Instituto Superior de Economia e Gestão

Análise de Dados para Finanças

Mestrado em Contabilidade, Fiscalidade e Finanças Empresariais

Ano lectivo 2019/ 2020

Objectivo Geral

A análise de dados, mais especificamente a análise de regressão, tem vindo a assumir um papel fundamental no domínio da modelação, previsão e interpretação de fenómenos de natureza económica, financeira e de gestão

Pretende-se preparar os alunos para planear e desenvolver estudos empíricos;

As técnicas a utilizar dependem, entre outros aspectos, da natureza dos dados e do objectivo da análise

Metodologia

A metodologia de ensino adoptada envolve a formação presencial em sala de computação. Consiste essencialmente na exposição de conteúdos teóricos, acompanhada de exemplos práticos e ilustrada pela via da exploração de conjuntos de dados reais com recurso a software (sobretudo o STATA)

Tópicos Programáticos

1. Análise de dados exploratória
2. Sintetização de informação via análise factorial
3. Regressão linear multipla (incluindo dummies, interacções, análise de especificação e estimação e inferência robusta à heteroscedasticidade)
4. Modelos para dados de painel
5. Modelos para dados binários
6. Modelos para dados censurados

Livros base

Wooldridge, J. M. (2016), *Introductory Econometrics: a Modern Approach*, 6th. ed., *Cengage Learning*

Newbold, P., Carlson, W. & Thorne, B. (2013), "Statistics for Business and Economics", 8th edition, Modern Prentice Hall

1. Análise exploratória de dados (Capítulos 1-2, 7-10 e 15 do Newbold)

1.1. Introdução

1.2. Principais tipos de dados

1.3. Descrição dos dados

1.4. Inferência paramétrica e não paramétrica

1.1. Introdução

- População: objecto de estudo do trabalho - é o conjunto de indivíduos cujas características se pretendem estudar
- Amostra: subconjunto finito da população que permite estimar quantidades/indicadores que descrevem as características da população de interesse
  - Se a amostra é recolhida ao acaso – amostra aleatória
  - Se toda a população é observada – censo
- Unidade amostral / observação / indivíduo: elemento da população cujas características se observam ⟶ *i*
- Dimensão da amostra: número de unidades amostrais na amostra ⟶ *n*
- Variável: característica das unidades amostrais, x_i, *i=1,…n*

1.2. Tipos de dados

Período de referência:

- Seccionais: n indivíduos observados no mesmo período de tempo
- Temporais: 1 indivíduo observado ao longo de vários momentos (T)
- Painel: n indivíduos observados ao longo de T momentos

Natureza):

- **Quantitativos**: assumem valores numéricos
  Discretos: assumem apenas valores inteiros (número de consultas médicas por paciente por mês)
  Contínuos: podem assumir qualquer valor (rendimento familiar)
- **Qualitativos**: categorias, níveis, …
  Nominal: categorias não têm ordenação natural
    - V. binária: duas categorias (género: masculino/feminino)
    - V. de categorias múltiplas (tipo de transporte: carro/metro/a pé)

  Ordinal: categorias revelam ordenação (rating de empresas - AAA, AA, A)

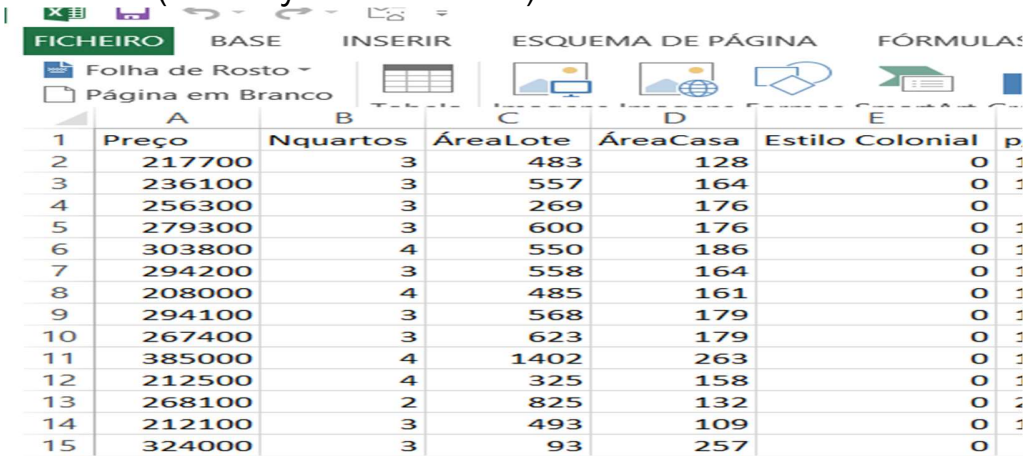
1.3. Descrição dos dados

Numa perspectiva univariada, considerar:

- tabelas de frequências, histogramas, etc.
- estatística descritiva: medidas de tendência central, dispersão, tendência não central, …

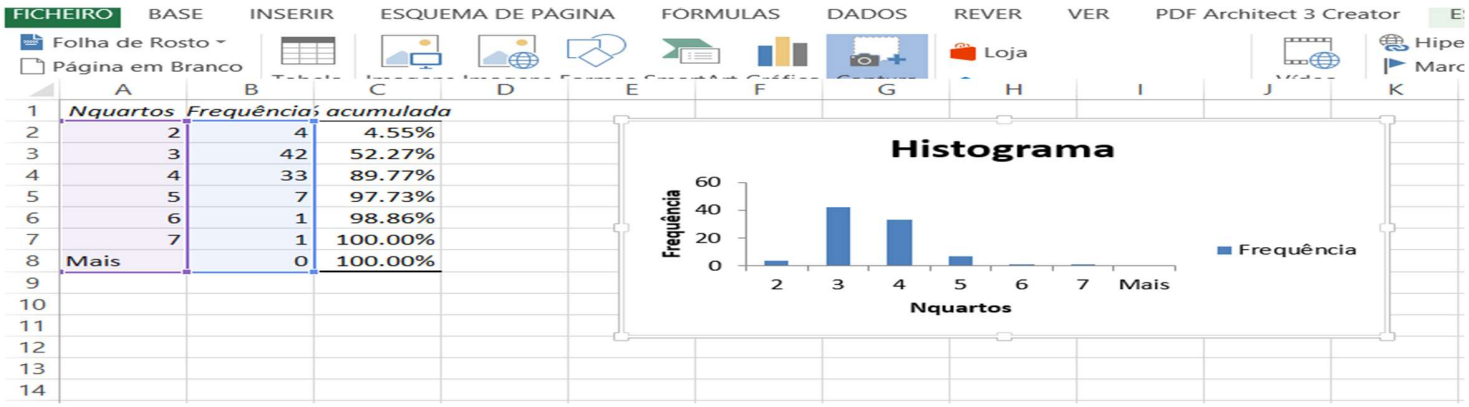
Exemplo: dados House Prices (Wooldridge, 2003), informação adaptada:

- Preço (dolares)
- Área do lote (square feet transformados em metros quadrados)
- Área da casa (idem)
- Nº de quartos (*bedrooms*)
- Colonial (dummy com 1 = sim)



