ESTATÍSTICA INFERENCIAL

Devemos ter o cuidado de não confundir os dados com as abstracções que utilizamos para os analisar.

William James (1842-1910)

Estatística inferencial

O objectivo da Estatística é caracterizar e eventualmente definir regras de decisão sobre uma população conhecendo apenas parte dela.

O objectivo usual é inferir sobre a forma ou os parâmetros da distribuição F_X.

Se estivermos interessados na **forma** podemos começar por comparar o histograma (ou gráfico de frequências) com os gráficos de $f(x)$ das distribuições usuais.

Seguidamente podemos construir gráficos de quantis (QQ-plot) ou de probabilidades (PP-plot). Estes gráficos também são designados papel de probabilidades.

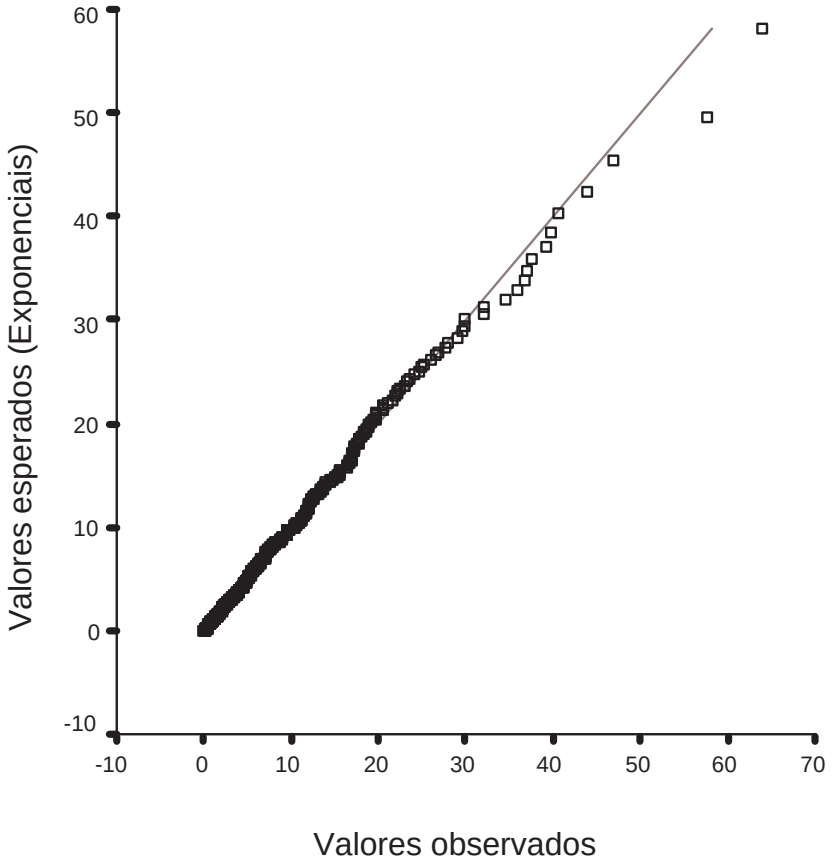
QQ-plots

Um **QQ-plot** é um gráfico de dispersão que confronta os quantis da amostra com os quantis de uma distribuição específica (usual). Se a amostra tiver sido retirada de uma população com aquela distribuição o gráfico deve assemelhar-se a um conjunto de pontos mais ou menos sobre uma recta. Caso contrário deverão surgir zonas de não-linearidade no gráfico.

No SPSS os QQ-plots estão disponíveis no menu `Graphs / QQ` para várias distribuições usuais. Em particular o QQ-plot da distribuição Normal também está disponível no menu `Analyze / Descriptive Statistics / Explore`, seleccionando o botão `Plots` e colcando um $\surd$ na opção `Normality tests with plots`.

Existem outros procedimentos para inferir sobre a forma de uma distribuição (a ver mais a diante).

Exemplo de um QQ-plot



Estimação pontual

Uma vez decidida a forma podemos estar interessados em inferir sobre os parâmetros.

Estimativa (pontual) de um parâmetro desconhecido - valor obtido a partir da amostra (através de uma estatística) que se destina a fornecer valores aproximados do parâmetro.

Exemplo: se uma amostra tiver média $\bar{x} = 5.1$, então esse valor é uma estimativa da média da população, μ.

Estimador - estatística que fornece estimativas pontuais.

Exemplo: a média de uma amostra, enquanto variável aleatória, $\bar{X}$, é uma estimador da média da população, μ.

