Quantitative Data Analysis

Master in Management - Academic Year 2023/2024

Ricardo Moura

ISEG

February 20, 2024

Introduction to Data Analysis

Read Chapters 1,2 and 15, Newbold, 2013

- Definitions and notation
- Types of data: Cross-sectional, Time series, Panel, Pooled data
- Data description techniques: Frequency tables, Histograms, Box-plots, etc.

Definitions and Notation

- **Population** - The entire group of individuals or instances about whom we hope to learn.
- **Sample** - A subset of the population from which we actually collect data.
- Sampling unit/observation/individual - i, where $i = 1, \ldots, n$
- Sample size: n
- Variable: X_j, where $j = 1, \ldots, p$ Observed value of the variable in the i-th individual: x_ij

Types of Data (just some)

- Cross sectional: n observations 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 observation 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
.	.	.	.	.	.
.	.	.	.	.	.

Types of Data (just some)

- 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

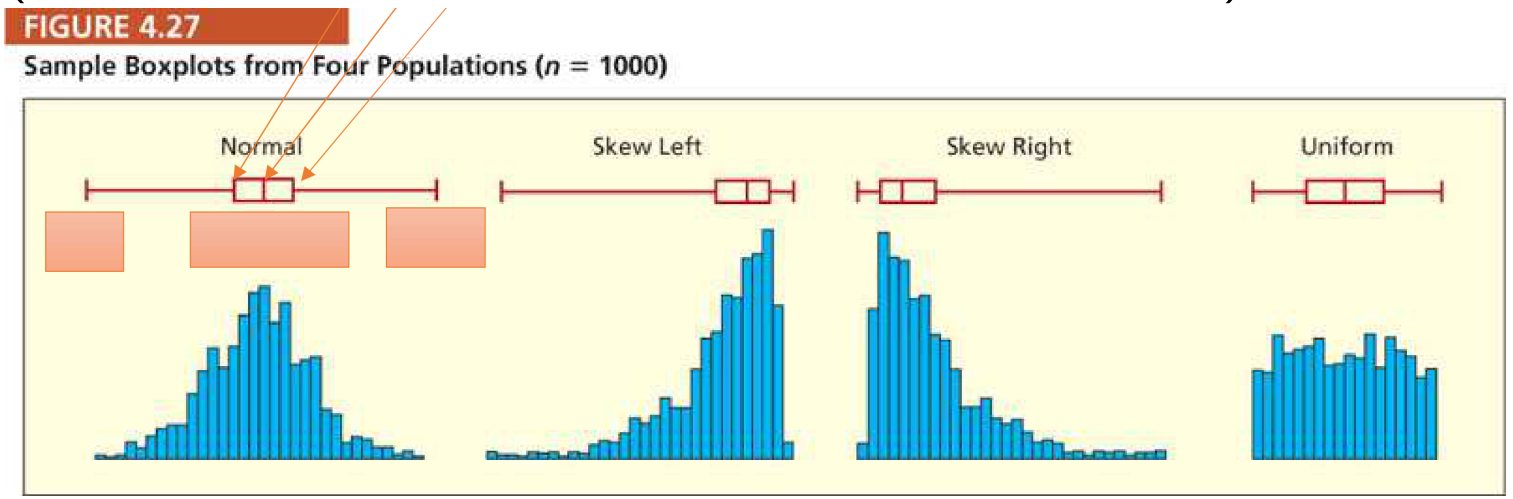
obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85,500	42	1600	3	2.0
2	1993	67,300	36	1440	3	2.5
3	1993	134,000	38	2000	4	2.5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
250	1993	243,600	41	2600	4	3.0
251	1995	65,000	16	1250	2	1.0
252	1995	182,400	20	2200	4	2.0
253	1995	97,500	15	1540	3	2.0
.	.	.	.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

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, ... ex: Q_1, Q_2, Q_3: 25%, 50%, 75% of observations have a smaller value)

For example: histogram and boxplot
(summarizes $x_{min} < Q_1 < Median < Q_3 < x_{max}$)



Data Description

Multiple variables:

- Correlation analysis (of pair of variables)
- Regression analysis: Is wage (**response variable)** explained by a set of (**explanatory**) variables (age, education, etc.)?

