DATA ANALYSIS IN ACCOUNTING

Master in Accounting

Academic year 2020/ 2021

1. Introduction to data analysis (Chapters 1-2, 7-10 e 15, Newbold, 2013)

1.1. Definitions / notation

1.2. Some types of data

1.3. Data description

1.4. Parametric and nonparametric inference

1.1. Definitions / Notation

- Population: object of analysis of the empirical work – includes the individuals (firms, workers, countries, ..) for which we wish to analyse some characteristics

- Sample: subset of the population from which we estimate the quantities of interest

  - A random sample is assumed to be available

  - In a census, the sample coincides with the population

- Sampling unit /observation/ individual → *i*

- Sample size → *n*

- Variable: x_i, *i=1,...n*

1.2. Some types of data

- Cross sectional: n individuals are observed at a given moment

TABLE 1.1 A Cross-Sectional Data Set on Wages and Other Individual Characteristics

obsno	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.

- Time series: 1 individual is observed over T periods

TABLE 1.3 Minimum Wage, Unemployment, and Related Data for Puerto Rico

obsno	year	avgmin	avgcov	prunemp	prgnp
1	1950	0.20	20.1	15.4	878.7
2	1951	0.21	20.7	16.0	925.0
3	1952	0.23	22.6	14.8	1015.9
.	.	.	.	.	.
.	.	.	.	.	.

- Panel: n individuals are observed over T periods

TABLE 1.5 A Two-Year Panel Data Set on City Crime Statistics

obsno	city	year	murders	population	unem	police
1	1	1986	5	350,000	8.7	440
2	1	1990	8	359,200	7.2	471
3	2	1986	2	64,300	5.4	75
4	2	1990	1	65,100	5.5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

- Pooled data: a cross sectional sample is available for several periods but the individuals at different periods are not necessarily the same

TABLE 1.4 Pooled Cross Sections: Two Years of Housing Prices

obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85,500	42	1600	3	2.0
2	1993	67,300	36	1440	3	2.5
3	1993	134,000	38	2000	4	2.5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
250	1993	243,600	41	2600	4	3.0
251	1995	65,000	16	1250	2	1.0
252	1995	182,400	20	2200	4	2.0
253	1995	97,500	15	1540	3	2.0
.	.	.	.	.	.	.

1.3. Data description

In an univariate approach, we may consider:

- Frequency tables, histograms, box-plots, etc.
- Descriptive statistics: central tendency measures (mean, median, mode), dispersion (standard error, variance, interquartile measures, …), noncentral tendency (quantiles, percentiles, …)

Descriptive statistics: location and dispersion

Location (central tendency)

- Mean (arithmetic, geometric, …)
- Median
- Mode

Mean

- Arithmetic: the typical location measure: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Geometric (only for positive variables): $\bar{x}_G = \left(\prod_{i=1}^{n} x_i\right)^{1/n} = \frac{\sum_{i=1}^{n} ln(x_i)}{n}$

Median

A value defined such that 50% of the observations are smaller and 50% are larger.

- The observations of the variable are ordered. Then, if n in odd the median is the central observation of the collection. If n is even, the median is the arithmetic mean of the two central observations of the ordered data.

Mode

The mode is the most frequent value of the variable

Dispersion

Variance

- Measures the square of the variation to the mean;

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

- The variance reflects the (squared) measurement unit of the variable in analysis. The magnitude is not informative.

Standard error

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

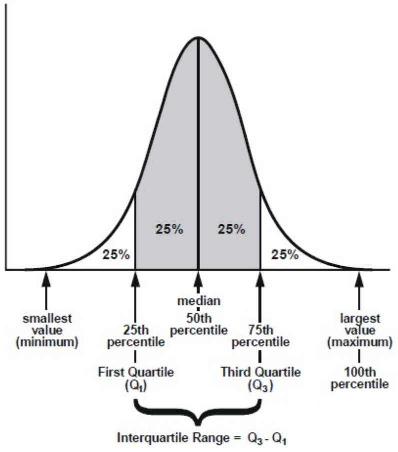
- The measurement unit is directly captured, but the magnitude is again dependent on that reference: for example, if one measures in euros, instead of hundred euros, s is 100 times larger.

Non central tendency measures: quartile, percentiles, quantiles,…

Quartiles and inter-quartile interval

- **1º Quartile** (Q1) – value for which 25% of the observations are inferior and 75% superior
- **2º Quartile** (Q2) - median
- **3º Quartile** (Q3) – value for which 75% of the observations are inferior and 25% superior

Interquartile range (IQR): Q3-Q1, containing, thus, 50% of the observations (includes the central values). It is considered a dispersion measure.