Habitualmente representa-se um estimador (ou uma estimativa) de um parâmetro colocando um acento circunflexo sobre a letra que o representa. $(\hat{\mu}, \hat{\sigma}, \hat{\theta})$

Exemplo:

$\hat{\mu} = \bar{X}$ representa um estimador da média da população μ.

$\hat{\mu} = \bar{x} = 5.1$ representa uma estimativa da média da população μ.

Um estimador é uma variável aleatória e como tal tem uma distribuição que o caracteriza - **distribuição de amostragem**.

Que propriedades deve ter um bom estimador?

- Um bom estimador deve ser tal que, ao tomarmos uma grande quantidade de amostras e calcularmos a médias das respectivas estimativas, esta deve aproximar-se do verdadeiro valor do parâmetro. Neste caso o estimador diz-se **centrado** ou **não enviesado**. Caso contrário diz-se **enviesado**.

- Um bom estimador deve ser tal que, ao aumentarmos a dimensão da amostra, as estimativas devem aproximar-se do verdadeiro valor do parâmetro. Neste caso o estimador diz-se **consistente**.

- Um bom estimador deve fornecer estimativas que não se afastem muito do verdadeiro valor do parâmetro (variância reduzida).

INTERVALOS DE CONFIANÇA

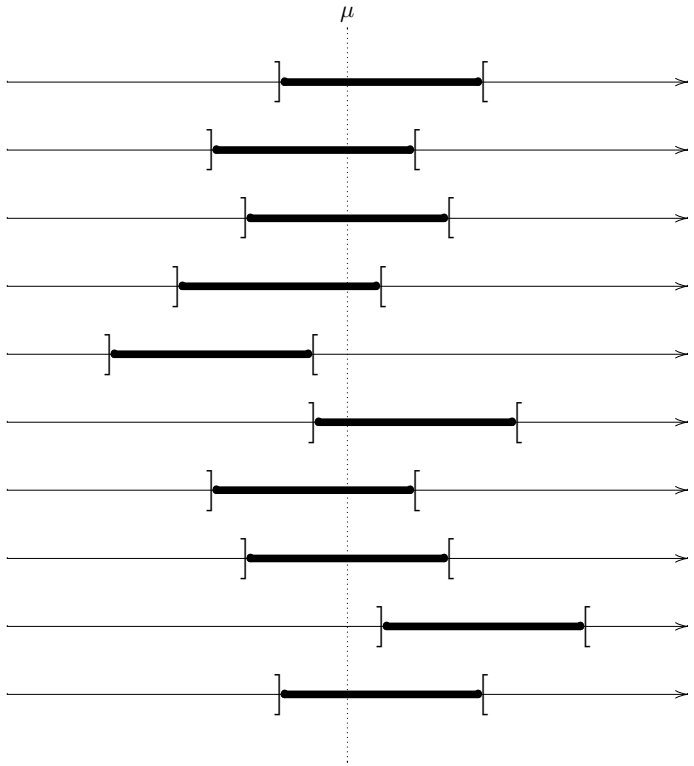
Uma estimativa pontual de um parâmetro não contém informação sobre a precisão do valor obtido. Uma forma mais completa de abordar a questão consiste em construir estimativas na forma de intervalos e conhecer a probabilidade de o intervalo conter o verdadeiro valor do parâmetro.

Um intervalo de confiança para um parâmetro θ, a um grau de confiança $1-\alpha$, é um intervalo aleatório (L_{inf}, L_{sup}) tal que

$$P(L_{inf} < \theta < L_{sup}) = 1-\alpha, \quad \alpha \in (0,1).$$

α deve ser um valor muito reduzido por forma a termos confianças elevadas. Valores usuais para o grau de confiança são 95%, 99% e 90%.

Para cada amostra que se observa obtém-se (em geral) um intervalo de confiança diferente para o mesmo parâmetro. Quando dizemos que um intervalo tem confiança $1-\alpha$ estamos a dizer que se observarmos muitas amostras distintas, os intervalos que se obtêm contêm o verdadeiro valor do parâmetro $(1-\alpha)*100\%$ das vezes.



Intervalo de confiança para a média μ de uma população Normal com variância conhecida σ^2

Pressupostos exigidos:

1. As observações devem ser independentes e retiradas da mesma população (amostra aleatória);
2. A população deve ter distribuição Normal;
3. A variância da população, σ^2, deve ser conhecida a priori.

