Metodologia de Diagnóstico e Elaboração de Relatório – FASHT

Inferência Estatística

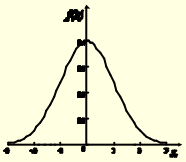
Profa. Cesaltina Pires
cpires@uevora.pt

Plano da Apresentação

- Duas distribuições importantes
  - Normal
  - T- Student
- Estimação
  - Distribuição por amostragem
  - Propriedades desejáveis dos estimadores
  - Estimação por intervalos
- Teste de hipóteses

Distribuição normal

- Muito importante em estatística porque
  - Vários fenómenos podem ser aproximadamente descritos por distribuição normal
  - É distribuição base em inferência estatística
- Características da distribuição normal
  - Tem forma de sino e é simétrica
  - Moda, mediana e média coincidem
  - Muito provável obter valores pouco afastados da média
  - Aproximadamente 2/3 das observações distam da média menos do que 1 desvio -padrão
  - Aproximadamente 95% das observações distam da média menos de 2 desvios padrões
  - Só depende da média e do desvio padrão



$X \sim N(\mu, s^2)$

média variância

Distribuição normal estandardizada

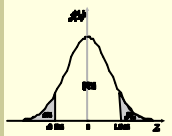
- É uma normal com média = 0 e desvio padrão =1
- Existem tabelas que nos indicam a probabilidade de uma variável aleatória com distribuição N(0,1) tomar um valor inferior ou igual a um dado valor.
  - Qual é a probabilidade de tomar um valor inferior ou igual a 0?
  - Qual é a probabilidade de tomar um valor inferior ou igual a 1.6?
- Se uma variável seguir a distribuição $N(\mu, s^2)$ podemos estandardizá-la

$$Z = (X - \mu)/s$$

Z tem distribuição N(0,1), podemos usar tabela.

- Exemplo notas têm distribuição N(13,2)
  - Qual é percentagem de alunos que tem mais de 17?
  - Qual a percentagem de alunos com nota entre 11 e 15?

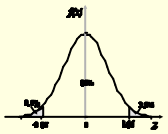
Distribuição normal estandardizada



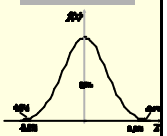
• A probabilidade de Z ser superior a 1.645 é 5%

• Por simetria, a probabilidade de Z ser inferior a -1.645 é 5%

• A probabilidade de Z estar entre -1.645 e +1.645 é 90%



• A probabilidade de Z estar entre -1.96 e 1.96 é 95%.



• A probabilidade de Z estar entre -2.575 e 2.575 é 99%.

Estes valores (1.645, 1.96, 2.575) são valores críticos.

Valor crítico z_a é o valor da variável normal tal que

a probabilidade de estar acima desse valor é a.

Distribuição t

- Tal como a normal N(0,1), a distribuição t é simétrica em torno do zero e tem forma de sino.
- Tem mais área nas abas e menos área no centro que a normal.
- A distribuição t depende apenas dos graus de liberdade
- À medida que o nº de graus de liberdade aumenta a t fica mais próxima da N(0,1).

Estimadores

Parâmetros da população (o que queremos realmente saber)	Estimadores (calculados usando a amostra)
Média na população – μ	Média na amostra – $\overline{X}$
Proporção na população – p	Proporção na amostra - $\hat{P}$
Diferença de Médias – $\mu_1 - \mu_2$	Diferença de médias na amostra
Variância na população – s^2	Variância na amostra – S^2

Distribuição por amostragem

- Os estimadores são variáveis aleatórias. Porquê? Porque o valor do estimador depende da amostra.
- Exemplo: quero estimar a idade média dos alunos de métodos quantitativos usando amostra de 5 alunos. A idade média na amostra depende dos 5 alunos seleccionados.
- Se os estimadores são variáveis aleatórias é muito importante saber qual é a sua função de distribuição. Será que seguem uma normal? Será que é uma t?...
- Toda a inferência estatística é baseada na distribuição do estimador.
- Diferença entre estimador (regra de cálculo) e estimativa (valor do estimador para uma amostra em concreto).

Distribuição da média na amostra - s^2 é conhecido

Se a população tiver uma distribuição normal $N(\mu,s^2)$ então o estimador média na amostra tem distribuição normal:

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}) \rightarrow \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

- A média das médias na amostra é igual á média na população.
- Quanto maior for a dimensão da amostra (maior n), menor é a variabilidade do estimador média na amostra. Ou seja, quanto maior n Maior é a precisão com que a média na amostra estima a média na população.
- Resultado verifica-se mesmo para pequenas amostras.

Distribuição da média na amostra - s^2 é conhecido

Seja qual for a distribuição da população (pode ser normal ou não) se a dimensão da amostra for elevada, então o estimador média na amostra tem distribuição aproximadamente normal:

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

- Este resultado chama-se Teorema do Limite Central
- Resultado só é verdadeiro para amostras grandes
- Este resultado admite que s^2 é conhecido. Se não for conhecido, pode calcular-se a variância na amostra e usá-la como estimador de s^2. Em grandes amostras não há problema.