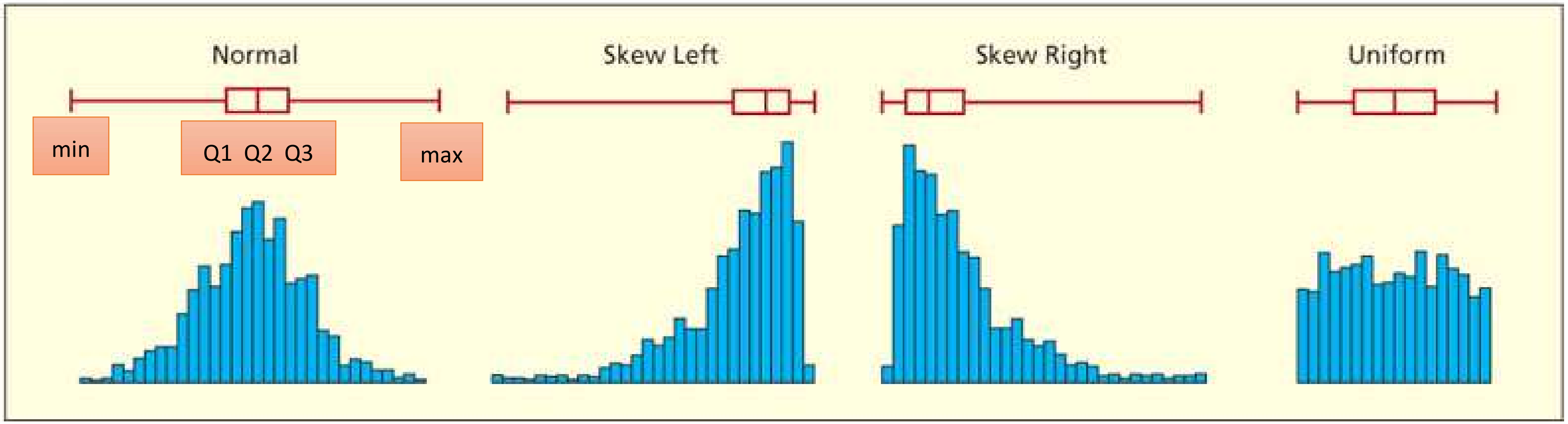
Summarizing five numbers: boxplots

The 5 numbers: $x_{min} < Q_1 < Median < Q_3 < x_{max}$

Boxplot (ou Box-and-whisker plot)

FIGURE 4.27

Sample Boxplots from Four Populations (n = 1000)

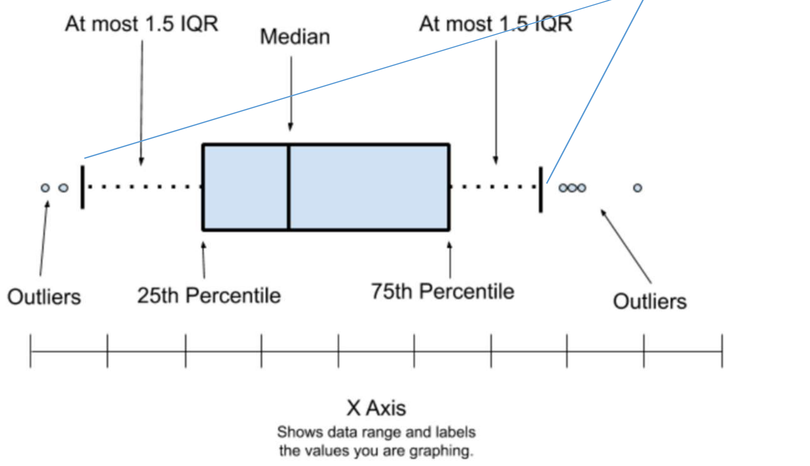


“*Outliers*”

Observations for which the value of the variable is distant form the others

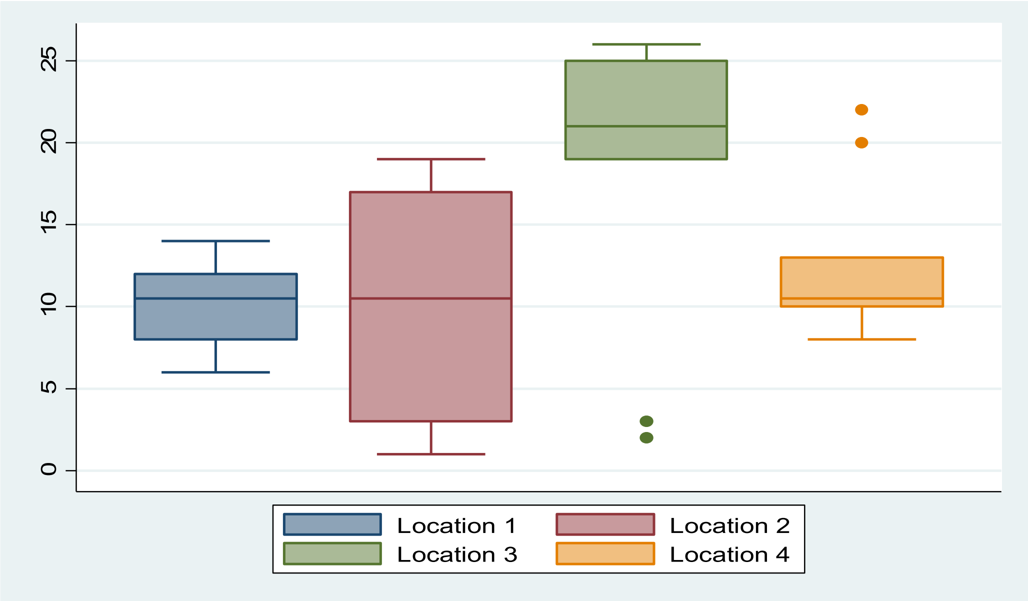
Usual definition of *outlier*

	Moderate	*Severe*
Top limit	larger than Q3 + 1.5 IQR	Larger than Q3 + 3.0 IQR
Lower limit	smaller than Q1 – 1.5 IQR	Smaller than Q1 – 3.0 IQR