Um intervalo de confiança para a média μ de uma população Normal com variância conhecida σ^2, a um grau de confiança $1-\alpha$, é dado por

$$(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}),$$

onde $z_{1-\frac{\alpha}{2}}$ representa o quantil de ordem $1-\frac{\alpha}{2}$ da distribuição Normal standard.

Propriedades deste intervalo de confiança

Quanto maior o grau de confiança maior a largura do intervalo.

Quanto maior a variância, maior a largura do intervalo,

Quanto maior a amostra, menor a largura do intervalo.

Intervalo de confiança para a média μ de uma população Normal com variância desconhecida

O intervalo de confiança para μ quando a variância é conhecida foi derivado do facto de

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \frown N(0,1).$$

Se o valor de σ é desconhecido tentamos substitui-lo por uma estimativa S. Neste caso tem-se

$$T = \frac{\bar{X} - \mu}{S_c/\sqrt{n}} \frown t_{n-1}.$$

Os intervalos que se obtêm agora têm maior largura do que se o valor de σ^2 fosse considerado conhecido, reflectindo a incerteza acrescida pelo desconhecimento deste parâmetro.

Pressupostos exigidos:

1. As observações devem ser independentes e retiradas da mesma população (amostra aleatória);
2. A população deve ter distribuição Normal com os dois parâmetros desconhecidos.

Um intervalo de confiança para a média μ de uma população Normal com variância desconhecida, a um grau de confiança $1-\alpha$, é dado por

$$(\bar{X} - t_{1-\frac{\alpha}{2},n-1}\,\frac{S_c}{\sqrt{n}}, \bar{X} + t_{1-\frac{\alpha}{2},n-1}\,\frac{S_c}{\sqrt{n}}),$$

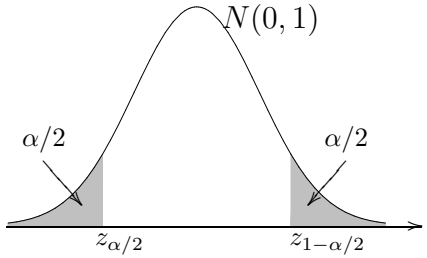
onde $t_{1-\frac{\alpha}{2},n-1}$ representa o quantil de ordem $1-\frac{\alpha}{2}$ da distribuição t de Student com $n-1$ graus de liberdade.

Nota: estes intervalos estão disponíveis no SPSS.

$X_1, X_2, \ldots, X_n$ é uma a.a. com distribuição Normal(μ, σ^2), σ conhecido.

$\bar{X}$ estima μ

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \frown N(0,1).$$



$$P(z_{\alpha/2} < Z < z_{1-\alpha/2}) = 1-\alpha \Leftrightarrow$$

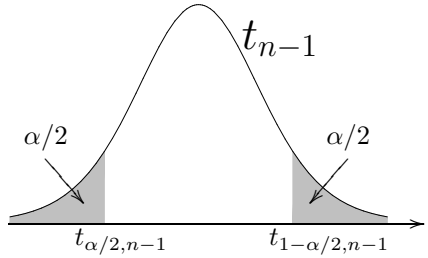
$$P(-z_{1-\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}) = 1-\alpha \Leftrightarrow$$

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

$X_1, X_2, \ldots, X_n$ é uma a.a. com distribuição Normal(μ, σ^2), σ desconhecido.

$\bar{X}$ estima μ e S_c estima σ

$$T = \frac{\bar{X}-\mu}{S_c/\sqrt{n}} \frown t_{n-1}.$$



$$P(t_{\frac{\alpha}{2},n-1} < T < t_{1-\frac{\alpha}{2},n-1}) = 1-\alpha \Leftrightarrow$$

$$P(-t_{1-\frac{\alpha}{2},n-1} < \frac{\bar{X}-\mu}{S_c/\sqrt{n}} < t_{1-\frac{\alpha}{2},n-1}) = 1-\alpha$$

$$\Leftrightarrow P(\bar{X} - t_{1-\frac{\alpha}{2},n-1}\frac{S_c}{\sqrt{n}} < \mu <$$

$$< \bar{X} + t_{1-\frac{\alpha}{2},n-1}\frac{S_c}{\sqrt{n}}) = 1-\alpha.$$

Intervalo de confiança para a diferença de médias $\mu_X - \mu_Y$ de duas populações Normais — amostras independentes.

Pressupostos exigidos:

1. Temos duas amostras $X_1, \ldots, X_n$, $Y_1, \ldots, Y_m$ independentes
2. Cada amostra deve ser constituída por observações independentes e retiradas da mesma população (amostras aleatórias)
3. As duas populações devem ter distribuição Normal com as variâncias **desconhecidas mas iguais**.