Study Correlation(degree of association between two variables):

- Quantitative variables: scatterplots, correlation coefficients
- Qualitative variables: contingency tables

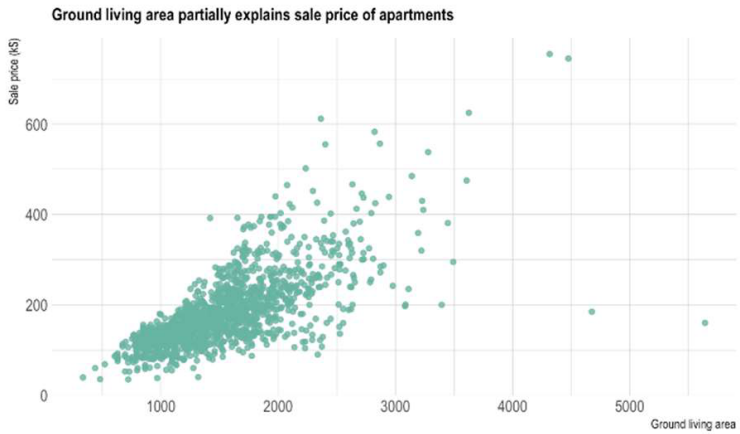


Table 1.2 HEI–2005 Participants' Activity Level (First Interview) by Gender (Component Bar Chart)

	Males	*Females*	*Total*
Sedentary	957	1,226	2,183
Active	340	417	757
Very active	842	678	1,520
Total	2,139	2,321	4,460

Data Description

Pearson Correlation Coefficient: informs on the linear association bewteen two variables Assumptions: linearity between variables; homoscedasticity (uniform variance); normality; independence between pairs of observations, random sample.

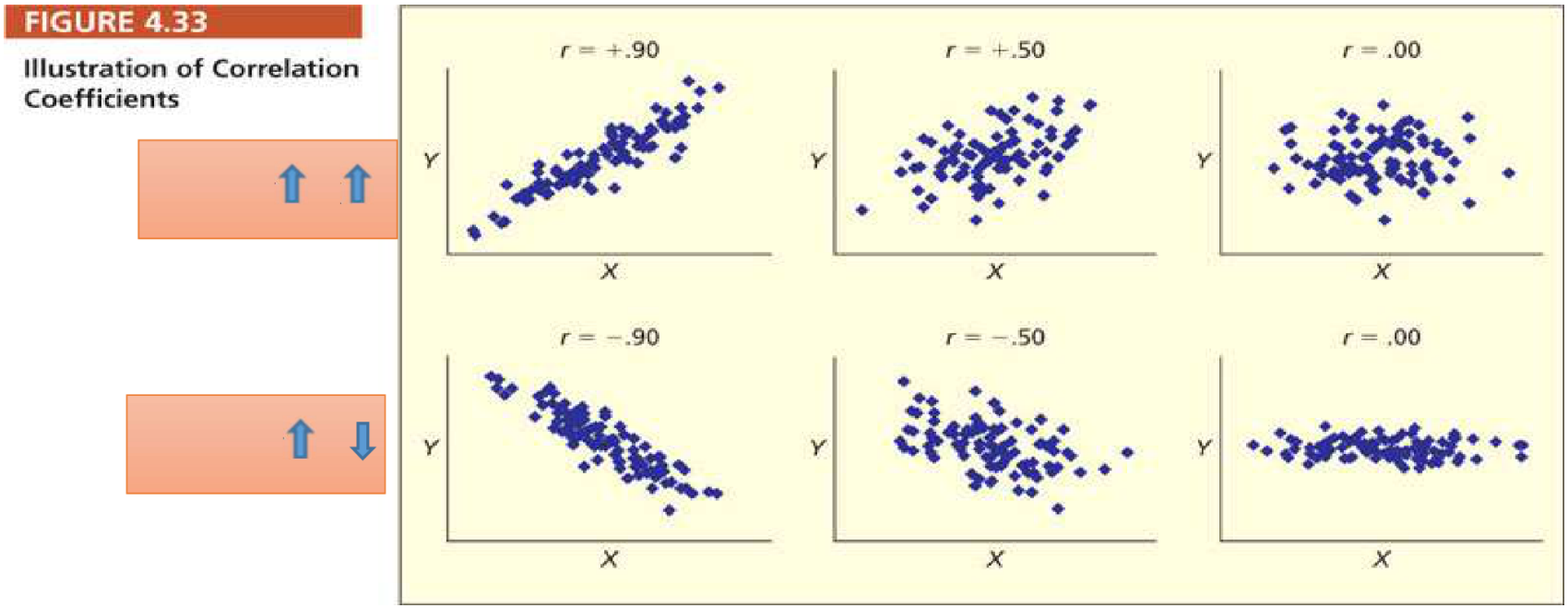
$$r_{yx} = \frac{s_{yx}}{s_y s_x}, -1 \leq r_{yx} \leq 1$$

