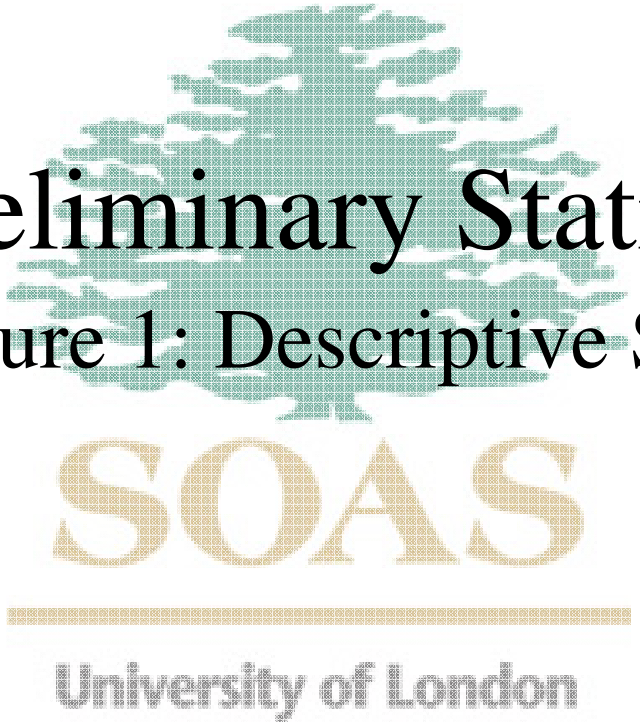
Preliminary Statistics

Lecture 1: Descriptive Statistic



Sophie van Huellen (sv8@soas.ac.uk), September 2013

Organisational

- Sessions:
  - 18.-25. Sep. 2012: 10.00-13.00, V111;
  - 26. Sep. Revision day;
  - 27. Sep. Examination: 10.00-13.00, V111&V211.
- Homework:
  - **DO!** – Will be discussed in the first 30min of the lecture
- Materials:
  - Website
  - Moodle
- Questions:
  - Are not answered individually, but send to the whole group
  - Use email (sv8@soas.ac.uk) or Moodle/website discussion board.
- **NO PHONES!!!**

Outline

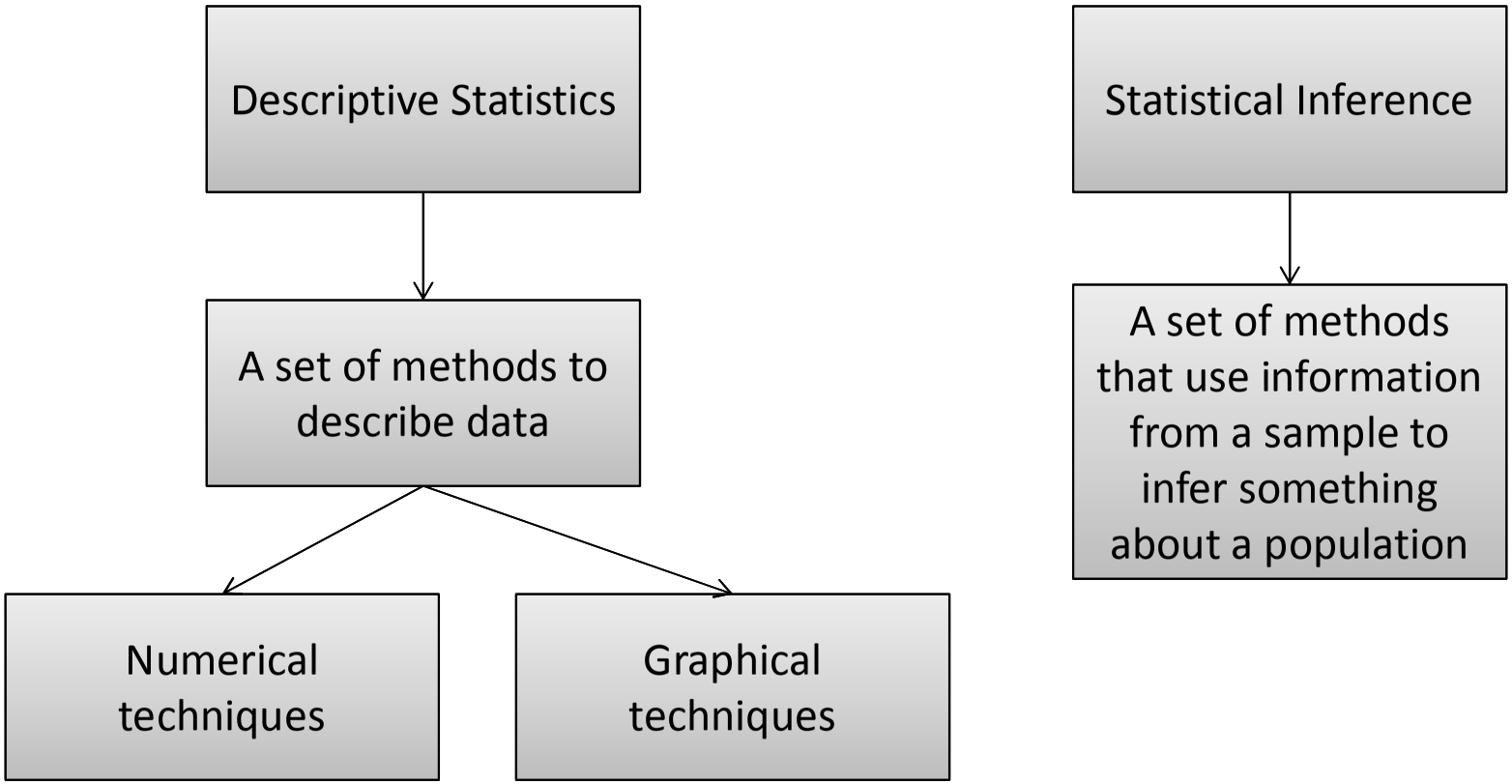
- Basic Concepts
- Summation Operator (digression)
- Descriptive Statistics
  - Numeric Summaries/Summary Statistics
    - Central Tendency
    - Dispersion
    - Shape
    - Measure of Association
    - Standardised Data
  - Graphical Techniques

