UNIVERSIDADE DE SÃO PAULO
Faculdade de Zootecnia e Engenharia de Alimentos

ZAB1111 – Estatística Básica

Aula 11
INTRODUÇÃO À INFERÊNCIA ESTATÍSTICA

Material preparado pelo Prof. Dr. César Gonçalves de Lima – FZEA/USP

1. INTRODUÇÃO À INFERÊNCIA ESTATÍSTICA

Já vimos:

- Análise descritiva: gráficos, tabelas e distribuições de frequências.
- Medidas de tendência central: média, moda e mediana.
- Medidas de dispersão: desvio médio, variância, desvio padrão, coeficiente de variação.
- Medidas de assimetria e de curtose. Percentis.
- Probabilidade
- Modelos probabilísticos para v.a. discretas (Bernouli, Binomial e Poisson) e v.a. contínuas (Uniforme, Exponencial e Normal).

A partir de agora:

Inferência Estatística: estuda como fazer afirmações sobre certas características de uma população, baseando-se em resultados obtidos em uma amostra.

Vale lembrar que:

- **População:** qualquer conjunto de indivíduos ou objetos que têm pelo menos uma variável comum observável.
- **Amostra:** qualquer subconjunto da população.

Exemplo 1.1. Consideremos uma pesquisa feita para estudar o ganho de peso dos bovinos de corte de um rebanho de 700 animais.

Selecionamos uma amostra de 40 animais e anotamos os seus pesos no início e no final do período experimental. A partir dessas informações calculamos o ganho de peso mensal de cada animal.

Esperamos que a distribuição dos ganhos de peso dos animais da amostra reflita bem a distribuição (é simétrica? *normal*?) e as principais características (mesma média? mesmo desvio padrão?) do ganho de peso dos animais de todo o rebanho.

Interesse:

i) Estimar o ganho de peso médio de todos os bovinos de corte.

ii) Testar se o ganho de peso médio desses bovinos, neste particular mês, foi superior a 10 kg.

Um problema anterior:

Como selecionar amostras representativas de uma população?

1.1. COMO SELECIONAR UMA AMOSTRA

Técnicas de amostragem: são formas diferentes de se obter uma amostra representativa da população.

Amostragem probabilística: quando os elementos da população têm probabilidades conhecidas e diferentes de zero de fazer parte da amostra.

- Implica em realizar um sorteio com regras determinadas quando a população for finita e totalmente acessível.
- As observações colhidas em uma amostra serão mais informativas quanto mais conhecermos sobre a população de onde a amostra foi retirada.

Amostragem não probabilística: os elementos da amostra são escolhidos de forma não aleatória porque são mais facilmente acessíveis ou se acredita que sejam representativos da população.

- É bastante utilizada, embora exista um grande risco de ser parcial.
- É perigoso usar uma amostra deste tipo para tirar alguma conclusão importante.

A amostragem não probabilística é necessária quando:

1) É difícil identificar a população alvo.

   Exemplo: Como estudar o comportamento dos *hackers* durante a pandemia? É difícil identificá-los.

2. A população designada é muito específica e de disponibilidade limitada.

   Exemplo: Para estudar o comportamento de executivos de empresas que empregam mais de 10 Engenheiros de Biossistemas, podemos ser obrigados a trabalhar somente com os executivos dispostos a participar.

3. A amostra é de um estudo piloto (inicial) que não será usada na pesquisa final e só temos um pequeno grupo de pessoas disponíveis.

Nota: As principais técnicas de Inferência Estatística pressupõem que as amostras utilizadas no estudo sejam probabilísticas.

1.2 TÉCNICAS DE AMOSTRAGEM PROBABILÍSTICA

i) **Amostragem casual simples ou aleatória (a.c.s.)**: cada elemento da população tem a mesma probabilidade de ser selecionado, ou seja, a mesma chance de fazer parte da amostra.

- O sorteio dos elementos para compor a amostra poderá ser feito de duas formas: com ou sem reposição.

  Exemplo: Sortear uma a.c.s. de n elementos de uma população finita de tamanho N. O número de amostras possíveis depende do tipo de sorteio:

  - Com reposição: N^n amostras possíveis
  - Sem reposição: $\binom{N}{n}$ amostras possíveis

ii) **Amostragem Sistemática:** é utilizada quando os elementos da população se apresentam ordenados de forma aleatória. A retirada dos elementos da amostra é feita *periodicamente*.