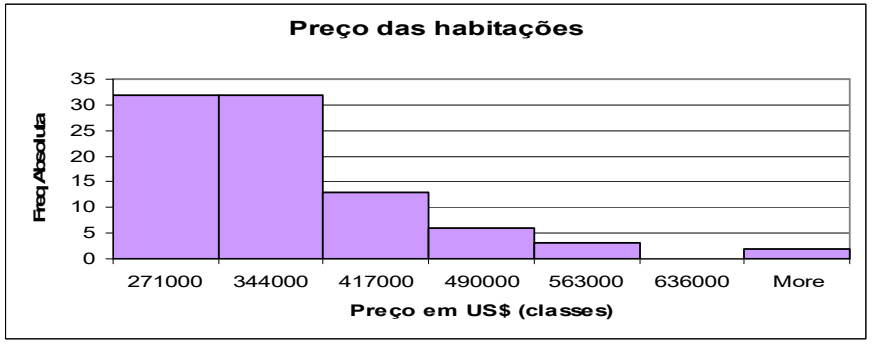
	A	B	C	D	E
1	Preço	Nquartos	ÁreaLote	ÁreaCasa	Estilo Colonial
2	217700	3	483	128	0
3	236100	3	557	164	0
4	256300	3	269	176	0
5	279300	3	600	176	0
6	303800	4	550	186	0
7	294200	3	558	164	0
8	208000	4	485	161	0
9	294100	3	568	179	0
10	267400	3	623	179	0
11	385000	4	1402	263	0
12	212500	4	325	158	0
13	268100	2	825	132	0
14	212100	3	493	109	0
15	324000	3	93	257	0

Diagramas de barras (variável discreta) – Nº de quartos



Nquartos	Frequência	acumulada
2	4	4.55%
3	42	52.27%
4	33	89.77%
5	7	97.73%
6	1	98.86%
7	1	100.00%
Mais	0	100.00%

Histograma (variável contínua) – Preço



Estatística descritiva: localização e dispersão

Medidas de localização (tendência central)

- Média (aritmética e geométrica)
- Mediana
- Moda

Média

Considerando dados não classificados:

- Aritmética: medida de localização por excelência; $\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$

- Geométrica: $m_g = \sqrt[n]{x_1\, x_2 \cdots x_n} = \left(x_1\, x_2 \cdots x_n\right)^{1/n}$

  * aplica-se apenas a variáveis positivas

Mediana

A mediana é o valor tal que 50% das observações são inferiores ou iguais e 50% das observações são superiores ou iguais à mediana. Para dados não classificados:

- Ordena-se a colecção;
- Se n for ímpar a mediana é a observação central na colecção ordenada. Se n for par a mediana é a média das 2 observações centrais na colecção ordenada.

Moda

A moda, se existir, é o valor que ocorre com maior frequência na colecção

A moda é um conceito que também se pode aplicar aos dados qualitativos. Aliás, se a variável for contínua, não é muito interessante usar a moda já que raramente é representativa.

Exemplo: retoma de HousePrices: a média e mediana do Preço é, respectivamente, 315736.364 e 290200

Medidas de Dispersão

Variância

- Trata-se da medida de dispersão associada com a média já que mede o quadrado das variações em torno da média;
- Cálculo para dados não classificados (VAR.S)

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

  - É habitual dividir por $n-1$ em vez de n (VAR.P)
    - exemplo de HousePrices: =VAR.S(A2:A89) $\rightarrow 9084841881$
- Como a variância está expressa em unidades ao quadrado, não é muito conveniente para comparar com a média. Utiliza-se assim o **desvio-padrão.**

- **Desvio-padrão** Raiz quadrada positiva da variância (STDEV.S e STDEV.P)

$$s = \sqrt{s^2}$$

Medidas de tendência não central: quartis, percentis, quantis,...

Quartis e intervalo inter-quartis

- **1º Quartil** – valor tal que aproximadamente 25% das observações lhe são inferiores e 75% superiores
- **2º Quartil - mediana**
- **3º Quartil** – valor tal que aproximadamente 75% das observações lhe são inferiores e 25% superiores; exemplo de HousePrices: =QUARTIL.EXC(A2:A89;3) → 353375

- O intervalo inter-quartis é o intervalo que medeia entre o 1º e o 3º quartil. Este intervalo vai conter 50% das observações e é composto pelos valores centrais. **A amplitude do intervalo inter-quartis**, AIQ, é utilizada como medida de dispersão.

.

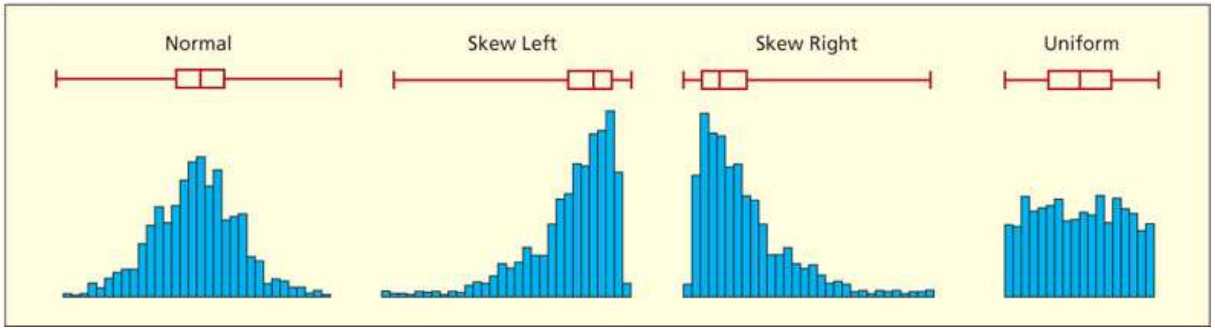
Resumo dos cinco números e gráfico *boxplot* (caixa de bigodes)

Resumo de cinco números: $x_{\min} < Q_1 < mediana < Q_3 < x_{\max}$

Boxplot (ou Box-and-whisker plot)

FIGURE 4.27

Sample Boxplots from Four Populations (n = 1000)

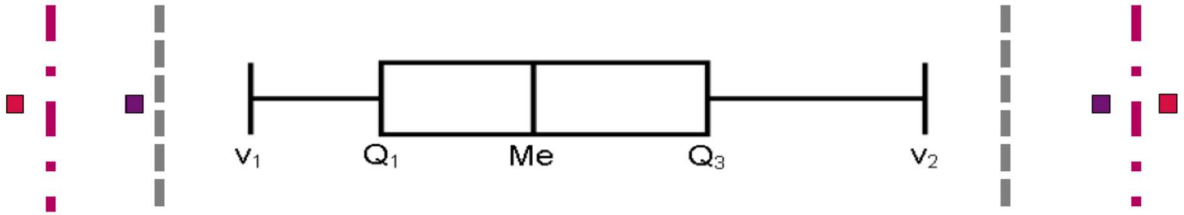


"*Outliers*"

Chamam-se *outliers* aos valores que são muito diferentes dos restantes;