Distribuição da média na amostra - s^2 desconhecido

Quando a variância na população é desconhecida, temos de a estimar com base na variância na amostra. Qual é a distribuição do estimador média na amostra quando se usa S^2?

- Se a amostra for grande a distribuição da média na amostra é aproximadamente normal.
- Para amostras pequenas, só se a população for normal é que sabemos a distribuição. O estimador média na amostra segue uma distribuição t com n-1 graus de liberdade.
- Porquê n-1 graus de liberdade?

Distribuição da proporção na amostra

Para amostras grandes, a proporção na amostra segue uma distribuição aproximadamente normal. Estandardizando:

Proporção na amostra

$$\frac{\hat{P} - p}{\sqrt{\hat{P}(1-\hat{P})/n}}$$

segue uma N(0,1)

Variância na amostra

- A média das proporções na amostra é igual á proporção na população
- A variância na população no caso da proporção é p(1-p)

Propriedades desejáveis de estimadores

- Não enviesamento – a média do estimador é igual ao parâmetro da população que queremos estimar. Ou seja, em média o estimador está correcto (às vezes sobrestima o parâmetro, outras vezes subestima, mas em média está correcto)
- Consistência – Á medida que a dimensão da amostra se torna mais elevada, o estimador dá-nos uma ideia cada vez mais precisa de qual é o verdadeiro valor do parâmetro
- Eficiência – No conjunto de estimador não enviesados o mais eficiente é aquele que tem menor variabilidade (que é mais preciso)

Estimação por intervalos

Exemplo: queremos construir um intervalo que contenha a média na população com probabilidade 95%. Sabe-se que a população é normal e tem variância = 16 e que vai ser recolhida amostra com 100 observações.

- Média na amostra tem média μ e variância 16/100
- A probabilidade de Z estar entre -1.96 e 1.96 é 95%

$$P\left(-1.96 \le \frac{\overline{X}-\mu}{4/\sqrt{100}} \le 1.96\right) = 0.95 \quad \Leftrightarrow \quad P\left(\overline{X} - 1.96 \times \frac{4}{\sqrt{100}} \le \mu \le \overline{X} - 1.96 \times \frac{4}{\sqrt{100}}\right)$$

Valor crítico $Z_{a/2}$

Nível de confiança (1-a)

Estimação por intervalos

Um estimador por intervalo é uma regra para determinar um intervalo que com certa probabilidade contém o parâmetro da população em que estamos interessados.

Exemplo: se a variância da população for conhecida e a população for normal ou se a amostra for grande, o intervalo com o nível de confiança de 100(1-a)% para μ é:

$$\left[\overline{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right),\ \overline{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

- O intervalo é centrado na média - soma-se e subtrai-se $z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$
- Quanto maior o nível de confiança desejado (maior (1-a)) maior é o valor crítico de z. Logo, maior é amplitude do intervalo (menor precisão).
- Quanto maior a dimensão da amostra, menor é a amplitude do inte rvalo (maior precisão).

Exemplo de um intervalo de confiança para μ

Precisamos de estimar o consumo médio de gasolina para um novo modelo de automóvel, com um nível de confiança de 99%. Obteve-se informação sobre 100 automóveis deste modelo. O consumo médio para nestes 100 automóveis foi 6.12 e o desvio padrão na amostra foi 0.4.

Como a amostra é grande podemos usar a normal. Valor crítico é 2.575.

[6.12 - 2.575×0.4/10 , 6.12 + 2.575×0.4/10]

⇓

[6.017 , 6.223]

O intervalo seria maior ou menor se nível de confiança fosse 95%?

Outro exemplo de um intervalo de confiança para μ

Precisamos de estimar o consumo médio de gasolina para um novo modelo de automóvel, com um nível de confiança de 99%. Sabe-se que o consumo de gasolina segue uma distribuição normal. Obteve-se informação sobre 6 automóveis deste modelo. O consumo médio para estes 6 automóveis foi 6.07 e o desvio padrão na amostra foi 0.18.

Como a amostra é pequena e a população é normal a distribuição a usar é a t com n-1 = 5 graus de liberdade. O valor crítico $t_{0.005}$ com 5 graus de liberdade é 4.032.