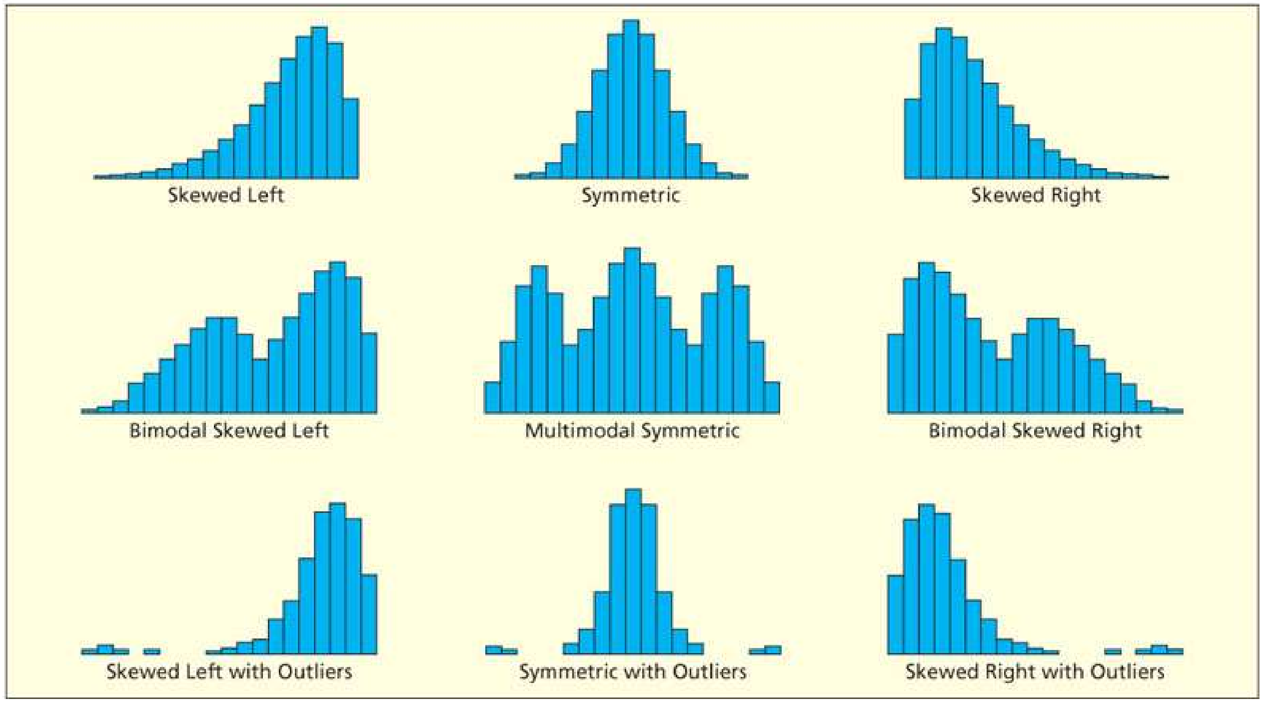
Exemplo 2.8 (Newbold) adaptado - Gilotti's Pizzeria

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. The box- plot is



Shape of the distribution (Doane and Seward)

FIGURE 3.7
Prototype Distribution Shapes



Correlation

Evaluates the degree association between two variables;

- Quantitative variables: scatterplots, correlation coefficient
- Qualitative variables (assuming a few diferente values): contingency table

Illustration for two quantitative variables. Contingency tables will be produced in Stata.

Correlation coefficient (Pearson): informs on the linear association between two variables

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

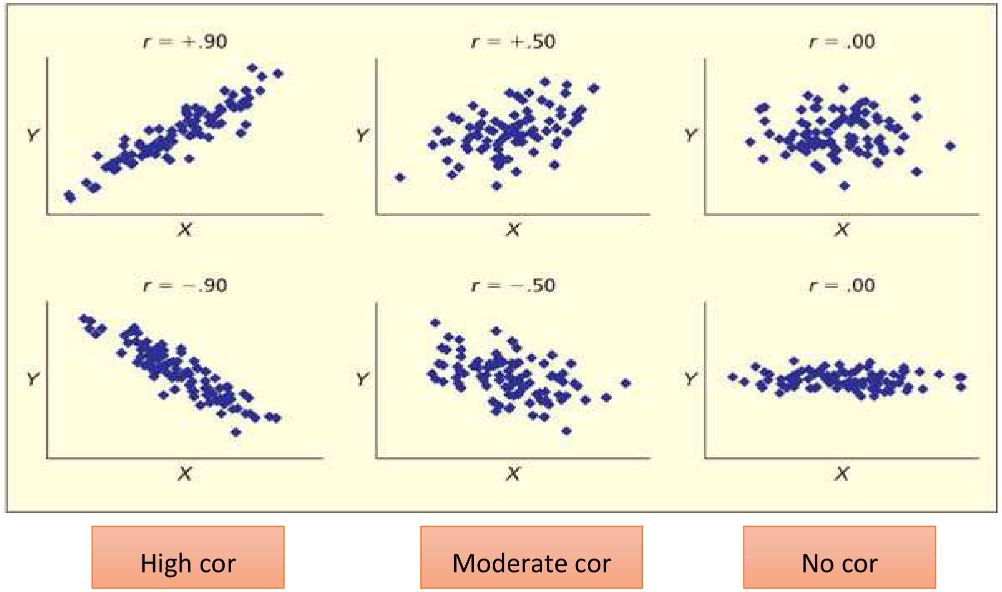
- ○ s_x standard deviation of x, s_y standard deviation of y
- ○ s_{xy} covariance between x and y, $s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