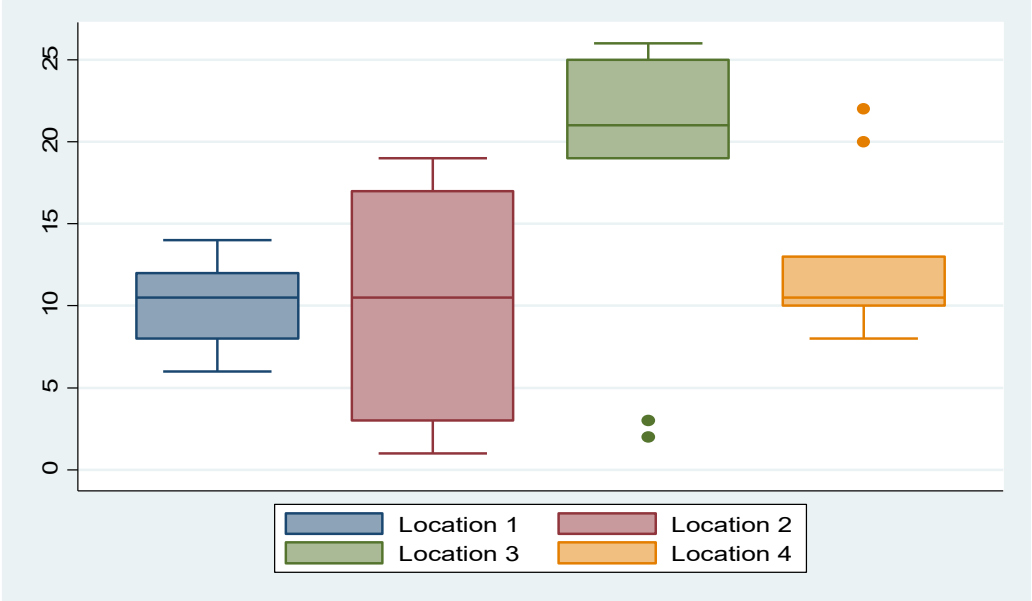
Definição habitual de *outlier*

	Moderado	*Severo*
Barreira superior	acima de Q3 + 1.5 AIQ	acima de Q3 + 3.0 AIQ
Barreira inferior	abaixo de Q1 – 1.5 AIQ	abaixo de Q1 – 3.0 AIQ



Exemplo 2.8 (Newbold) adaptado - Gilotti's Pizzeria

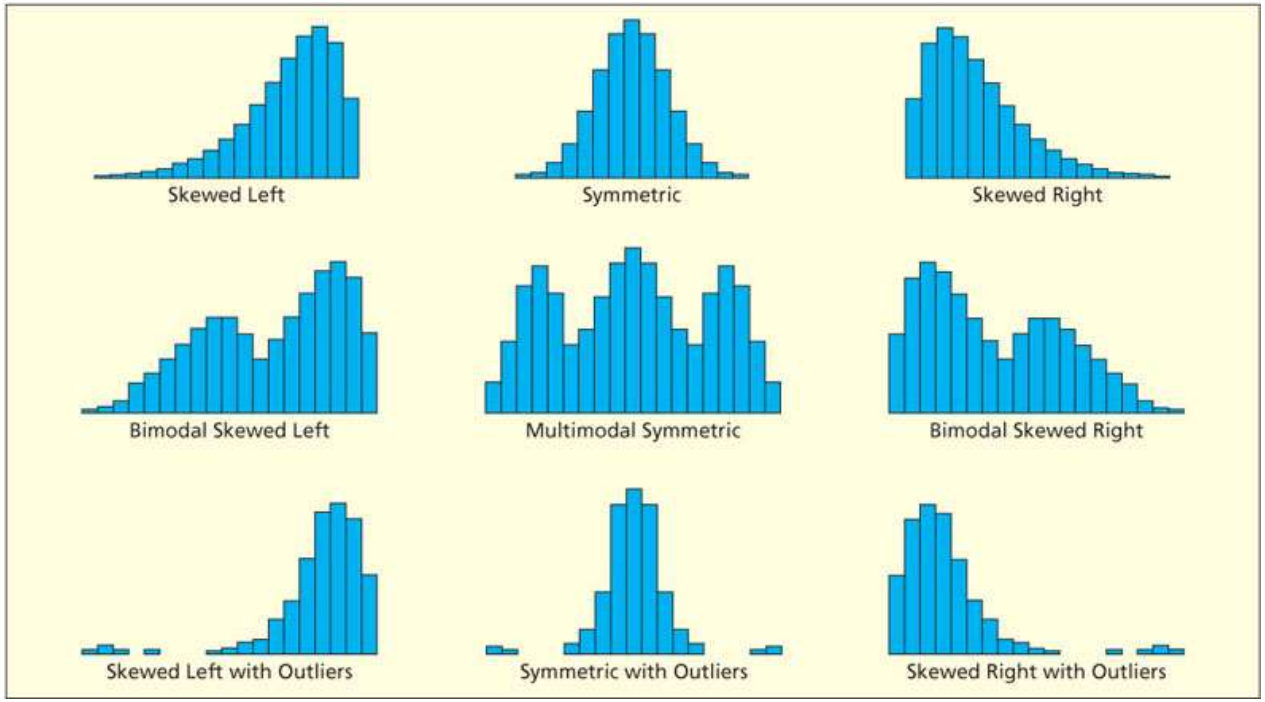
Gilotti's Pizzeria has 4 locations in one large metropolitan area. Daily sales (in hundreds of dollars) from a random sample of 10 weekdays from each of the 4 locations are given in Table 2.2. Plot the data with a box-and-whisker plot.



Forma (“shape”) da distribuição (gráfico tirado do Doane and Seward)

FIGURE 3.7

Prototype Distribution Shapes



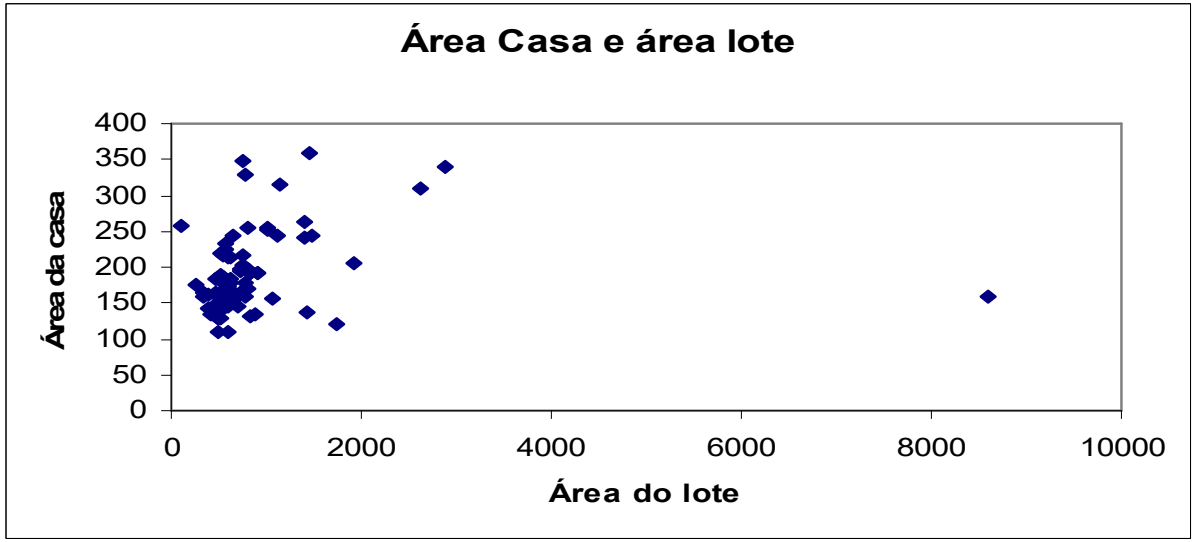
Correlação

Avalia o grau de associação entre as características;

- Variáveis quantitativas: diagrama de dispersão e coeficiente de correlação
- Variáveis qualitativas (sobretudo com poucos níveis): tabela de dupla entrada;

No que se segue, considera-se que ambas as características são quantitativas. Na componente prática estudar-se-ão as tabelas de dupla entrada.

