intervalar

Principais estimadores e procedimentos para intervalos de confiança

Prof. Me. Lineu Alberto Cavazani de Freitas

Departamento de Estatística
Laboratório de Estatística e Geoinformação

UNIVERSIDADE FEDERAL DO PARANÁ

Inferência

Inferência

- Temos interesse em determinada característica (parâmetro) na **população**, alguma medida tal como uma média, variância, proporção, etc.
- Com base nos dados (evidência amostral), precisamos **estimar** os **parâmetros**.
- Uma **estimativa**/**estatística** é uma quantidade calculada **a partir dos dados**.
- A distribuição de probabilidades de uma estimativa/estatística é chamada **distribuição amostral**.

Inferência

- Uma **estimativa pontual** fornece apenas um valor plausível de ser o verdadeiro valor do parâmetro.

- Uma **estimativa intervalar/intervalo de confiança** leva em conta a incerteza devido a termos apenas uma amostra.
    - É uma faixa de valores com determinada probabilidade de conter o parâmetro.

Em resumo

1. Definimos a variável aleatória de interesse na população (Y, por exemplo).
2. Esta variável tem o comportamento dado pela sua distribuição ($Y \sim f(\theta)$).
3. Tomamos uma amostra aleatória ($Y_1, Y_2, ..., Y_n$) cujos elementos são independentes e identicamente distribuídos, seguindo a mesma distribuição de Y. Ou seja, $Y_i \sim f(\theta)$.
4. Estamos interessados em estudar algum parâmetro (θ) na população.
5. Estimamos esta característica com base na amostra usando algum estimador ($\hat{\theta} = f(Y_1, Y_2, ..., Y_n)$).
6. Expressamos a incerteza associada a esta estimativa (por estarmos usando uma amostra).
7. Avaliamos hipóteses sobre esta estimativa.
8. Interpretamos e tiramos conclusões.

Propriedades dos estimadores

Propriedades dos estimadores

- Mais de uma função da amostra pode ser proposta para estimar um parâmetro de interesse (por exemplo, no caso da variância).
- Para facilitar a escolha entre estimadores, é importante verificar se possuem algumas características desejáveis.
- Algumas características a serem verificadas são: vício, consistência e eficiência.
- Um bom estimador é: não viciado, consistente e eficiente.

Vício

- Um estimador $\hat{\theta}$ é não viciado ou não viesado para um parâmetro θ se seu valor esperado coincide com o parâmetro de interesse.

- $E(\hat{\theta}) = \theta$.

- Independente do tamanho amostral esta propriedade deve ser válida.

Consistência

- Um estimador $\hat{\theta}$ é consistente, se:
  - À medida que aumenta-se o tamanho amostral, o valor esperado converge para o parâmetro de interesse: $\underset{n\to\infty}{lim} E(\hat{\theta}) = \theta$.
  - À medida que aumenta-se o tamanho amostral, a variância converge para 0: $\underset{n\to\infty}{lim} Var(\hat{\theta}) = 0$.
- Note como o conceito de consistência está diretamente ligado ao tamanho amostral, diferentemente do conceito de vício.

Eficiência

Dados dois estimadores $\hat{\theta}_1$ e $\hat{\theta}_2$, ambos não viciados para um parâmetro θ, dizemos que $\hat{\theta}_1$ é mais eficiente que $\hat{\theta}_2$ se $Var(\hat{\theta}_1) < Var(\hat{\theta}_1)$, ou seja, quanto menor a variância, maior a eficiência.

Propriedades dos estimadores

$$\hat{\mu} = \bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

$$\hat{p} = \frac{\text{número de sucessos}}{n}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- Os estimadores apresentados para μ e p são não viciados e consistentes.
- A expressão da variância populacional (n no denominador) é viciada e consistente, já a expressão amostral (n-1 no denominador) é não viciada e consistente

Distribuições amostrais

Média e variância

- Se σ^2 é conhecido
  - $\hat{\mu} \sim N(\mu; \sigma^2/n)$.
  - $(\hat{\mu} - \mu)/(\sigma/\sqrt{n}) \sim N(0; 1)$.
- Se σ^2 é desconhecido
  - $(\hat{\mu} - \mu)/(S/\sqrt{n}) \sim t_{n-1}$.
  - $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

Proporção

- $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$.