- s_x and s_y sample standard deviations of x and y, respectively.
- s_{xy} - sample covariance between x and y

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

High linear correlation: $|r_{yx}|$ close to 1
Absence of linear correlation: $r_{yx} = 0$, meaning that either there is no correlation or nonlinear correlation

Data Description



(From Doane and Seward)

Note: As a rule of thumb, you may say that the correlation is significant if $|r_{yx}| > \frac{2}{\sqrt{n}}$ (5% significance level)

Inference

Some parametric and nonparametric inference
The parametric approach considered here relies on the assumption of the normal distribution. This assumption is not required by the nonparametic approach

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

Main interest here: **hypothesis testing**

Hypothesis Testing

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

$$H_0 : \mu = a \quad vs. \quad \mu \neq a$$

$$H_0 : \mu \leq a \quad vs. \quad \mu > a$$

$$H_0 : \mu \geq a \quad vs. \quad \mu < a$$

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

- Define a test statistic
- Define the rejection (critical) region that depends on the significance level (5%, 10% or 1%, are the most usual - that is the probability of rejection the null hypothesis when in fact it is true - Type 1 error)

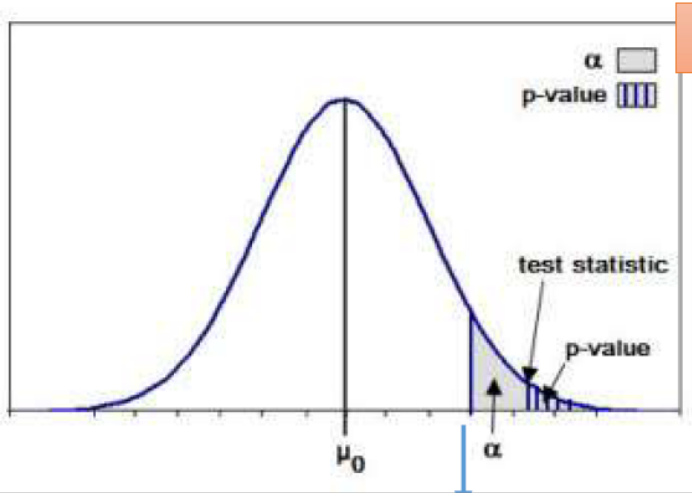
Hypothesis Testing

Using the p-value:

If p-value greater than α, we do not reject H_0

If p-value less than α, we reject H_0

For an unilateral test: $H_0 : \mu \leq a \;\; vs. \;\; \mu > a$



α=0.05, grey area

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

2) Consider the p-value: reject

Testing the equality of p means

Analysis of variance (ANOVA)

Aim: comparing means of p normal populations

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$$

Assumptions of ANOVA

- p independent samples are available (one for each population), each one with size n_j, with theoretic observations $X_{1j}, \ldots, X_{n_j j}$, for $j = 1, \ldots, p$
- these p populations are normally distributed with unknown means μ_j and common unknown variance σ^2.

Hypothesis Testing

Test Statistic: $F = \frac{MS1}{MS2} = \frac{SS1/(p-1))}{SS2/(n-p)} \sim F(p-1, n-p)$

$SS1$ - "Sum of Squares Between Groups" - measures the variance or variability between the groups that we are comparing. It quantifies how much the group means, $\bar{X}_{.j}$, deviate from the overall mean of the data, $\bar{X}_{..}$.