Um intervalo de confiança para a diferença de médias $\mu_X - \mu_Y$ de duas populações Normais com variâncias desconhecidas mas iguais, obtido a partir de duas amostras independentes, a um grau de confiança $1-\alpha$, é dado por

$$\left(\bar{X} - \bar{Y} - t_{1-\frac{\alpha}{2},n+m-2} \sqrt{\frac{1}{n}+\frac{1}{m}} \sqrt{\frac{(n-1)S^2_{X_c}+(m-1)S^2_{Y_c}}{(n+m-2)}}, \right.$$

$$\left. , \bar{X} - \bar{Y} + t_{1-\frac{\alpha}{2},n+m-2} \sqrt{\frac{1}{n}+\frac{1}{m}} \sqrt{\frac{(n-1)S^2_{X_c}+(m-1)S^2_{Y_c}}{(n+m-2)}} \right) .$$

Nota: estes intervalos estão disponíveis no SPSS.

Intervalo de confiança para a diferença de médias $\mu_X - \mu_Y$ de duas populações Normais — amostras emparelhadas.

Quando queremos comparar a localização de duas populações com base em amostras dependentes não sabemos especificar (em geral) qual a distribuição da diferença das médias amostrais.

Iremos considerar apenas a situação em que as amostras são dependentes na medida em que têm a mesma dimensão e cada observação X_i depende da observação Y_i mas os pares (X_i, Y_i) e (X_j, Y_j) são independentes $(i \neq j)$. Este tipo de amostras chamam-se amostras emparelhadas.

O procedimento a seguir é o seguinte:

Dadas duas amostras aleatórias emparelhadas $(X_1, \ldots, X_n)$, $(Y_1, \ldots, Y_n)$ provenientes de populaccões Normais consideram-se as diferenças

$$D_i = X_i - Y_i \frown N(\mu_D, \sigma_D^2),$$

onde μ_D é igual à diferença das médias das populações e σ_D representa o desvio padrão das diferenças D_i.

A variável

$$T = \frac{\bar{D} - \mu_D}{S_{D_c}/\sqrt{n}} \frown t_{n-1}$$

onde S_{D_c} representa o desvio padrão amostral corrigido das diferenças.

Em seguida determina-se um intervalo de confiança para a média da população das diferenças como se fez anteriormente para uma só amostra.

Pressupostos exigidos:

1. Temos duas amostras $X_1, \ldots, X_n$, $Y_1, \ldots, Y_n$ emparelhadas, i.e., formando pares (X_i, Y_i).
2. Cada amostra deve ser constituída por observações independentes e retiradas da mesma população (amostras aleatórias)
3. As duas populações devem ter distribuição Normal

Um intervalo de confiança para a diferença de médias $\mu_X - \mu_Y = \mu_D$ de duas populações Normais, obtido a partir de duas amostras emparelhadas, a um grau de confiança $1 - \alpha$, é dado por

$$\left(\bar{D} - t_{1-\frac{\alpha}{2},n-1} \frac{S_{D_c}}{\sqrt{n}}, \bar{D} + t_{1-\frac{\alpha}{2},n-1} \frac{S_{D_c}}{\sqrt{n}} \right).$$

Nota: estes intervalos estão disponíveis no SPSS.

Intervalo de confiança para a média μ de uma população genérica com variância conhecida σ^2

Duma forma geral, conhecendo a variância duma distribuição e considerando válidas as condições do Teorema do Limite Central (n elevado) tem-se que $\bar{X} \frown N(\mu, \sigma^2/n)$, pelo que podemos obter um intervalo de confiança para μ.

Pressupostos exigidos:

1. As observações devem ser independentes e retiradas da mesma população (amostra aleatória);
2. A variância da população é conhecida.
3. A amostra tem dimensão elevada.

Um intervalo de confiança aproximado para a média μ de uma população genérica com variância conhecida, σ^2, a um grau de confiança $1-\alpha$, é dado por

$$(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}).$$

Esta aproximação será tanto melhor quanto maior a dimensão da amostra.

Quando não se conhece a variância σ^2 é usual substiuir σ por S_c e utilizar o mesmo intervalo. Chama-se a atenção que este procedimento só deve ser utilizado em grandes amostras.

Pressupostos exigidos:

1. As observações devem ser independentes e retiradas da mesma população (amostra aleatória);
2. A amostra tem dimensão elevada.