Exemplo: Com o objetivo de estudar a qualidade da refeição oferecida no Campus vamos usar uma amostra de $n = 50$ alunos, asbendo que são consumidas, em média, 500 refeições por dia.

1) Sorteamos um número (k) de 1 a 10 (note que $500/50 = 10$) e aplicamos o questionário ao k-ésimo aluno da fila.

2) Os demais 49 alunos serão escolhidos sistematicamente, de 10 em 10 até que a amostra de 50 alunos esteja completa.

Exemplo: Sorteando o número 5, farão parte da amostra de 50 números o 5º, 15º, 25º, 35º, 45º, ⋯, 495º aluno.

iii) **Amostragem por Conglomerados:** é utilizada quando a população apresenta uma subdivisão natural em pequenos grupos ou conglomerados. Espera-se que esses grupos sejam heterogêneos internamente e reproduzam bem a população.

Para retirarmos uma amostra, sorteamos um número suficiente de conglomerados e os seus elementos constituirão a amostra.

Exemplo: Preciso de uma amostra de 70 bovinos Nelore para realizar uma pesquisa e os 700 animais do rebanho estão distribuídos aleatoriamente em 70 piquetes com 10 animais/piquete.

Solução: Numero os piquetes de 1 a 70, sorteio 7 piquetes e uso os animais desses piquetes sorteados para compor a amostra.

iv) **Amostragem Estratificada**: é utilizada quando a população pode ser dividida em diferentes subpopulações, classes ou estratos. A técnica consiste em especificar quantos elementos da amostra serão retirados de cada estrato.

Supõe-se que a variável de interesse apresente um comportamento diferente de estrato para estrato e um comportamento homogêneo dentro de cada estrato.

- Se o sorteio dos elementos da amostra não considerar tais estratos pode ocorrer que os diversos estratos não sejam convenientemente representados na amostra.
- A amostra pode ser mais influenciada pelas características da variável nos estratos mais favorecidos pelo sorteio.

Exemplo: Desejamos saber a opinião dos alunos de Graduação sobre a qualidade da refeição servida no refeitório do Campus.

Admitindo que a opinião dos alunos sobre a qualidade das refeições pode ser influenciada pelo tempo que eles frequentam o refeitório, podemos reagrupá-los pelo ano de ingresso, formando 5 estratos de tamanhos diferentes. A seguir, sorteamos certa quantidade de alunos dentro de cada estrato, de forma a tornar a amostra representativa da população.

Se não levarmos este aspecto em conta, podemos favorecer subgrupos de alunos que já têm opinião consolidada sobre a qualidade da refeição.

A amostragem estratificada pode ser de três tipos:

i) **Uniforme**: retira-se igual número de elementos em cada estrato, independente do seu tamanho.

ii) **Proporcional**: o número de elementos sorteados em cada estrato é proporcional ao número de elementos existentes no estrato

iii) **Ótima**: retiramos, em cada estrato, um número de elementos proporcional ao número de elementos que o compõem e à variabilidade da variável de interesse no estrato, medida por seu desvio padrão.

1.3 TÉCNICAS DE AMOSTRAGEM NÃO PROBABILÍSTICA

i) **Amostragem por conveniência:** Envolve a obtenção de respostas das pessoas que estão disponíveis e dispostas a participar da pesquisa. O problema principal desta abordagem é que a opinião dessas pessoas pode diferir muito da opinião dos que não estão dispostos a participar da pesquisa.

Exemplo: Nos sítios de compra da web os clientes que têm reclamações podem estar mais dispostos a responder um questionário do que os que estão satisfeitos com a compra do produto ou do serviço.

ii) **Amostragem "bola de neve" (SnowBall):** Algumas pessoas são convidadas a participar da pesquisa e solicita-se a elas que convidem outras pessoas a participar. A amostragem continua até que o número exigido de respostas seja obtido. Essa técnica é frequentemente usada quando a população é difícil de ser identificada ou acessada pelos pesquisadores.