Diagrama de dispersão (exemplo de HousePrices)



Coeficiente de correlação (Pearson): mede a associação **linear** entre as variáveis.

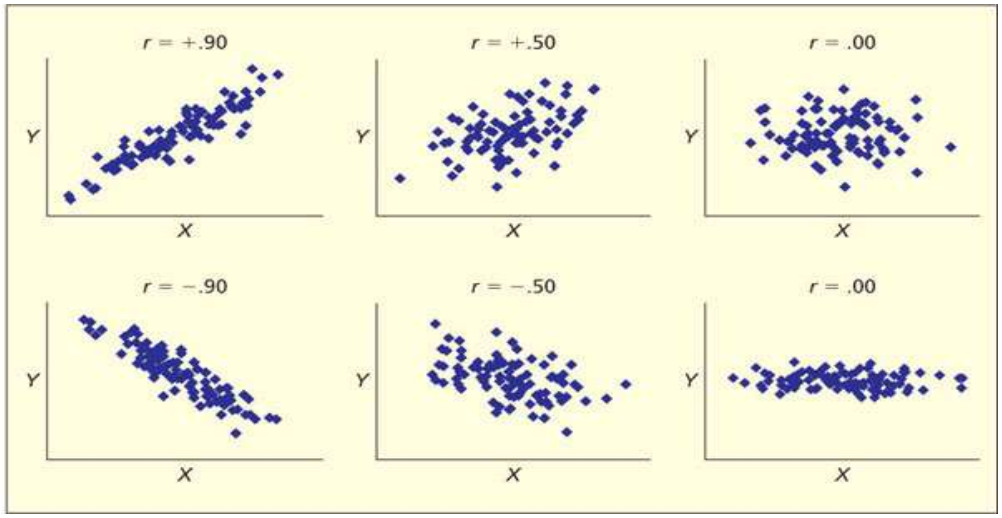
$$r_{yx} = \frac{s_{yx}}{s_y s_x}, -1 \leq r_{yx} \leq 1,$$

- s_x desvio padrão da variável x, s_y desvio padrão da variável y
- s_{xy} covariância entre x e y, $s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

- $r_{yx} = 1$ ($r_{yx} = -1$): correlação linear perfeita positiva (negativa)
- $0 < r_{yx} < 1$ ($-1 < r_{yx} < 0$): correlação linear positiva (negativa), não perfeita
- $r_{yx} = 0$: correlação nula ou não linear

- Graficamente (Doane & Seward)

FIGURE 4.33
Illustration of Correlation Coefficients



- Quando o coeficiente de correlação é calculado com base numa amostra de dimensão n, uma regra empírica consiste em considerar significativas os casos $|r_{yx}| > {}^{2}/_{\sqrt{n}}$.
- Exemplo de House prices: o coeficiente de correlação entre a área da casa e área do lote é 0.1842, valor positivo (como seria de esperar) mas porventura mais baixo do que o esperado.

1.4. Inferência paramétrica e não paramétrica

A inferência pode ser paramétrica (estuda-se aqui apenas com base na distribuição normal) ou não paramétrica

Inferência paramétrica:

- Estimação
  - Pontual: atribui-se um valor (estimativa) ao(s) parâmetro(s) desconhecido(s)
  - Intervalos: obtem-se um intervalo que contenha o parâmetro de interesse com um grau de confiança fixado pelo investigador;

$$\text{Estimativa pontual} \pm \text{margem de erro}$$

- Teste de hipóteses

- **Princípio base da inferência estatística:** Tratando-se de inferência indutiva (do particular para o geral) todas as conclusões estão sujeitas a incerteza.

Estimação por intervalos

Rever Estatística II

Exemplo: $X \sim n(\mu;\sigma)$- **Intervalo de confiança para a média (μ), variância desconhecida**

O IC de grau $1-\alpha$ é $\left(\bar{x}-t_{\alpha/2}\frac{s}{\sqrt{n}};\bar{x}+t_{\alpha/2}\frac{s}{\sqrt{n}}\right)$

- Construir um intervalo de confiança a 95% para μ a partir da amostra $(6.3;7.4;9.2;12.3;5.2;3.1;15.1;6.2;3.5;6.7)$.

O IC vem $\left(7.5-2.262\frac{3.77}{\sqrt{10}};7.5+2.262\frac{3.77}{\sqrt{10}}\right)$, isto é, $(4.803;10.197)$

Exemplo: grandes amostras, sem assumir normalidade - **Intervalo de confiança para a média (μ), variância desconhecida**

O IC de grau $1-\alpha$ é $\left(\overline{x} - z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} ; \overline{x} + z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right)$

Exemplo: De um universo X com variância finita, recolheu-se uma amostra de dimensão $n = 1000$, tendo-se observado $\overline{x} = 123.4$ e $s = 25.4$. Construir um intervalo de confiança a 90% para μ.

Como σ é desconhecido, utiliza-se $\hat{\sigma} = s$ e vem

$\left(123.4 - 1.645 \frac{25.4}{\sqrt{1000}} ; 123.4 + 1.645 \frac{25.4}{\sqrt{1000}} \right)$, isto é, $(122.07 ; 124.73)$

Exemplo: Universos de Bernoulli ($\mu = \pi$) - Intervalo de confiança para uma proporção:

- IC de grau $1-\alpha$: $\left(\overline{x} - z_{\alpha/2}\sqrt{\frac{\overline{x}(1-\overline{x})}{n}}; \overline{x} + z_{\alpha/2}\sqrt{\frac{\overline{x}(1-\overline{x})}{n}} \right)$

Exemplo: Para antecipar a votação num referendo e admitindo que todos os votos eram expressos em "sim" ou "não" recolheu-se uma amostra casual simples de dimensão 900, tendo-se observado 600 votos "sim". Construa um IC a 95% para π, proporção de votos "sim" no universo.

$1-\alpha = 0.95$ logo $z_{0.025} = 1.96$. $\overline{x} = 600/900 = 2/3$. O IC vem

$\left(\frac{2}{3} - 1.96\sqrt{\frac{(2/3)\times(1/3)}{900}}; \frac{2}{3} + 1.96\sqrt{\frac{(2/3)\times(1/3)}{900}} \right)$, isto é, $(0.636; 0.697)$

Como dimensionar a amostra?

- Ideia: fixar a margem de erro, M, e o grau de confiança, $1-\alpha$, e ver qual a dimensão da amostra necessária, n (sendo n um número inteiro)

Caso geral: $n \geq z_{\alpha/2}^2 \dfrac{\sigma^2}{M^2}$. Se σ^2 desconhecido, substituir por uma estimativa

Universo de Bernoulli: Pode-se utilizar a expressão anterior ou optar por uma visão pessimista e fazer $n \geq z_{\alpha/2}^2 \dfrac{0.25}{M^2}$ já que 0.25 é o maior valor possível para σ^2.

Exemplo: Dimensionar uma amostra num universo de Bernoulli para obter uma margem de erro não superior a 3% com uma confiança de 95%.

$M = 0.03$, $1-\alpha = 0.95$ logo $z_{\alpha/2} = 1.96$. Assim $n \geq (1.96/0.03)^2 \times 0.25 \approx 1067.11$ e portanto toma-se $n = 1068$.