Basic Concepts

- Origin Of Statistics:
  - The collection of information on the population by the state
- Definition Of Statistics:
  - An arithmetic measure derived from a sample set of data
  - Commonly used as an estimate of a population parameter
- Functions:
  - Description
  - Inference
  - Prediction

Basic Concepts (cont.)

Functions of Statistics



Basic Concepts (cont.)
Functions of Statistics

- Time Series
  - The same data is collected repeatedly over a number of time periods (e.g.: *UK annual inflation*)

- Cross-sectional
  - Data is collected from the elements of the sample at one point in time (e.g.: *Living Standards Measurement Survey SA 1993*)

- Panel/Longitudinal
  - The same data is collected from the same elements over a period of time (e.g.: *British Household Panel Survey*)

Basic Concepts (cont.)
Levels and Quality of Measurement

Characteristic	Nominal	Ordinal	Interval	Ratio
Example	Gender	Preferences	Temperature	Length, age
Distingctiveness	X	X	X	X
Ordered by size	0	x	X	X
Equal intervals	0	0	X	X
Absolute zero	0	0	0	X

Basic Concepts (cont.)
Levels and Quality of Measurement

- Reliability
  - A measurement instrument is reliable if in repeated trials it presents the same measure. The measure may be wrong, but it is the same each time.
  - *E.g.: 2003 GDP 2nd quarter growth underestimated due to incorrect construction figures.*
- Validity
  - A measurement instrument is valid if it measures the concept that is intended.
  - *E.g.: 2001 Employment Survey found and „extra“ 750,000 workers*.

(S.S. Stevens, 1946; S. Briscoe, 2006, (FT Sep 28))

Summation Operator

$$\sum_{i=1}^{n} X_i$$

Summing a Dataset

$$\sum_{i=1}^{5} X_i = X_1 + X_2 + X_3 + X_4 + X_5$$

Index	X_i
1	17,000
2	19,000
3	20,000
4	22,000
5	29,000
SUM	107,000

Summation Operator (cont.)
Common Expression

	Operator	Summation
Multiply the matched pairs X and Y, the sum	$\sum X_i Y_i$	$(X_1 Y_1 + X_2 Y_2 + \cdots + X_n Y_n)$
Square each value of X then sum	$\sum X_i^2$	$(X_1^2 + X_2^2 + \cdots + X_n^2)$
Sum the values of X then square	$\left(\sum X_i\right)^2$	$(X_1 + X_2 + \cdots + X_n)^2$
Double Summation	$\sum_{i=1}^{n} \sum_{j=1}^{m} X_i Y_j$	$\begin{bmatrix} (X_1 Y_1 + X_1 Y_2 + \cdots + X_1 Y_m) \\ +(X_2 Y_1 + \cdots + X_2 Y_m) \\ +(X_n Y_1 + \cdots + X_n Y_m) \end{bmatrix}$

Summation Operator (cont.)