Exemplo: Admitindo que os hackers de software se conheçam, se acharmos um hacker para participar da pesquisa, podemos solicitar a ele que indique/convide outros possíveis participantes.

iii) **<u>Amostragem por cota</u>:** é a versão não probabilística da amostragem (aleatória) estratificada. A população alvo é dividida em estratos apropriados, baseados em subgrupos conhecidos (sexo, grau de instrução *etc*.). Cada estrato é amostrado, usando amostragem por conveniência ou bola de neve, de forma que o número de respondentes em cada estrato corresponde à sua proporção na população.

Exemplo: Numa pesquisa de opinião sobre o atendimento em um açougue de supermercado, aplica-se o questionário até que se obtenha a opinião de 200 clientes, sendo 80 solteiros e 120 casados.

Para maiores detalhes sobre Técnicas de Amostragem consultar:

Bolfarine, H. Bussab, W.O. **Elementos de Amostragem**. São Paulo, SP. Edgard Blücher, 2005.

Cochran, W.G. **Técnicas de Amostragem**, Fundo de Cultura, Rio de Janeiro, 1955.

Mendehall, W. Scheaffer, R.L. Ott. L. **Elementary Sampling**, 6th ed. Southbank: Thomson, 2006.

1.3. INFERÊNCIA

Vamos diferenciar algumas medidas utilizadas para descrever características importantes num conjunto de dados.

- **Parâmetro** é qualquer medida numérica usada para descrever uma característica da população.
- **Estatística** é qualquer medida usada para descrever uma característica da amostra, ou seja, *qualquer* função dos elementos da amostra.

Notação: Geralmente os parâmetros são denotados por letras gregas ou letras maiúsculas do alfabeto romano. Já para as estatísticas são utilizadas letras minúsculas do alfabeto romano ou letras gregas com sinal circunflexo.

Exemplos de parâmetros (populacionais) e estatísticas (amostrais):

Descrição	Parâmetro	Estatística
Número de elementos	N	n
Média	μ	$\bar{x}$
Variância	σ^2	s^2
Desvio padrão	σ	s
Proporção	p	$\hat{p}$
Coeficiente de correlação	$\rho(X,Y)$	$r(X,Y)$

1.4. DISTRIBUIÇÕES AMOSTRAIS

A inferência estatística está interessada em tomar decisões sobre algum parâmetro da população com base na informação contida em uma amostra aleatória desta população.

Nas amostras são calculadas as estatísticas, usando expressões matemáticas chamadas de estimadores.

Como toda estatística ($\bar{x}$, s^2 etc.) é função dos valores de variáveis aleatórias, é necessário estudar a distribuição de probabilidades dessas estatísticas.

A distribuição de probabilidades de uma estatística é chamada distribuição amostral daquela estatística

Vamos conhecer a distribuição amostral da média e da proporção amostrais.

1.4.1. A DISTRIBUIÇÃO AMOSTRAL DA MÉDIA

Teorema 1. Seja X uma população com média μ e variância σ^2 e seja $\{x_1, x_2, \ldots, x_n\}$ uma $a.c.s.$ de tamanho n, retirada desta população. Então, a esperança matemática e a variância da média amostral são obtidas como:

$$E(\bar{x}) = \mu \qquad\qquad var(\bar{x}) = \sigma^2/n$$

O erro padrão da média é definido como:

$$ep(\bar{x}) = \sqrt{var(\bar{x})} = \sigma/\sqrt{n}$$

Dúvida: Qual a distribuição probabilística que podemos associar à média amostral ($\bar{x}$)?

1) Se a amostra aleatória de tamanho n é retirada de uma distribuição normal, $X \sim N(\mu, \sigma^2)$, a estatística $\bar{x}$ também será uma distribuição normal de média, μ, mas com variância σ^2/n, isto é:

$$\text{Se } X \sim N(\mu, \sigma^2) \quad \Rightarrow \quad \bar{x} \sim N(\mu, \sigma^2/n\,)$$

2) **Teorema do Limite Central**

Se a $a.c.s.$ de tamanho n é retirada de **qualquer** população com média μ e variância σ^2, a distribuição amostral da média $(\bar{x})$ aproxima-se de uma **distribuição normal** com média μ e variância σ^2/n, quando o tamanho da amostra, n, tender para infinito.

$$\text{Se } X \sim ?\,(\mu, \sigma^2) \quad \Rightarrow \quad \bar{x} \sim N(\mu, \sigma^2/n) \text{ quando } n \to \infty$$

- A rapidez dessa convergência depende da distribuição da população da qual a amostra é retirada: se a distribuição for simétrica e unimodal a convergência é bastante rápida.
- De um modo geral, admitimos que para amostras com mais de 30 elementos, a aproximação pela distribuição normal já pode ser considerada boa.