- $r_{yx} = 1$ ($r_{yx} = -1$): perfect positive (negative) linear correlation
- $0 < r_{yx} < 1$ ($-1 < r_{yx} < 0$): linear positive (negative) correlation
- $r_{yx} = 0$: absence of linear correlation (either no correlation or nonlinear correlation)

- Illustration (Doane & Seward)

FIGURE 4.33

Illustration of Correlation Coefficients



- For a sample of size n, it is considered that the correlation is significant for $|r_{yx}| > {}^{2}/_{\sqrt{n}}$.

1.4 Parametric and nonparametric estimation and inference

The parametric approach considered here relies on the assumption of the normal distribution. This assumption is relaxed by the nonparametic approach

Parametric approach:

- Estimation
  - Point: an estimate is obtained for the (unknown) parameter of interest
  - Interval: an interval is obtained, that contains the true value of the parameter of interest at a given confidence level (chosen by the researcher):

$$\text{Point estimate} \pm \text{margin of error}$$

- Hypothesis testing

Basic principle of statistical inference: as an inductive inference procedure (particular to general) all the conclusions are subject to uncertainty

Interval estimation

Statistics II - review

Example: for $X \sim n(\mu;\sigma)$ **the confidence interval (CI) for the mean** (μ), unknown variance, with level $(1-\alpha)$ is

$$\left(\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}; \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right)$$

- CI at the 95% level for μ, using the data (6.3; 7.4; 9.2; 12.3; 5.2; 3.1; 15.1; 6.2; 3.5; 6.7) or which $\bar{x} = 7.5$ and $s = 3.77$

$$\left(7.5 - 2.262\frac{3.77}{\sqrt{10}}; 7.5 + 2.262\frac{3.77}{\sqrt{10}}\right)$$, that is $(4.803;10.197)$

Example: large samples, without the normality assumption **– CI for the mean**, (μ), unknown variance at the $(1-\alpha)$ level;

$$\left(\bar{x} - z\alpha/_2 \frac{s}{\sqrt{n}} ; \bar{x} + z\alpha/_2 \frac{s}{\sqrt{n}}\right)$$

- Consider a sample of size n of variable X with $\bar{x} = 123.4$ and $s = 25.4$. The CI at the 90% level for μ is

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

using $\hat{\sigma} = s$, because σ is unknown

Example: consider a Bernoulli variable X (Xϵ{0,1}, hence the mean equals the proportion of successes $\mu = \pi$) - **CI for a proportion** at level $1-\alpha$:

$$\left(\bar{x} - z_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}\,;\,\bar{x} + z_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}\right)$$

- With the aim of anticipating the voting result for a given decision, where votes are expressed as “yes” or “no”, a sample of size 900 was collected, with 600 votes for “yes”. The CI at 95% for π, the proportion of votes for “yes” is

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

Note: we use $\bar{x} = 600/900 = 2/3$

How to choose the sample size n?

- Idea: define both the margin of error M , and confidence level $1-\alpha$, and then obtain n (n defined as an integer)

General case: $\downarrow n \geq z_{\alpha/2}^2 \frac{\sigma^2}{M^2 \uparrow}$. Note: if σ^2 is unknown, replace by an estimate

Bernoulli case: the previous result is considered with σ^2=0.25 (0.25 is the largest possible value for σ^2): $n \geq z_{\alpha/2}^2 \frac{0.25}{M^2}$.

Example: choose the sample size in the framework of a Bernoulli population to obtain a margin of error not larger than 3% with a confidence of 95%.

Because $M = 0.03$ and $1-\alpha = 0.95$, $z_{\alpha/2} = 1.96$.

Thus, $n \geq (1.96/0.03)^2 \times 0.25 \approx 1067.11$, which means that $n = 1068$.

Parametric hypothesis testing (Statistics II: review)

An hypothesis test is a statistical procedure that allows to reject or not to reject, using a sample, a given "theory". Procedure:

1) **Formulate the hypothesis of the test**. Usually:

- $H_0 : \mu = a$ versus $H_1 : \mu \neq a$

but it is also possible to consider

- $H_0 : \mu \leq a$ versus $H_1 : \mu > a$
- $H_0 : \mu \geq a$ versus $H_1 : \mu < a$

2) **Specify a decision rule** that, for a given sample, allows to reject or not H_0

- Define a suitable **test statistic**
- Define the **rejection (critical) region** depends on the significance level α (typically 0.05, or 0.01 or 0.1 - type I error – probability of rejection of a true null hypothesis)