Testes de hipóteses paramétricas

Rever Estatística II

O teste de hipótese é um procedimento estatístico que permite rejeitar ou não, com base numa amostra, uma dada "teoria". Procedimento:

1) **Formular as hipóteses** em teste. 3 situações habituais são:
   - $H_0 : \mu = a$ contra $H_1 : \mu \neq a$
   - $H_0 : \mu \leq a$ contra $H_1 : \mu > a$
   - $H_0 : \mu \geq a$ contra $H_1 : \mu < a$
2) **Especificar uma regra de** decisão que permita, com uma amostra, rejeitar ou não H_0
   - Definir uma **estatística de teste** adequada
   - Definir uma **região de rejeição**

Valor-p

O valor-*p* ou p_{obs}, é a probabilidade de obter um valor tão ou mais desfavorável para a hipótese nula, admitindo que esta hipótese é verdadeira. Se esta probabilidade for muito pequena rejeita-se então H_0.

Para interpretar o valor-*p*

- Valor-p>α: não rejeitar H_0
- Valor-p<α: rejeitar H_0

Exemplo: Seja μ o preço médio por m2 de uma casa. Com base na amostra observada (ficheiro dados1.xls), teste $H_0 : \mu \leq 1600$ contra $H_1 : \mu > 1600$

- $T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$

Fazer estatística descritiva no excel e depois

=(1699.65607-1600)/(52133.492/88)^.5 →4.094

=INV.T(0.05;87) →1.66 Rejeita-se H_0

Exemplo: retomar o exemplo anterior e testar $H_0 : \mu = 1600$ contra $H_1 : \mu \neq 1600$

=INV.T(0.025;87) →1.987: Rejeita-se H_0

Exemplo: Bernoulli, grandes amostras

$$Z = \frac{\bar{X} - \pi_0}{\sqrt{\pi_0 (1-\pi_0)/n}} \overset{a}{\sim} N(0,1)$$

Numa sondagem à opinião pública, em dado país, foram inquiridas 1000 pessoas e houve 53% que se disseram favoráveis a determinado projecto de lei submetido a referendo. Testar para $\alpha = 0.05$, $H_0 : p \leq 0.5$ contra $H_1 : p > 0.5$.

- =(.53-.5)/(.5*(1-.5)/1000)^.5=1.9
- Rejeita-se H_0

Exemplo – Amostra grande – teste da igualdade de duas proporções. $H_0 : \pi_1 = \pi_2$ contra uma das alternativas habituais

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\bar{X}(1-\bar{X})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \overset{a}{\sim} N(0,1) \qquad \text{com } \bar{X} = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2}{n_1 + n_2}$$

$X_1 \sim Ber(\pi_1)$, $X_2 \sim Ber(\pi_2)$, $n_1 = 110$, $n_2 = 100$, $\bar{x}_1 = 0.43$, $\bar{x}_2 = 0.45$, $H_0 : \pi_1 = \pi_2$ e $H_1 : \pi_1 \neq \pi_2$

=(0.43-0.45)/(((110*0.43+100*0.45)/210)*(1-((110*0.43+100*0.45)/210))*(1/110+1/100))^0.5

$\rightarrow -0.29$

=INV.NORMAL(0.025;0;1) $\rightarrow -1.96$ Não se rejeita H_0

Análise da variância com classificação simples (ANOVA)

Considere-se o problema da comparação entre médias de *m* populações normais,

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_m .$$

Pressupostos da ANOVA

- Dispõe-se de *m* amostras casuais, **independentes** (uma para cada população), $X_{i1}, X_{i2}, \ldots, X_{in_i} \ (i = 1, 2, \ldots, m)$
- As *m* populações têm distribuição normal com médias desconhecidas e **variância comum** também desconhecida,

$$X_{ij} \sim N(\mu_i, \sigma^2) \ (i = 1, 2, \ldots, m; j = 1, 2, \ldots, n_i).$$

- Estatística de teste: $F = \dfrac{\text{MS1}}{\text{MS2}} = \dfrac{\text{SS1}/(m-1)}{\text{SS2}/(n-m)} \sim F(m-1, n-m)$,
  - $\text{SS1} = \sum_{i=1}^{m} n_i (\overline{X}_{i\circ} - \overline{X}_{\circ\circ})^2$, $\text{n} = \sum_{i=1}^{m} n_i$, $\overline{X}_{i\circ} = \dfrac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$ $(i = 1, 2, \ldots, m)$, $\overline{X}_{\circ\circ} = \dfrac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} X_{ij}}{\sum_{i=1}^{m} n_i} = \dfrac{\sum_{i=1}^{m} n_i \overline{X}_{i\circ}}{n}$
  - $\text{SS2} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i\circ})^2$
- Rejeita-se quando $F_{obs} > F_{\alpha}$.

- Tabela da ANOVA

Origem da variação	Soma de quadrados	Graus de liberdade	Médias quadráticas
Entre amostras	SS1	$m-1$	$\text{MS1} = \text{SS1}/(m-1)$
Dentro das amostras	SS2	$n-m$	$\text{MS2} = \text{SS2}/(n-m)$
Total	SST	$n-1$	$F = \text{MS1}/\text{Ms2}$

Quando se rejeita H_0 (existem diferenças estatisticamente significativas entre as m médias) é, muitas vezes, interessante avaliar se estes contrastes são generalizados ou se apenas se verificam entre alguns grupos.

- Quando se efectua uma sucessão de testes recomenda-se corrigir os valores-p (ou regiões de rejeição) para evitar uma sobre rejeição (com 20 testes em cadeia e $\alpha = 0.05$ rejeitar-se-á indevidamente uma vez). Das várias alternativas a mais conhecida é a de Bonferroni

Exemplo: para amostras casuais independentes de 3 populações, pretende-se saber se as médias dos universos donde provêm podem ser consideradas iguais

Pop 1	13	27	26	22	26		
Pop 2	43	35	47	32	31	37	
Pop 3	33	37	33	26	44	33	54

```
oneway variable population ,bonferroni
                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      760.453968      2   380.226984      6.78     0.0080
 Within groups      841.157143     15   56.0771429
------------------------------------------------------------------------
    Total           1601.61111     17   94.2124183
Bartlett's test for equal variances:  chi2(2) =   1.1727  Prob>chi2 = 0.556

                    Comparison of Variable by Population
                                (Bonferroni)
Row Mean-|
Col Mean |          1          2
---------+----------------------
       2 |       14.7
         |      0.016
         |
       3 |    14.3429   -.357143
         |      0.015      1.000
```

A 5% de significância, rejeita-se a igualdade das 3 médias. Contudo, não se rejeita a igualdade da média 2 com 3

Testes de hipóteses não paramétricas – Introdução

Testar-se se se a localização (geralmente a mediana) das populações é igual sem assumir nenhuma distribuição; A diferença com os testes sob normalidade (ou grandes amostras) é que as estatísticas de teste se baseiam no "rank" das observações.

- Testes mais conhecidos
  - Amostras emparelhadas (comparam a mediana das diferenças das observações)
    - Teste do sinal (*sign test*)
    - Teste (do posto-sinal) de Wilcoxon (*Wilcoxon signed rank*)
  - 2 amostras independentes
    - Teste U de Man-Whitney (com a variante *Wilcoxon rank sum*)
  - *k* amostras independentes
    - Teste de Kruskal-Wallis

Teste de Kruskal-Wallis

Muitas vezes considerado uma versão não paramétrica da ANOVA, permite testar:

- Igualdade de médias (ou medianas) em *m* populações contínuas.
- Igualdade das funções de distribuição nas *m* populações.