$SS2$ - "Sum of Squares Within Groups" - measures the variance or variability within each group. This sum of squares quantifies how much the individual observations within each group deviate from their respective group mean.

$$SS1 = \sum_{j=1}^{p} n_j (\bar{X}_{.j} - \bar{X}_{..})^2$$

$$SS2 = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{.j})^2$$

Hypothesis Testing

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

ANOVA table:

Source of Variation	**SS**	**Degrees**	**Mean Squares**
Between samples	$SS1$	$p-1$	$MS1 = \frac{SS1}{p-1}$
Within samples	$SS2$	$n-p$	$MS2 = \frac{SS2}{n-p}$
Total	SST	$n-1$	

Table: ANOVA Table

Hypothesis Testing

In cases where H_0 is rejected (indicating statistically significant differences between the p means), it may be interesting to investigate for which particular pairs of means the differences are significant.

Idea: Test pairs of means. If a sequence of tests is implemented, their p-values must be corrected to avoid over-rejection. We use the Bonferroni approach (implementation in Stata is addressed).

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

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

where $p = 3$.

Hypothesis Testing

```
oneway variable population ,Bonferroni                          H0:μ1=μ2=μ3
                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      760.453968      2   380.226984      6.78     0.0080 α=0.05
 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
                                                        H0:σ1=σ2=σ3

                    Comparison of Variable by Population
                                (Bonferroni)
Row Mean-|
Col Mean |          1           2
---------+----------------------
       2 |       14.7  H0:μ1=μ2
         |      0.016
         |
       3 |    14.3429   -.357143
         |      0.015      1.000   H0:μ2=μ3
                H0:μ1=μ3
```

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 that in the test for variance equality, the null hypothesis of equal variances is not rejected

Kruskal-Wallis test

Considered the non-parametric version of ANOVA (Extension of Mann-Whitney U test for more than 2 groups).

Aim: compare the mean (or median, central locations) equality across p populations. (In fact, we are comparing if there is no significant difference among groups)

Idea: for p independent samples (one for each population), $X_{1j}, \ldots, X_{n_j j}$, for $j = 1, \ldots, p$, construct the observations "rank" for each sample and check whether the rank distribution is similar across the different populations.

Example:

Sample A: 3.2, 4.1, 5.5, 3.8, 4.5 [6, 10, 15, 8, 12.5]

Sample B: 2.5, 2.0, 3.1, 2.2, 2.9 [3, 1, 5, 2, 4]

Sample C: 4.0, 4.5, 5.0, 3.5, 4.2 [9, 12.5, 14, 7, 11]

Hypothesis Testing

Kruskal-Wallis test The test is based on the sums of the ranks R_1, R_2, ..., R_p for the p samples.

Test Statistic:

$$W = \frac{12}{n(n+1)} \sum_{j=1}^{p} \frac{R_j^2}{n_j} - 3(n+1)$$

where $n = \sum n_j$; $R_j = \sum_{i=1}^{n_j} r_{ij}$ with r_{ij} as the rank of every observation of the j-th sample. When there are ties, there is a correction for W which is significant when there are an high number of ties.
For large samples (every group has at least 5 observations), under the null hypothesis,

$$W \sim \chi^2_{p-1}.$$

Hypothesis Testing

Kruskal-Wallis test Example:
Sample A: 3.2, 4.1, 5.5, 3.8, 4.5 [6, 10, 15, 8, 12.5]
Sample B: 2.5, 2.0, 3.1, 2.2, 2.9 [3, 1, 5, 2, 4]
Sample C: 4.0, 4.5, 5.0, 3.5, 4.2 [9, 12.5, 14, 7, 11]

$$R_A = 51.5; \quad R_B = 15; \quad R_C = 53.5$$

Thus,

$$w_{obs} = \frac{12}{15 \times 16} \times \left(\frac{51.5^2}{5} + \frac{15^2}{5} + \frac{53.5^2}{5} \right) - 3 \times 16 \approx 9.395 (9.412)$$

Since $P(\chi^2_2 > 9.412) \approx 0.0090$ (command in STATA - di chi2tail(2,9.412)) then for $\alpha = 0.05$ we have enough statistical information to reject the equality of the means.

Kruskal-Wallis test

```
. kwallis y, by(sample)

Kruskal-Wallis equality-of-populations rank test
```

sample	Obs	Rank sum
1	5	51.50
2	5	15.00
3	5	53.50

```
chi2(2) =  9.395
   Prob = 0.0091

chi2(2) with ties =  9.412
             Prob = 0.0090
```