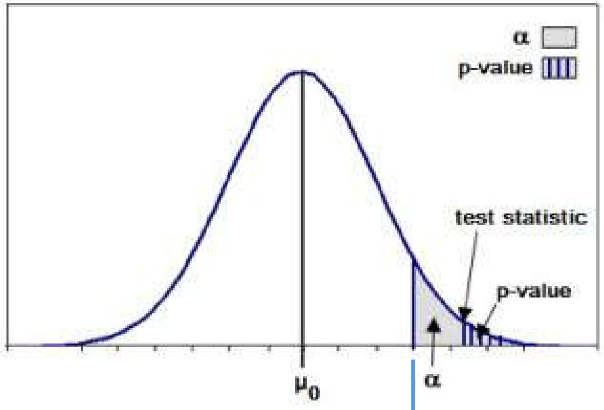
p-value

The p-value (p_{obs}), is the probability of obtaining a test result at least as extreme as that observed for H_0, under the assumption that H_0 is true. Small p-values suggest the rejection of H_0.

Interpretation of p-values

- P-value> α : do not reject H_0
- P-value < α : reject H_0

- For an unilateral test where H_1 involves >



Critical value – statistical table

α=0.05, grey area

1) Consider the test statistics and the table critical value: reject

2) Consider the p-value: reject

Example: consider μ as the average price of a residence by m². Using a sample with n=88, for which the mean and the variance is 1699.656 and 52133.492, test $H_0: \mu \leq 1600$ versus $H_1: \mu > 1600$

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

Tobs=(1699.656-1600)/(52133.492/88)^.5 →4.094
Critical value at (0.05;87) →1.66 Reject H_0

Example: consider the previous example but test $H_0: \mu = 1600$ versus $H_1: \mu \neq 1600$

Critical value at (0.025;87) →1.987: Reject H_0

Example: Bernoulli, large sample

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

In a given region, 1000 individuals were asked about their opinion on the implementation of a project and 53% of them expressed their agreement. Test, for $\alpha = 0.05$, $H_0 : p \leq 0.5$ versus $H_1 : p > 0.5$.

- Zobs=(.53-.5)/(.5*(1-.5)/1000)^.5=1.9
- Critical value at (0.05)=1.645
- H_0 is rejected

Example – proportions equality, large samples.

$H_0 : \pi_1 = \pi_2$

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

Consider $X_1 \sim Ber(\pi_1)$, $X_2 \sim Ber(\pi_2)$, $n_1 = 110$, $n_2 = 100$, $\bar{x}_1 = 0.43$, $\bar{x}_2 = 0.45$. Test

$H_0 : \pi_1 = \pi_2$ against $H_1 : \pi_1 \neq \pi_2$

=(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

$\rightarrow -0.29$

Critical value at (0.025) $\rightarrow -1.96, 1.96$

H_0 is not rejected

Analysis of variance (ANOVA)

Aim: comparing averages of m normal populations

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

Assumptions of ANOVA

- *m* independent samples are available (one for each population), each one with size n_i, with observations $X_{i1}, X_{i2}, \ldots, X_{in_i} \ (i = 1, 2, \ldots, m)$
- *These m* populations follow a normal distribution with unknown means and a common unknown variance,

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

- Test statistics: $F = \dfrac{\text{MS1}}{\text{MS2}} = \dfrac{\text{SS1}/(m-1)}{\text{SS2}/(n-m)} \sim F(m-1, n-m)$,

where $\boldsymbol{SS1} = \sum_{i=1}^{m} \boldsymbol{n_i}(\bar{\boldsymbol{X}}_{i\cdot} - \bar{\boldsymbol{X}}_{\cdot\cdot})^2$, $\bar{\boldsymbol{X}}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} \boldsymbol{X}_{ij}$, $\bar{\boldsymbol{X}}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{m} \boldsymbol{n_i}\bar{\boldsymbol{X}}_{i\cdot}$, $\boldsymbol{SS2} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_{i\cdot})^2$

- Rejection: $F_{obs} > F_{\alpha}$.

- ANOVA table

Source of the variation	Sum of squares	Degrees of freedom	Squared means
Between samples	SS1	$m-1$	$\text{MS1} = \text{SS1}/(m-1)$
Within sample	SS2	$n-m$	$\text{MS2} = \text{SS2}/(n-m)$
Total	SST	$n-1$	

In cases where H_0 is rejected (there are statistically significant differences between the m means) it may be interesting to investigate **for which particular means the differences are significant.**

- Idea: test pairs of means - if a sussession of tests is implemented, their p-values must be corrected to avoid over rejection. One of those corrections is that of Bonferroni (no details are given but implementation in Stata is addressed)

Example: consider the following samples of 3 populations and test the equality of means