- Ideia base: para *m* amostras casuais, independentes (uma para cada população), $X_{i1}, X_{i2}, \ldots, X_{in_i}$ $(i = 1, 2, \ldots, m)$: passar das observações aos "ranks" (em termos da amostra no seu todo) e verificar se a distribuição destes "ranks" é "semelhante" pelas várias populações.

Descrição geral do teste

Para $n = \sum_{i=1}^{m} n_i$ o "rank" da observação X_{ij} é dado por r_{ij}. Defina-se $S_i = \sum_{j=1}^{n_i} r_{ij}$ (soma dos "ranks" referentes à população *i*). Se as populações fossem todas de igual dimensão esperar-se-ia que estas somas fossem semelhantes. Calcule-se

$$S_P = \sum_{j=1}^{n_i} \left(S_i^2 / n_i \right),\ S_R = \sum_{i=1}^{m} \sum_{j=1}^{n_i} r_{ij}^2 \text{ e } C = \frac{n \times (n+1)^2}{4}$$

- A estatística de teste vem $Q = \dfrac{(n\text{-}1) \times (S_P\text{-}C)}{(S_R - C)}$
- Distribuição de *Q*
  - Amostras pequenas: valores críticos de Q em tabela especifica
  - Amostras moderadas ou grandes, a distribuição de *Q* pode ser aproximada por uma qui-quadrado com *m*-1 graus de liberdade.

Exemplo: retoma do caso tratado por ANOVA

```
. kwallis pop, by(type)

Kruskal-Wallis equality-of-populations rank test
  +-------------------------+
  | type | Obs | Rank Sum |
  |------+-----+----------|
  |    1 |   5 |    17.00 |
  |    2 |   6 |    72.50 |
  |    3 |   7 |    81.50 |
  +-------------------------+

chi-squared =     9.061 with 2 d.f.
probability =     0.0108

chi-squared with ties =     9.146 with 2 d.f.
probability =     0.0103
```

A decisão não vem diferente daquela que se tomou com base na ANOVA

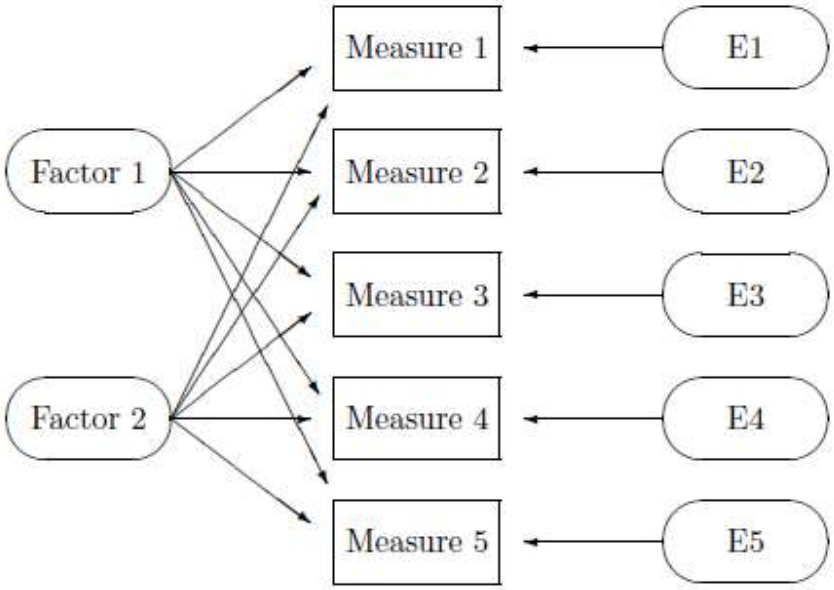
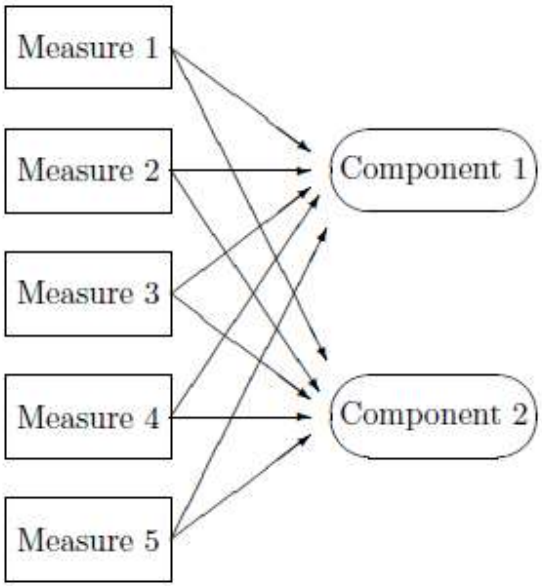
Análise Factorial: introdução

O objectivo é identificar um conjunto de factores (não observáveis) que possam explicar a correlação entre as variáveis disponíveis, de forma a sintetizar a informação nelas contida. Basicamente, sumaria-se a informação de um conjunto de variáveis num menor número de variáveis latentes, os factores

Estas técnicas baseiam-se na análise da matriz de correlações do conjunto de variáveis disponíveis, as qual se baseia no coeficiente de correlação linear de Pearson:

- $r_{yx} = \frac{s_{yx}}{s_y s_x}$, $-1 \leq r_{yx} \leq 1$, onde s_x e s_y são o desvio padrão de X e Y e s_{xy} a covariância
- Para um conjunto de p variáveis, tem-se uma matriz de correlações p x p com 1's na diagonal principal

Análise Factorial / Análise de componentes principais

AF	ACP
	

*Equações estruturais: descrevem casos onde os factores se influenciam

Análise Factorial / Análise de componentes principais

Análise factorial:

$$X_i = \underbrace{a_{i1}FC_1 + a_{i2}FC_2 + \ldots + a_{im}FC_m}_{parte\ comum} + e_1$$

Análise de componentes principais

$$CP_i = a_{1i}X_1 + a_{2i}X_2 + \ldots + a_{pi}X_p$$

onde

p=#variáveis, m=#factores, i=1,...m

FC=factores comuns, e=erro

a=coeficientes designados de loadings (não são coeficientes de regressão, pois os factores não são observados)

Etapas da análise

1. Construção e análise da matriz de correlações das variáveis observadas
2. Extração de factores e decisão sobre quantos manter na análise
3. Eventual rotação dos factores
4. Interpretação do significado de cada um dos factores
5. Possível utilização dos factores como variáveis noutra análise (por exemplo como variáveis explicativas num modelo de regressão)

Análise de correlação

Indicador de Kaiser-Meyer-Olken (KMO) – sumaria o nível de correlação entre as variáveis e permite verificar se as correlações são relevantes de acordo com

- *0.00 to 0.49 unacceptable*
- *0.50 to 0.59 miserable*
- *0.60 to 0.69 mediocre*
- *0.70 to 0.79 middling*
- *0.80 to 0.89 meritorious*
- *0.90 to 1.00 marvelous*

Extração de factores

Para obter os factores pode-se utilizar o método das componentes principais ou o método da máxima verosimilhança, por exemplo

Escolha do nº de factores:

- Deve-se ver qual a parte da variância das variáveis originais que fica explicada. Esta análise é facilitada pelo facto do software apresentar os factores por ordem decrescente de capacidade explicativa (os primeiros explicam mais). Utilizando o eighenvalue de cada factor (tem-se pelo somatório dos respectivos loadings ao quadrado) obtém-se a proporção de variância explicada pelo factor dividindo o eighevalue respectivo por p.