Corolário 1. Se $\{x_1, x_2, \ldots, x_n\}$ é uma $a.c.s.$ de tamanho n de uma população X que tem média μ e variância σ^2, então a variável:

$$\mathrm{Z} = \frac{\bar{x}-\mu}{\sqrt{\sigma^2/n}} = \frac{(\bar{x}-\mu)\sqrt{n}}{\sigma} \sim N(0,1) \text{ quando } n \to \infty$$

Nota: Este resultado será usado no cálculo de probabilidades sobre a média amostral.

1.4.2. DISTRIBUIÇÃO AMOSTRAL DA PROPORÇÃO

Seja $\hat{p}$ a proporção (frequência relativa) de indivíduos que têm uma característica de interesse na amostra. Podemos provar que:

$$E(\hat{p}) = p \qquad var(\hat{p}) = \frac{p(1-p)}{n}$$

Para amostras grandes ($n \to \infty$) utilizamos o Teorema do Limite Central para garantir que:

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

A seguir vamos conhecer mais algumas distribuições de probabilidade que serão úteis quando fizermos inferências sobre a média, a variância e a proporção.

2. OUTRAS DISTRIBUIÇÕES PROBABILÍSTICAS IMPORTANTES

2.1. DISTRIBUIÇÃO QUIQUADRADO (χ^2)

É utilizada em tabelas de contingência, na construção de intervalos de confiança e testes de hipóteses sobre a variância de uma população normal.

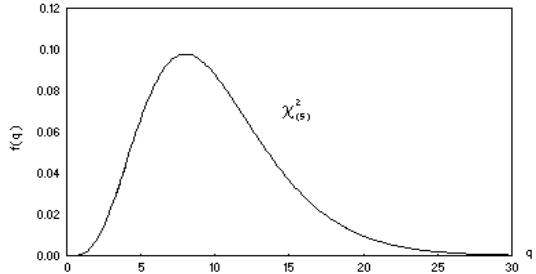


Figura 1. Distribuição $\chi^2_{(5)}$

A distribuição χ^2 é assimétrica à direita e definida somente para valores positivos.

Para o cálculo de probabilidades usamos a Tábua II.

2.2. DISTRIBUIÇÃO t DE STUDENT

É uma das mais importantes distribuições probabilísticas usadas nas inferências sobre médias de populações normais.

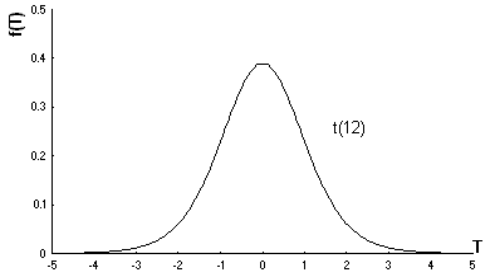


Figura 2. Distribuição $t(12)$

O gráfico da distribuição t-Student (Figura 2) é parecido com o gráfico da distribuição $N(0;1)$ mas tem as caudas mais pesadas

A Tábua III fornece valores críticos t_c tais que $P(T > t_c) = p$, para alguns valores de p e de ν (número de graus de liberdade)

2.3. DISTRIBUIÇÃO F DE SNEDECOR

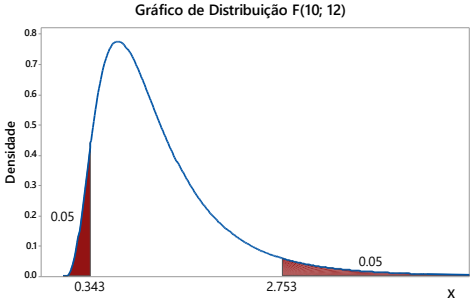
É bastante usada nas inferências sobre as variâncias de duas populações com distribuição normal e nos testes associados à Análise de Variância (ANOVA).

- A distribuição F tem dois parâmetros: os números de graus de liberdade do numerador (v_1) e do denominador (v_2).

- Os valores críticos f_c tais que $P(F > f_c) = 0{,}05$, para alguns valores de v_1 e v_2 podem ser encontrados na Tábua IV.