Um intervalo de confiança aproximado para a média, μ, de uma população genérica com variância desconhecida, σ^2, a um grau de confiança $1-\alpha$, é dado por

$$(\bar{X} - z_{1-\alpha/2}\frac{S_c}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{S_c}{\sqrt{n}}).$$

Esta aproximação será tanto melhor quanto maior a dimensão da amostra.

Intervalo de confiança para a variância σ^2 de uma população Normal

Pressupostos exigidos:

1. As observações devem ser independentes e retiradas da mesma população (amostra aleatória);
2. A população deve ter distribuição Normal.

Um intervalo de confiança para a variância σ^2 de uma população Normal, a um grau de confiança $1-\alpha$, é dado por

$$\left(\frac{(n-1)S_c^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}, \frac{(n-1)S_c^2}{\chi^2_{\frac{\alpha}{2},n-1}} \right),$$

onde $\chi^2_{1-\frac{\alpha}{2},n-1}$ representa o quantil de ordem $1-\frac{\alpha}{2}$ da distribuição χ^2_{n-1}.

Intervalo de confiança para o desvio padrão σ de uma população Normal

Pressupostos exigidos:

1. As observações devem ser independentes e retiradas da mesma população (amostra aleatória);
2. A população deve ter distribuição Normal.

Um intervalo de confiança para o desvio padrão σ de uma população Normal, a um grau de confiança $1-\alpha$, é dado por

$$\left(\sqrt{\frac{(n-1)S_c^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}}, \sqrt{\frac{(n-1)S_c^2}{\chi^2_{\frac{\alpha}{2},n-1}}} \right),$$

onde $\chi^2_{1-\frac{\alpha}{2},n-1}$ representa o quantil de ordem $1-\frac{\alpha}{2}$ da distribuição χ^2_{n-1}.

Intervalo de confiança para a razão de variâncias $\frac{\sigma_X^2}{\sigma_Y^2}$ de duas populações Normais

Pressupostos exigidos:

1. Temos duas amostras $X_1, \ldots, X_n$, $Y_1, \ldots, Y_m$ independentes
2. Cada amostra deve ser constituída por observações independentes e retiradas da mesma população (amostras aleatórias)
3. As duas populações devem ter distribuição Normal.

Um intervalo de confiança para a razão de variâncias $\frac{\sigma_X^2}{\sigma_Y^2}$ de duas populações Normais, a um grau de confiança $1-\alpha$, é dado por

$$\left(\frac{S_{X_c}^2}{S_{Y_c}^2} f_{\frac{\alpha}{2},m-1,n-1}, \frac{S_{X_c}^2}{S_{Y_c}^2} f_{1-\frac{\alpha}{2},m-1,n-1} \right),$$

onde $f_{\frac{\alpha}{2},m-1,n-1}$ representa o quantil de ordem $\alpha/2$ da distribuição de Fisher com $(m-1, n-1)$ graus de liberdade.

Para consultar a tabela da distribuição de Fisher é útil saber que

$$f_{1-\alpha,v,w} = \frac{1}{f_{\alpha,w,v}}.$$

Intervalo de confiança para uma proporção p

Podemos utilizar o Teorema do Limite Central para obter intervalos de confiança aproximados para uma proporção p.

Seja $\hat{p} = X/n$ a proporção de indivíduos com uma certa característica de interesse numa amostra aleatória de dimensão n, e p a proporção de indivíduos com essa característica na população. Um intervalo de confiança aproximado para p, a um grau de confiança $1 - \alpha$, é dado por

$$\left(\hat{p} - z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) .$$

Validação de pressupostos

1. Para averiguar se uma amostra é aleatória é necessário conhecer o forma como foi recolhida para se poder avaliar se as observações são independentes e retiradas da mesma população.

2. Para averiguar se uma amostra provém duma população Normal utilizam-se várias ferramentas em conjunto:

- constroem-se gráficos: histograma, boxplot e QQ-plot (Normal); Atenção que em amostras pequenas ($n < 30$) os histogramas ficam com poucas classes, estão sujeitos a muita variabilidade e consequente são pouco fidedignos. Os boxplots apresentam a mesma limitação em amostras muito pequenas ($n < 15$). Também os QQ-plots ficam sujeitos a muita variabilidade quando as amostras são pequenas. Duma forma geral, é muito difícil (senão impossível) inferir sobre a forma de uma distribuição com base numa amostra pequena.

- realizam-se teste de ajustamento (a conhecer mais adiante).