- Pode-se utilizar o critério de Kaiser: manter os factores com eighenvalues >1, já que o eighevalue do factor pode ser visto como o número de variáveis que o factor sumaria

Rotação de factores

Os *loadings* associados aos factores não são unicos (existem múltiplas soluções). Assim, é comum proceder à rotação dos factores, que consiste essencialmente na imposição de restrições adicionais e que facilita a interpretação dos factores, extremando a contribuição da cada variável para o factor

- Rotação ortogonal: gera factores que não se correlacionam entre si e *loadings* entre ±1 (varimax, quartimax, equimax,…)
- Rotação obliqua: gera factores que podem estar correlacionados (oblimax, quartimin, …)

Interpretação de factores

Procedendo à análise dos *loadings*, muitas vezes é possível identificar o significado de um factor.

Exemplo: Considere os dados em factorial.sav, recolhidos junto de 30 indivíduos. O objectivo é determinar os benefícios que os consumidores procuram ao mudarem a sua residência para outra região. Os entrevistados tinham de indicar o seu acordo ou desacordo através de uma escala de 7 pontos (1= total desacordo e 7= totalmente de acordo) sobre as seguintes afirmações:

V1 = É muito importante ter em conta o local onde vive a família
V2 = Eu mudo para o local onde possa ter um salário mais elevado
V3 = Eu procuro um local com melhores condições em termos de infraestruturas (escolas, acessibilidade, hospitais,…)
V4 = Eu prefiro um local onde o custo de vida seja mais baixo
V5 = A qualidade de vida que se pode disfrutar num local não é um factor determinante na escolha do local de residência
V6 = O principal motivo da mudança é a possibilidade de progredir na carreira profissional

Análise de correlação e da medida KMO

```
. cor v1 v2 v3 v4 v5 v6
             |       v1       v2       v3       v4       v5       v6
-------------+------------------------------------------------------
          v1 |   1.0000
          v2 |  -0.0532   1.0000
          v3 |   0.8731  -0.1550   1.0000
          v4 |  -0.0862   0.5722  -0.2478   1.0000
          v5 |  -0.8576   0.0197  -0.7778  -0.0066   1.0000
          v6 |   0.0042   0.6405  -0.0181   0.6405  -0.1364   1.0000
```

```
. quietly factor  v1 v2 v3 v4 v5 v6
. estat kmo
Kaiser-Meyer-Olkin measure of sampling adequacy
    ----------------------
        Variable |     kmo
    -------------+--------
              v1 |  0.6206
              v2 |  0.6973
              v3 |  0.6787
              v4 |  0.6367
              v5 |  0.7687
              v6 |  0.5612
    -------------+--------
         Overall |  0.6600
    ----------------------
```

Verifica-se que várias variáveis apresentam correlação elevada. A estatística de KMO está perto de um valor aceitável. Valerá a pena enveredar por uma análise factorial.

Análise Factorial: método da componentes principais

```
. factor  v1 v2 v3 v4 v5 v6, pcf
(obs=30)
Factor analysis/correlation                       Number of obs    =         30
    Method: principal-component factors           Retained factors =          2
    Rotation: (unrotated)                         Number of params =         11

    --------------------------------------------------------------------------
        Factor  |   Eigenvalue   Difference        Proportion   Cumulative
    ------------+-------------------------------------------------------------
       Factor1  |      2.73119      0.51307            0.4552       0.4552
       Factor2  |      2.21812      1.77652            0.3697       0.8249
       Factor3  |      0.44160      0.10034            0.0736       0.8985
       Factor4  |      0.34126      0.15863            0.0569       0.9554
       Factor5  |      0.18263      0.09742            0.0304       0.9858
       Factor6  |      0.08521            .            0.0142       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(15) =  115.57 Prob>chi2 = 0.0000
```

Sugere a extracção de 2 factores. De facto, os dois primeiro factores têm eigenvalues superiores a 1, sendo a diferença do primeiro para o segundo de 0.513 e do segundo para o terceiro de 1.777. A proporção da variância das variáveis captada pelo Factor 1 é de 45.52% (2.73119/6) e a do segundo é de 36.97% (2.21812/6), de tal forma que a variância acumulada explicada pelos primeiros dois factores é de 82.49%

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
v1	**0.9283**	0.2532	0.0741
v2	-0.3005	**0.7952**	0.2773
v3	**0.9362**	0.1309	0.1064
v4	-0.3416	**0.7890**	0.2609
v5	**-0.8688**	-0.3508	0.1222
v6	-0.1766	**0.8712**	0.2099

Este quadro faz a decomposição dos eighenvalues por variável, informando quais as variáveis que mais contribuem para a formação do factor e permitindo a sua interpretação. O primeiro factor tem pesos elevados em v1, v3 e v5, que parecem ser variáveis associadas à "qualidade de vida" e o Factor 2 com v2, v4 e v6, que parece traduzir "questões profissionais".

Alguma simulação de cálculos:

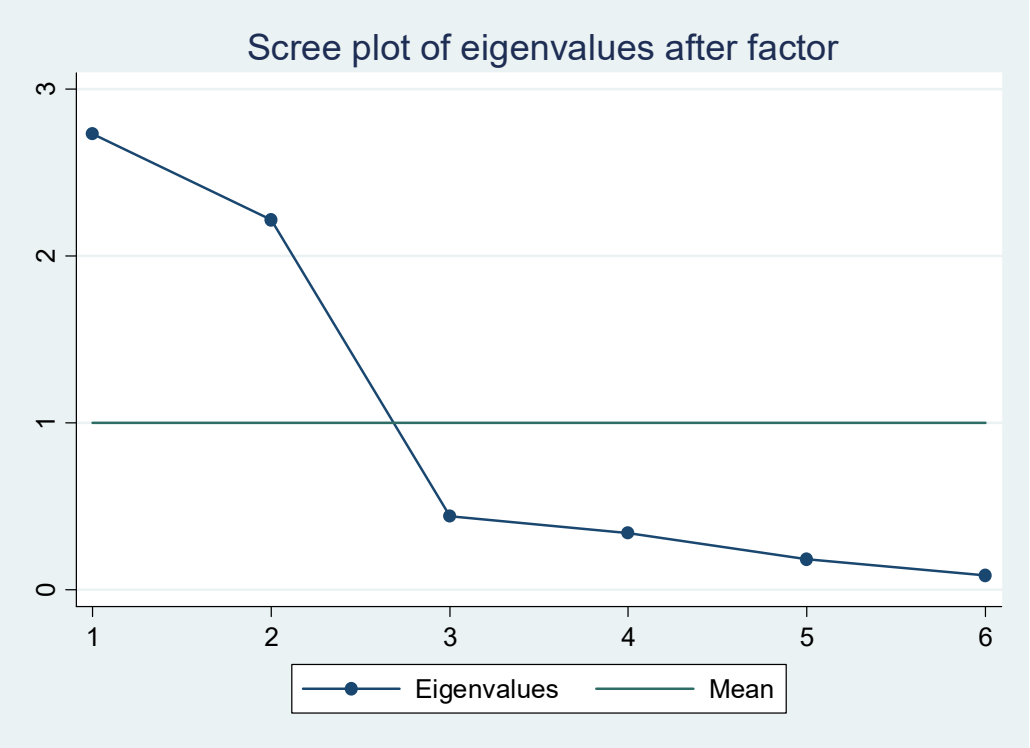
```
. display (0.9283^2+0.3005^2+0.9362^2+0.3416^2+0.8688^2+0.1766^2)
2.7312031
. display (0.9283^2+0.2532^2)
.92585113
. display(1-.92585113)
.07414887
```