- Os valores críticos (f_c), tais que $P(F < f_c) = 0{,}05$ não são obtidos diretamente na Tábua IV. Para tanto utiliza-se a relação:

$$F(v_1;\ v_2) = 1/\ F(v_2;\ v_1)$$



A Figura 3 apresenta o gráfico da distribuição $F(10;12)$ e os valores críticos tais que:

$$P(F < 0{,}343) = 0{,}05$$

$$P(F > 2{,}753) = 0{,}05$$

Note que a distribuição F é bastante assimétrica à direita (assimetria positiva)

3. ESTIMAÇÃO

Na produção de generalizações sobre a população com base em resultados obtidos de uma amostra, estão envolvidos a estimação de parâmetros e os testes de hipóteses.

A estimação de parâmetros pode ser feita pontualmente e por intervalo. Para obtenção de bons estimadores pontuais existem alguns métodos, como:

- Método dos Mínimos Quadrados (MMQ)
- Método da Máxima Verossimilhança (MMV)
- Método dos Momentos (MM)

Exemplos de estimadores:

- O estimador de mínimos quadrados da média populacional (μ) é a média amostral:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Da Álgebra Linear sabemos que os estimadores de mínimos quadrados dos parâmetros do modelo de regressão linear simples, $y_i = a + bx_i + \varepsilon_i$, são:

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \hat{a} = \bar{y} - \hat{b}\bar{x}$$

3.1. CARACTERÍSTICAS DE UM BOM ESTIMADOR PONTUAL

Um bom estimador deve ser justo, consistente e eficiente.

i) T é um estimador justo, não viesado ou não tendencioso do parâmetro θ se $E(\mathrm{T}) = \theta$.

ii) T é um estimador consistente do parâmetro θ se ele for justo e se $\lim_{n\to\infty} var(T) = 0$.

iii) Se T_1 e T_2 são estimadores justos do parâmetro θ e se $var(T_1) < var(T_2)$, então T_1 é dito ser mais eficiente que o estimador T_2.

Exemplo 3.2. Queremos comprar um rifle e temos três opções, r_1, r_2 e r_3. Para avaliar a qualidade de cada arma vamos atirar 30 vezes num alvo e analisar a distribuição das marcas próximas da mosca (ponto central).

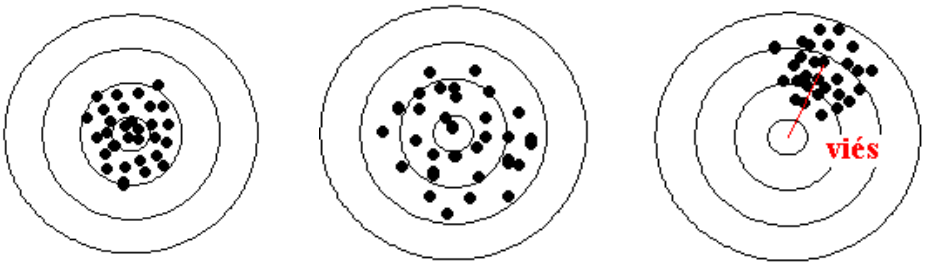


Figura 5. Qualidades de um bom estimador.

Observamos que:

i) As armas r_1 e r_2 buscam atingir a mosca. A mira de r_3 não está bem calibrada, porém, se conseguirmos calibrar r_3 podemos ter um bom rifle.

ii) O rifle r_1 é melhor que r_2 porque os seus tiros formam uma nuvem de pontos com menor dispersão ao redor da mosca.

Conclusão: O rifle r_1 reúne as melhores características.

Analogia: A mosca do alvo é o valor do parâmetro que pretendemos estimar. Os rifles são os estimadores que usamos para tentar acertar o verdadeiro valor do parâmetro. Os tiros dos rifles correspondem às estimativas geradas pelos estimadores em amostras diferentes.

Então podemos dizer que:

i) Os estimadores r_1 e r_2 são estimadores justos ou não viesados.

ii) O estimador r_3 é um estimador viesado porque, em média, não acerta o valor do parâmetro.

iii) Dentre os estimadores justos, o estimador r_1 é mais eficiente que r_2, pois tem menor variância.

Exemplos de bons estimadores:

Média amostral: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Variância amostral: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$