Intervalos de confiança

Intervalos de confiança

- Uma estimativa pontual é um valor candidato ao parâmetro de interesse baseado em uma amostra.
- Por ser baseado na amostra, a estimativa pontual carrega consigo uma incerteza.
- Um intervalo de confiança é uma faixa de valores com determinada probabilidade de conter o parâmetro.
- Usa a estimativa pontual e informações da distribuição amostral.

Intervalos de confiança

- Para média, proporção e variância as operações são razoavelmente simples.
- O ponto mais importante é a fixação do nível de confiança $(1 - \alpha)$.
- O nível de confiança é um número entre 0 e 1 que determina os valores limites a serem usados nos intervalos.
- É uma quantidade que define a probabilidade do intervalo conter o parâmetro.

Intervalos de confiança

- O intervalo de confiança é calculado a partir de uma amostra, logo o intervalo também é aleatório.
- O valor do parâmetro é fixo, quem varia são as estimativas e os intervalos (de acordo com a amostra).
- Como o valor do parâmetro é fixo, é o intervalo que deve conter o valor do parâmetro, e não o contrário.

Intervalos de confiança

Interpretações, considerando um nível de confiança de $(1 - \alpha) \times 100\%$:

- ERRADA: temos $(1 - \alpha) \times 100\%$ de confiança de que o parâmetro se encontra no intervalo.
- CORRETA: temos $(1 - \alpha) \times 100\%$ de confiança de que o intervalo contém o parâmetro.

Forma alternativa de interpretação:

- Se pudermos obter 100 amostras e calcular um intervalo de confiança de 95% para cada uma das 100 amostras, esperamos que 5 destes intervalos não contenham o verdadeiro valor do parâmetro.
- A interpretação é análoga para outros níveis de confiança.

Intervalos de confiança

Podemos definir o nível de confiança como 100%?

- Quanto maior o nível de confiança, maior será o intervalo de confiança associado.
- Um intervalo muito grande deixa de ser informativo.
- Por isso, o nível de confiança age como uma espécie de compromisso entre risco de errar a inferência e fornecer uma informação com uma precisão e interpretação razoável.

Intervalos de confiança

Passos gerais

1. Verificar os requisitos.
2. Determinar o nível de confiança.
3. Encontrar os valores críticos.
4. Calcular os limites superior e inferior do intervalo.
5. Interpreta os resultados.

Veremos

- Intervalo de confiança para média com variância populacional conhecida.
- Intervalo de confiança para média com variância populacional desconhecida.
- Intervalo de confiança para proporção.
- Intervalo de confiança para variância.

Intervalo de confiança para a média (μ) com variância populacional (σ^2) conhecida

IC para a média com variância conhecida

- Supondo que σ^2 é conhecido:

$$\bar{Y} \sim N\left(\mu; \frac{\sigma^2}{n}\right) \text{ ou } \frac{(\bar{Y} - \mu)}{(\sigma/\sqrt{n})} \sim N(0, 1).$$

- Para obter um intervalo de confiança para média com σ conhecido, as seguintes condições devem ser atendidas:
  - A amostra deve ser aleatória simples.
  - σ deve ser conhecido.
  - A população deve seguir distribuição Normal ou $n > 30$ (regra empírica TLC).

IC para a média com variância conhecida

- Ao fixar uma probabilidade $1 - \alpha$ podemos encontrar os limites inferior ($\bar{y}_{LI}$) e superior ($\bar{y}_{Ls}$) do intervalo de confiança, tal que

$$P(y_{LI} < \mu < y_{LS}) = 1 - \alpha.$$

- Vários intervalos podem gerar o resultado acima, trabalharemos com intervalos simétricos em relação à estimativa pontual.