(eighenvalue do factor 1)

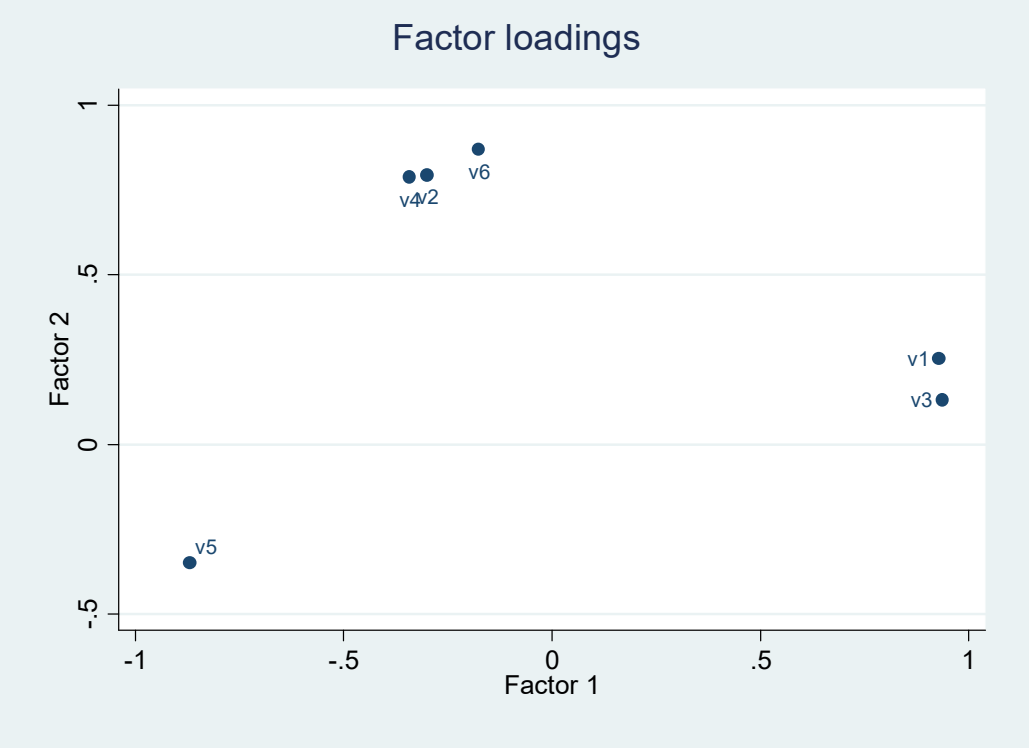
(parte da variável explicada pelos dois factores)

(uniqueness: parte da variável que ficou por explicar)

```
. screeplot, mean
```



```
. loadingplot
```



Rotação de factores para acentuar os loadings e tornar a interpretação mais fácil

```
. rotate
Factor analysis/correlation                          Number of obs    =         30
    Method: principal-component factors              Retained factors =          2
    Rotation: orthogonal varimax (Kaiser off)        Number of params =         11
    ------------------------------------------------------------------------------
        Factor  |     Variance   Difference        Proportion   Cumulative
    ------------+-----------------------------------------------------------------
       Factor1  |      2.68990      0.43048            0.4483       0.4483
       Factor2  |      2.25941            .            0.3766       0.8249
    ------------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(15) =  115.57 Prob>chi2 = 0.0000
```

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
v1	**0.9620**	-0.0205	0.0741
v2	-0.0626	**0.8478**	0.2773
v3	**0.9349**	-0.1401	0.1064
v4	-0.1037	**0.8535**	0.2609
v5	**-0.9326**	-0.0899	0.1222
v6	0.0778	**0.8855**	0.2099

Factor rotation matrix

	Factor1	Factor2
Factor1	0.9589	-0.2837
Factor2	0.2837	0.9589

- As contribuições estão agora mais acentuadas. No output também consta a matriz que permite passar de umas contribuições para outras
- Pode-se utilizar outro método de extração

Geração dos factores para cada um dos indivíduos da amostra (temos agora 2 variáveis latentes que sumariam as 7 variáveis iniciais – ver nos dados)

```
. predict factor1 factor2

(regression scoring assumed)
Scoring coefficients (method = regression; based on varimax rotated factors)
    ---------------------------------
        Variable |  Factor1   Factor2
    -------------+-------------------
              v1 |  0.35833   0.01304
              v2 | -0.00380   0.37501
              v3 |  0.34543  -0.04066
              v4 | -0.01902   0.37656
              v5 | -0.34988  -0.06141
              v6 |  0.04940   0.39496
```

Análise Factorial: método da máxima verosimilhança

```
. factor  v1 v2 v3 v4 v5 v6, ml
(obs=30)

number of factors adjusted to 3
Iteration 0:   log likelihood = -4.9672274
…
Factor analysis/correlation                       Number of obs    =        30
    Method: maximum likelihood                    Retained factors =         3
    Rotation: (unrotated)                         Number of params =        15
                                                  Schwarz's BIC    =   51.6724
    Log likelihood = -.3272396                    (Akaike's) AIC   =   30.6545
    Beware: solution is a Heywood case
            (i.e., invalid or boundary values of uniqueness)
    ------------------------------------------------------------------------------
         Factor  |   Eigenvalue   Difference        Proportion   Cumulative
    -------------+----------------------------------------------------------------
        Factor1  |      1.83935     -0.71754            0.3821       0.3821
        Factor2  |      2.55688      2.13956            0.5312       0.9133
        Factor3  |      0.41732            .            0.0867       1.0000
    ------------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(15) =  115.57 Prob>chi2 = 0.0000
    (the model with 3 factors is saturated)
```

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Uniqueness
v1	0.0042	**0.9852**	0.0547	0.0262
v2	**0.6405**	-0.0773	0.3009	0.4933
v3	-0.0181	**0.9004**	-0.2552	0.1238
v4	**0.6405**	-0.1169	0.5085	0.3176
v5	-0.1364	**-0.8694**	-0.0093	0.2255
v6	**1.0000**	-0.0000	-0.0000	0.0000

```
. rotate
Factor analysis/correlation                       Number of obs    =        30
    Method: maximum likelihood                    Retained factors =         3
    Rotation: orthogonal varimax (Kaiser off)     Number of params =        15
                                                  Schwarz's BIC    =   51.6724
    Log likelihood = -.3272396                    (Akaike's) AIC   =   30.6545
    Beware: solution is a Heywood case
            (i.e., invalid or boundary values of uniqueness)
```

Factor	Variance	Difference	Proportion	Cumulative
Factor1	2.56011	0.75442	0.5319	0.5319
Factor2	1.80569	1.35794	0.3751	0.9070
Factor3	0.44775	.	0.0930	1.0000

```
    LR test: independent vs. saturated:  chi2(15) =  115.57 Prob>chi2 = 0.0000
```

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Uniqueness
v1	**0.9838**	-0.0378	0.0669	0.0262
v2	-0.0555	**0.6329**	0.3210	0.4933
v3	**0.9023**	-0.0464	-0.2446	0.1238
v4	-0.0979	**0.6278**	0.5279	0.3176
v5	**-0.8740**	-0.1005	-0.0245	0.2255
v6	0.0403	**0.9986**	0.0331	0.0000

Factor rotation matrix

	Factor1	Factor2	Factor3
Factor1	0.0403	0.9986	0.0331
Factor2	0.9991	-0.0407	0.0123
Factor3	-0.0136	-0.0326	0.9994