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
```

At the 5% significance level, the equality of the 3 means is rejected. However, the equality of means 2 and 3 is not rejected. Note in the test for variance equality, the null hypothesis of equal variances is not rejected

Nonparametric hypothesis tests - Introduction

Aim: testing whether the location (typically the median) for different populations is the same, without assuming a distribution. A major difference relative to previous (parametric) tests (require normality or the presence of large samples), is that tests statistics rely on the ranking of the observations.

- Most well known tests
  - Paired samples
    - *Sign test*
    - *Wilcoxon signed rank* test
  - 2 independent samples
    - Man-Whitney U test
  - *m* independent samples
    - Kruskal-Wallis test

Kruskal-Wallis test

Often considered the nonparamentric version of ANOVA. Tests:

- the mean (or median) equality across *m* populations.
- the distribution functions equality across m populations.

- Idea: for m independent samples (one for each population), $X_{i1}, X_{i2}, \ldots, X_{in_i}$ $(i = 1, 2, \ldots, m)$: construct the observations "rank" for each sample and check whether the rank distribution is similar across the different populations.

Test description

For $n=\sum_{i=1}^{m} n_i$ the rank of observation X_{ij} is given by r_{ij}. Define $S_i=\sum_{j=1}^{n_i} r_{ij}$ (sum of the ranks for sample i). The test statistics is

$$Q=\frac{(n\text{-}1)\times(S_P\text{-}C)}{(S_R-C)}$$

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

The distribution of Q, for large samples, is a qui-squared with m-1 degrees of freedom. For small samples, see the specific table

Example: consider the ANOVA case

```
. 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
```

At the 5% significance level, similarly to ANOVA, the mean equality is rejected.

Tests based on paired samples

- Paired sample of size n: the same individuals are observed before (X_i) and after a programme/treatment/change (Y_i): for example, productivity of firm workers is measured before and after a training programme

- Idea: test whether the median of the difference $Z_i = Y_i - X_i$ is 0, by checking if the probability *p* of a positive Z_i equals that of a negative

$$H_0 : p = 0.5$$

  - *Sign test*
  - *Wilcoxon signed rank* test

- Example 14.8 (Newbold) – Product Preference

An Italian restaurant created a new recipe for the sauce used on its pizza. A random sample of eight students was chosen, and each was asked to rate the the original and the new sauce on a scale 1 to 10. Scores are shown below, with higher numbers indicating a greater liking of the product.

Test whether tastes are different

	Rating		
Student	Original Pizza Sauce	New Pizza Sauce	z=original-new
A	6	8	-2
B	4	9	-5
C	5	4	1
D	8	7	1
E	3	9	-6
F	6	9	-3
G	7	7	0
H	5	9	-4

- *Sign test*

```
. signtest original=new

Sign test

        sign |    observed    expected
-------------+------------------------
    positive |           2         3.5
    negative |           5         3.5
        zero |           1           1
-------------+------------------------
         all |           8           8
…
Two-sided test:
  Ho: median of original - new = 0 vs.
  Ha: median of original - new != 0
      Pr(#positive >= 5 or #negative >= 5) =
      min(1, 2*Binomial(n = 7, x >= 5, p = 0.5)) =  0.4531
```

Do not reject Ho

- *Wilcoxon signed rank* test

```
. signrank original=new

Wilcoxon signed-rank test

        sign |      obs   sum ranks    expected
-------------+---------------------------------
    positive |        2           5        17.5
    negative |        5          30        17.5
        zero |        1           1           1
-------------+---------------------------------
         all |        8          36          36