Summation Rules

- 1. Rule
  - The sum of a constant:

$$\sum_{i=1}^{n} a = a + a + \cdots + a = na$$

- 2. Rule
  - The sum of a constant times a variable:

$$\sum_{i=1}^{n} aX_i = aX_1 + aX_2 + \cdots + aX_n = a(X_1 + X_2 + \cdots + X_n) = a\sum_{i=1}^{n} X_i$$

Summation Operator (cont.)
Summation Rules

- 3. Rule
  - Summation is cummutative over addition (but not over multiplication):

$$\sum_{i=1}^{n}(X_i + Y_i) = [(X_1+Y_1) + (X_2 + Y_2) + \cdots + (X_n + Y_n)]$$

$$= [(X_1 + X_2 + \cdots + X_n) + (Y_1 + Y_2 + \cdots + Y_n)] = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i$$

$$\sum_{i=1}^{n}(X_i + Y_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i$$

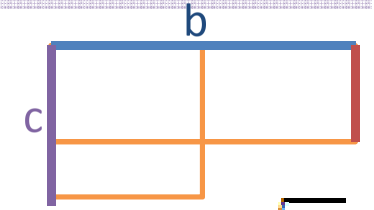
$$\sum_{i=1}^{n} X_i Y_i \neq \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i$$

Numeric Summaries

- Central Tendencies
  - Mean, Median, Mode
- Dispersion
  - Variance, Standard Deviation, Range, Percentiles, Inter-Quartile Range
- Shape
  - Skewness, Kurtosis
- Measure of Association
  - Covariance, Correlation
- Standardised Data

Numeric Summaries (cont.)
Central Tendencies

The Mean	
Arithmetic mean i=1,…,n With n=sample size	$\bar{x} = \frac{\sum x_i}{n}$
Grouped data With C=number of classes f_i=frequency (number of observations in each class)	$\tilde{x} = \frac{\sum_{i=1}^{C} f_i x_i}{\sum_{i=1}^{C} f_i}$
Weighted mean Some data point contribute more than others to the final average. e.g.: mean of grade for two school classes with different size.	$\bar{x}_w = \sum_{i=1}^{n} w_i X_i, with\ w_i = \frac{k_i}{k}$
Geometric mean Normalizes the ranges being averaged. Useful for average of two indicators of different scale.	$\bar{x}_g = \sqrt[n]{x_1 x_2 \ldots x_n}$



$cc = ab, so\ that\ \bar{x}_g = \sqrt{ab} = c$

Numeric Summaries (cont.)
Central Tendencies

	The Median
Definition	The median is the middle value of the distribution (50th percentile) i.e. 50 percent of the values of the variable are below and above the median respectively.
Calculation	1. Put the observations in an ascending/descending order, 2. Find the midpoint observation, 3. Location of Median: (n+1)/2. If odd sample: Median is the value of the middle observation. If even sample: Median is the average value of the two middle observations.
Quartiles	Quartiles are found by dividing the distribution into four parts (same method as for generation of median).

Numeric Summaries (cont.)

Central Tendencies

The Mode

- Is the most frequently occurring value among the entire sample.

Life Expectancy at birth 2003			
Swaziland	32.5	Côte d'Ivoir	45.9
Lesotho	36.3	Tanzania, U. Rep. of	46
Zambia	37.5	Kenya	47.2
Central Afric an Republic	39.3	Burkina Fas o	47.5
Malawi	39.7	Ethiopia	47.6
Angola	40.8	Mali	47.9
Sierra Leon	40.8	Haiti	51.6
Mozambique	41.9	Mauritania	52.7
Congo, Dem Rep. of the	43.1	Djibouti	52.8
Nigeria	43.4	Guinea	53.7
Burundi	43.6	Eritrea	53.8
Chad	43.6	Benin	54
Rwanda	43.9	Madagascar	55.4
Niger	44.4	Gambia	55.7
Guinea-Bissau	44.7	Senegal	55.7
Cameroon	45.8	Yemen	60.6

Numeric Summaries (cont.)

Central Tendencies

Relative Strength

- Mean:
  - Interval/ratio data
  - Sensitive to outliers
  - Useful in further statistics
  - A reasonable measure for symmetrically distributed variables
- Median
  - Ordinal, interval, and ratio variables
  - Robust in terms of the shape of the distribution and outliers
- Mode:
  - Nominal, ordinal, interval, and ratio
  - A dataset can be bi- or multi-modal

Numeric Summaries (cont.)
Measures of Dispersion

- Range
  - Simply the spread of values of the data set
  - Measured by: Maximum – Minimum value
- Inter-Quartile Range
  - Difference between the third and the first quartile
  - Measured by: Upper Quartile – Lower Quartile

Numeric Summaries (cont.)
Measures of Dispersion

- Variance
  - The variance is the average of all squared deviations from the mean:

$$\hat{\sigma}^2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

  - With *n=sample size*
  - Unbiased estimator for the variance:

$$sd^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

Numeric Summaries (cont.)
Measures of Dispersion

- Standard Deviation
  - The standard deviation is given by the square root of the variance:

$$\hat{\sigma} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}} \qquad sd = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

  - With n-1 degree of freedom*
  - The standard deviation is measured in the same units as the data.