[6.07 – 4.032×0.18/√6 , 6.07 + 4.032×0.18/√6]

Teste de hipóteses

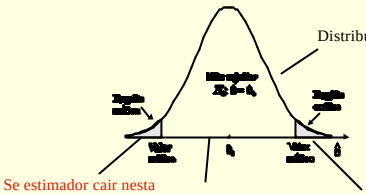
- Como é que a informação na amostra pode ser usada para testar «conjecturas» ou «hipóteses» sobre os parâmetros da população.
- Exemplo: quero testar se a média do salário é a mesma para homens e mulheres (assumindo igual qualificação e experiência)
- Exemplo: quero testar se a proporção de eleitores que vai votar num dado candidato é superior a 50%.
- Hipótese nula – é a hipótese que temos sobre a população e que continuaremos a admitir a não ser que a evidência na amostra sugira que essa hipótese é falsa.
  - No caso do salário a hipótese nula é que não há descriminação.
- Hipótese alternativa

Teste de hipóteses

- Suponhamos que queremos testar H_0: $\mu = 10$.
- Recolhemos uma amostra e calculamos a média na amostra.
- A ideia é: se a média na amostra é próxima de 10, não rejeitamos a hipótese nula. Se a média na amostra é muito afastada de 10 rejeitamos a hipótese nula (a evidência na nossa amostra sugere que H_0 é falsa).
- O que é «afastado» de 10? O afastamento deve ser medido em termos de desvios padrões da variável (2 desvios padrões é afastado, 3 é mesmo muito afastado)
- Definir região crítica – região em que se rejeita a hipótese nula.

Teste de hipóteses

Teste da hipótese nula H_0 : ? = $?_0$



Se estimador cair nesta zona rejeitamos H_0

Se estimador cair nesta zona não rejeitamos H_0

Se estimador cair nesta zona rejeitamos H_0

Exemplo – Peso das embalagens de detergente

Um produtor de detergentes argumenta que o peso médio de cada embalagem do seu detergente é 500 gramas. Sabe-se que o peso segue uma distribuição normal com desvio padrão igual a 12.5.
Numa amostra de 20 caixas o peso médio foi 485 gramas. Será que o produtor tem razão?

$H_0 : \mu = 500$
$H_1 : \mu \neq 500$

Intuitivamente, devemos rejeitar a hipótese nula se a média na amostra for muito diferente de 500.

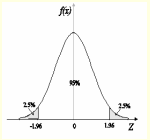
Teste de hipóteses não são 100% correctos

Num teste de hipóteses há alguma probabilidade de tirarmos conclusões erradas (porquê?)

- Podemos rejeitar hipótese nula, quando na realidade ela é verdadeira (a informação na nossa amostra sugere que a hipótese nula é falsa, mas H_0 é mesmo verdadeira).
  A probabilidade desse erro chama-se nível de significância do teste – a.
- Podemos não rejeitar a hipótese nula quando na realidade ela é falsa (a informação na amostra sugere que H_0 é verdadeira, mas na realidade não é).
  A probabilidade deste erro depende da hipótese alternativa – designa-se por ß.

Passos no teste de hipóteses

1. **Identificar a distribuição seguida pelo estimador**
   Neste exemplo, como a população é normal e a variância é conhecida, a média na amostra tem distribuição normal.
2. **Determinar região crítica.**
   A região crítica depende do nível de significância do teste.
   Para a = 0.05, devemos rejeitar a hipótese nula se
   z < -1.96 ou se z > 1.96



Passos no teste de hipóteses

3. **Calcular o valor da estatística z com base nos resultados na amostra e verificar se o seu valor caí ou não na região de rejeição.**
   Neste exemplo o valor de z é:

$$z = \frac{485 - 500}{12.5/\sqrt{20}} = -5.37$$

Este valor caí na Região crítica

Conclusão: como -5.37 < -1.96, rejeitamos a hipótese nula de que o peso médio das embalagens é 500 gramas. A evidência sugere que O peso médio é inferior a 500.

Exemplo – Poluição baixou?

Uma empresa produtora de papel tomou medidas para reduzir descarga de poluentes na água. Antes dessas medidas a média era 400ppm. Para testar se a média baixou a empresa recolheu amostras de água em 25 dias consecutivos. Para essas observações a média foi 208.8 e o desvio padrão foi 115.5. Como testar se a poluição baixou?

$H_0 : \mu = 400$
$H_1: \mu < 400$ (teste unilateral)

Intuitivamente, devemos rejeitar a hipótese nula se a média na amostra for muito inferior a 400.

Exemplo – Poluição baixou?

1. **Identificar a distribuição do estimador**
   Admitindo que a população é normal, a distribuição apropriada é a t-student com n-1 graus de liberdade (porque a variância na população é desconhecida e amostra é pequena).

2. **Determinar região crítica**
   Como o teste é unilateral só devemos rejeitar para valores muito baixos de t. Para a = 0.01, o valor crítico de uma t com n-1 = 24 graus de liberdade é

Exemplo de teste de hipóteses

3. **Calcular estatística t usando resultados na amostra e verificar se caí ou não na região crítica.**

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} = \frac{208.8 - 400}{115.15/5} = -8.3$$

Conclusão: como -8.3 < -2., rejeitamos a hipótese nula de que a poluição não baixou. A evidência sugere que a empresa conseguiu reduzir a poluição na água.