..
Ho: original = new
             z =  -1.757
    Prob > |z| =   0.0789
```

Do not reject Ho, at the 5% significance level

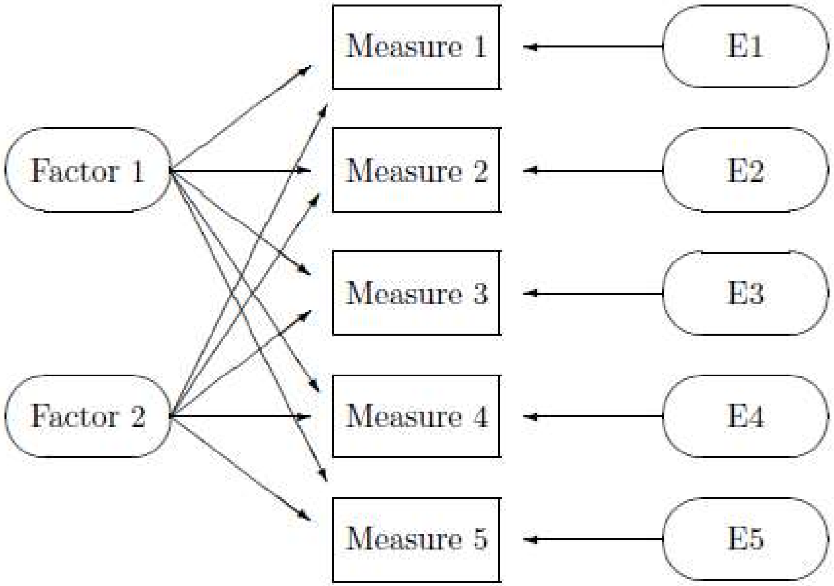
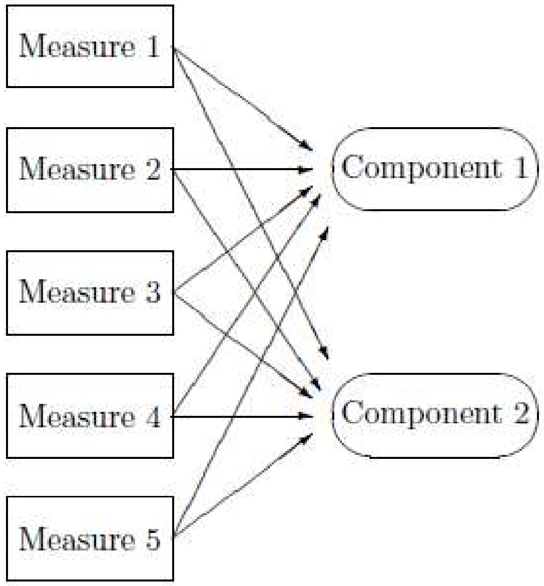
Factor analyis: Introduction

Aim: obtain a set of factors (nonobservable variables) that may explain the initial set of variables. Basically, the information of a set of variables is summarized by a smaller number of latent variables: the factors.

These techniques are based on the correlation matrix of the available variables, which contains the Pearson linear correlation coefficient:

- $r_{yx} = \frac{s_{yx}}{s_y s_x}$, $-1 \leq r_{yx} \leq 1$, where s_x e s_y are standard deviations and s_{xy} is the covariance
- This matrix, naturally contains 1's is the principal diagonal

Factor analysis / Principal component analysis

Factor analysis	Principal component analysis
	

*Structural equations: allow related factors

Factor analysis / Principal component analysis

Factor analysis:

$$X_i = a_{i1}CF_1 + a_{i2}CF_2 + \cdots + a_{im}CF_m + e$$

$$X_i = common\ part + e$$

Principal component analysis

$$PC_i = a_{1i}X_1 + a_{2i}X_2 + \cdots + a_{pi}X_p$$

where

p=#variáveis, m=#factors, i=1,…n, FC=common factor, PC=principal component, e=error

a=coefficients designated as loadings (they are not regression coefficients as the factors are not observed)

Procedure:

1. Obtain the correlation matrix of the variables
2. Extract factors and choose how much of them we wish to keep
3. Factor rotation
4. Interpretation of each factor
5. Possible use of the factor in other analysis (for example, as explanatory variables of a regression model)

Correlation analysis

Kaiser-Meyer-Olken (KMO) index – summarizes the level of correlation between the variables, allowing to check whether the correlations are relevant

- *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*

Factor extraction

Factors may be extracted by the principal component methods or by the method of maximum likelihood, for example

Selecting the # of factors:

- Depend on the proportion of the variance of the original variables that is explained by the factors
  - Typically the software displays the most important factor first, then the second and so on
  - The proportion of variance explained by each factor is the respective eigenvalue (equals the sum of the square of the loadings) divided by p

- One may use the Kaiser criteria: keep the factors for which the eigenvalues >1

Factor rotation

The *loadings* associated to the factor are not unique (there are multiple solutions). Thus, often the factors are rotated, which consists essentially on the imposition of additional restrictions. This makes the interpretation of the factors easier, as the loadings that represent the contribution of each variable to the factor are more extreme:

- Orthogonal rotation: yields independent factors and *loadings bounded by* ±1 (varimax, quartimax, equimax,…)
- Oblique rotation: yields factors that may be correlated (oblimax, quartimin, …)

Factor interpretation

A designation can be issued to each factor by analysing the *loadings*: a higher loading in absolute value, implies a higher contribution of the variable to the factor. By considering the set of variables with higher loadings, a designation may emerge for the factor

Example: Consider a sample of 30 individuals that provided information on the determinants of changing their region of residence. For each question, the degree of agreement is indicated in a 7-level discrete scale (1= total disagreement - 7= total agreement) about the following topics:

V1 = The residence location of the family is important
V2 = A better wage is a major motivation for moving residence
V3 = Good infrastructures and public goods are important (schools, hospitals, roads…)
V4 = Choosing a location where the cost of life is lower is important
V5 = Quality of life is not an important issue
V6 = The major motivation for changing location is career progression

Correlation analysis and KMO index

```
.  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
    -----------------------
```

Several variables display a high correlation. The KMO index is close to a suitable value. Factor analysis will be implemented.

Factor analysis: principal component method