IC para a média com variância conhecida

- Para obtenção do intervalo basta definir os limites de Z na distribuição amostral padronizada.

$$P\left(z_{LI} < \frac{\bar{y}-\mu}{\sigma/\sqrt{n}} < z_{LS}\right) = 1-\alpha.$$

- Isolando μ e garantindo intervalos simétricos temos que:

$$P\left(\bar{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

- $Z_{\alpha/2}$ é o quantil da distribuição Normal padrão para o valor de $1-\alpha$ fixado.
- Valores comuns para $1-\alpha$ são 90, 95, 99, contudo qualquer valor é possível.

IC para a média com variância conhecida

Logo:

$$IC(\mu)_{1-\alpha} = \bar{y} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Intervalo de confiança para a média (μ) com variância populacional (σ^2) desconhecida

IC para a média com variância desconhecida

- Supondo que σ^2 é desconhecido:

$$t = \frac{(\bar{Y} - \mu)}{(S/\sqrt{n})} \sim t_{n-1}.$$

- A notação t_{n-1} denota a distribuição t-Student com n - 1 graus de liberdade.

IC para a média com variância desconhecida

- Para obter um intervalo de confiança para média com σ desconhecido, as seguintes condições devem ser atendidas:
  - A amostra deve ser aleatória simples.
  - σ é desconhecido mas exite uma estimativa s.
  - A população deve seguir distribuição Normal ou $n > 30$ (regra empírica TLC).

- A expressão do intervalo de confiança é similar à do caso para variância conhecida, apenas a distribuição se altera:

$$P\left(\bar{y} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{y} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1 - \alpha.$$

IC para a média com variância desconhecida

Logo:

$$IC(\mu)_{1-\alpha} = \bar{y} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

Intervalo de confiança para proporção

IC para proporção

- Seja $Y_i \sim Bernoulli(p)$:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right).$$

- Para obter um intervalo de confiança para a proporção, as seguintes condições devem ser atendidas:
  - A amostra deve ser aleatória simples.
  - A variável é binária (sucesso ou fracasso).
  - As tentativas são independentes e a probabilidade de sucesso constatnte (binomial).
  - $np \geq 5$ e $np(1-p) \geq 5$ (garante aproximação com a distribuição normal).

IC para proporção

- De maneira análoga aos casos anteriores temos que:

$$P\left(\hat{p} - Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha.$$

IC para proporção

$$P\left(\hat{p} - Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha.$$

- Perceba um problema: a expressão do intervalo de confiança depende do p verdadeiro.
- Existem duas alternativas:
  - Otimista: usar $\hat{p}$ dentro da raiz.
  - Conservativo: usar 0.5 no lugar de p dentro da raiz. Esta alternativa vai conduzir ao maior intervalo de confiança possível para a proporção.

IC para proporção

Logo:

$$IC(p) = \hat{p} \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Intervalo de confiança para variância

IC para variância

- Seja $Y_i \sim N(\mu, \sigma^2)$:

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$

em que $n-1$ representa os graus de liberdade.

IC para variância

- Neste caso, o intervalo fica dado por:

$$IC_{1-\alpha}(\sigma^2) = \left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}; \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right),$$

em que $\chi^2_{\alpha/2,n-1}$ e $\chi^2_{1-\alpha/2,n-1}$ são os quantis da cauda direita e da cauda esquerda da distribuição χ^2 com $n-1$ graus de liberdade.

- Neste caso o intervalo de confiança não é simétrico como no caso da média e da proporção pois a distribuição amostral não é simétrica.

Intervalos

De forma geral:

- IC p/ μ com σ^2 conhecido: $IC(\mu) = \bar{y} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

- IC p/ μ com σ^2 desconhecido: $IC(\mu) = \bar{y} \pm t_{\alpha/2}\frac{S}{\sqrt{n}}$

- IC p/ p: $IC(p) = \hat{p} \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$

- IC p/ σ^2: $IC_{1-\alpha}(\sigma^2) = \left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}; \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right)$

O que foi visto:

- Principais estimadores e estimativas.
- Propriedades dos estimadores.
- Intervalos de confiança.
  - Média com variância conhecida.
  - Média com variância desconhecida.
  - Proporção (otimista e conservativo).
  - Variância

Próximos assuntos:

- Tamanho amostral.
  - Média com variância conhecida.
  - Média com variância desconhecida.
  - Proporção.