```
. 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
```

- Two factors appear to be relevant (eigenvalues>1)
- The proportion of the variance captured by factor 1 is 45.52% (2.73119/6) and captured by factor 2 is (2.21812/6), in such a way that, together, the two first factors explain 82.49% of the variance of the 6 variables

Factor loadings (pattern matrix) and unique variances

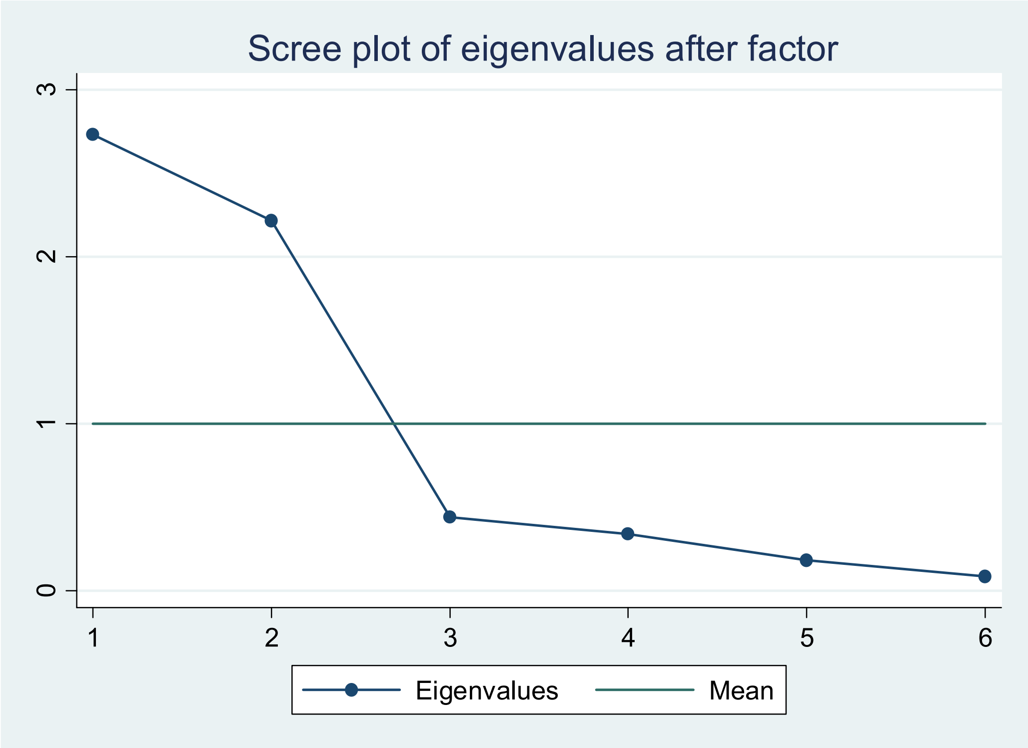
Variable	Factor1	Factor2	Uniqueness
v1	**0.9283**	0.2532	0.0741
v2	-0.3005	**0.7952**	0.2773
v3	**0.9362**	0.1309	0.1064
v4	-0.3416	**0.7890**	0.2609
v5	**-0.8688**	-0.3508	0.1222
v6	-0.1766	**0.8712**	0.2099

- eigenvalues are decomposed by variable: it is possible to identify which variables have a higher contribution to the factor and, thus, interpret the factor. Loadings are high for v1, v3 and v5 for factor 1 and for v2, v4 and v6 for factor 2. Factor 1 appears to summarize quality of life aspects and factor 2 appears to capture professional aspects.

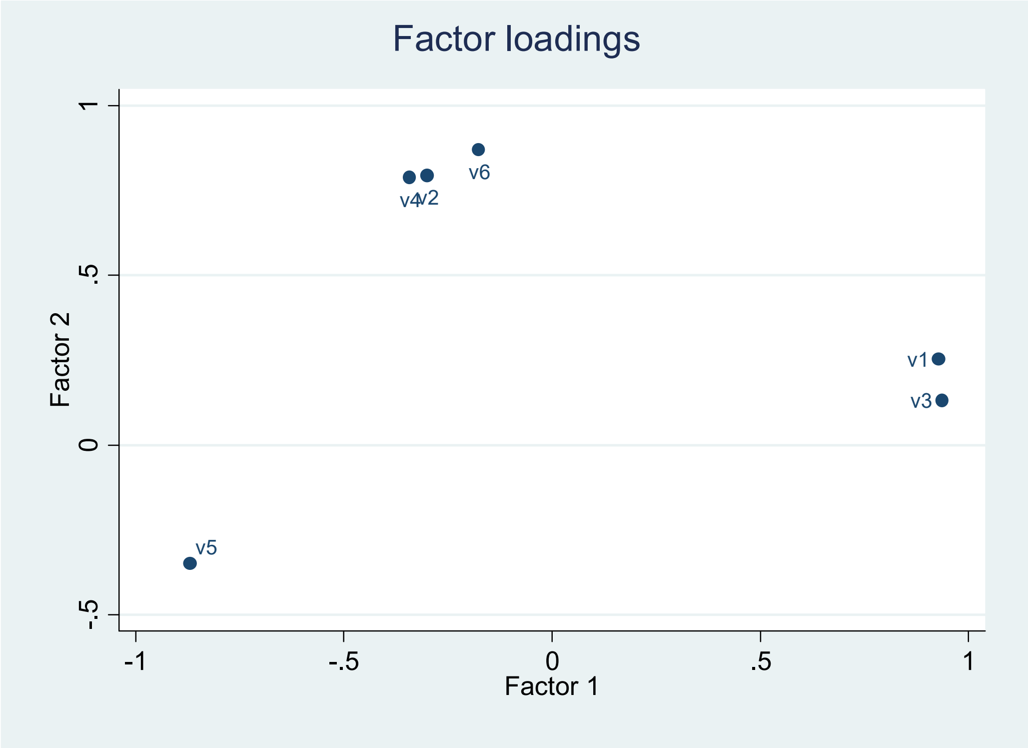
Some calculations:

```
. display (0.9283^2+0.3005^2+0.9362^2+0.3416^2+0.8688^2+0.1766^2)
2.7312031       (eigenvalue of factor 1)
. display (0.9283^2+0.2532^2)
.92585113       (communality: part of the variables explained by the 2 factors)
. display(1-.92585113)
.07414887       (uniqueness: unexplained part of the variable)
```

```
. screeplot, mean
```



```
. loadingplot
```



Factor rotation to extremate loadings and make factor interpretation easier:

```
. 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

- Loadings are now more extreme. The output provides the matrix that allows obtaing a set of loadings from the others
- Other rotation forms are available

Factor generation for each of the individuals (2 additional (latent) variables are added to the database – open the dataset)

```
. 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
```

Factorial analysis: maximum likelihood

```
. 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