*degree of freedom is the number of values in the final calculation of a statistic that are free to vary.

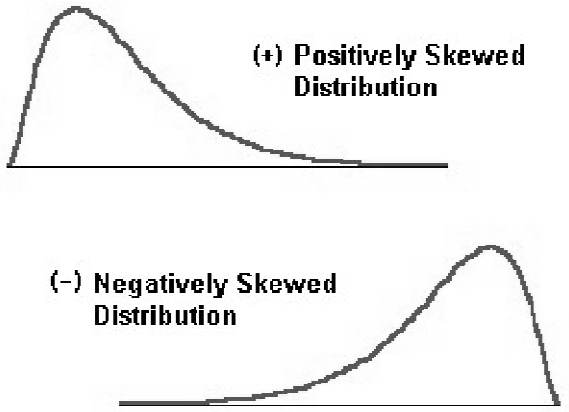
Numeric Summaries (cont.)
Measures of Dispersion

- Coefficient of Variation
  - Measure of relative dispersion
  - Provides a method of comparing the variation of variables measured in different units
  - Expresses the standard deviation as a proportion of the mean: $\frac{\sigma}{\mu}$
  - (the s.d. is about 80% of the mean)

Numeric Summaries (cont.)
Measures of Shape

- Skewness
  - Show how asymmetric a distribution is
  - Coefficient of skewness:

$$S = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{sd}\right)^3$$

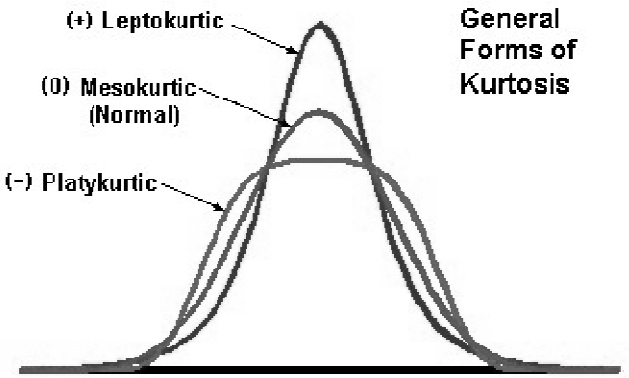


  - Distributions can be:
    - Zero symmetric
    - Positively skewed, skewed to the right (long right tail)
    - Negatively skewed, skewed to the left (long left tail)

Numeric Summaries (cont.)
Measures of Shape

- Kurtosis
  - Measure of peakedness
  - Coefficient of excess kurtosis:

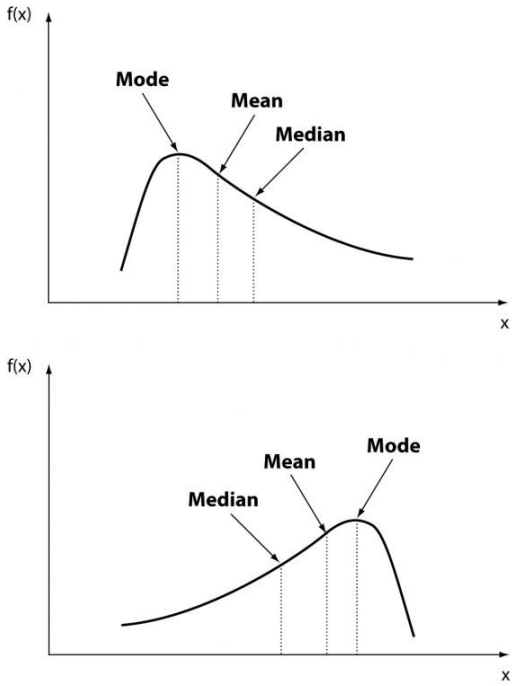
$$K = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{sd}\right)^4 - 3$$



  - Distributions can be:
    - Mesokurtic (zero-normal distribution)
    - Leptokurtic (sharp peak and slim tails)
    - Platykurtic (flat and fat tails)

Numeric Summaries (cont.)
Central Tendency and Shape

- If the data is unimodal* and
  - Right hand skewed
    - Mode ≤ Median ≤ Mean
  - Left hand skewed
    - Mode ≥ Median ≥ Mean
  - If the data is symmetrical
    - Mode = Median = Mean



*There is only one maximum of the distribution function.

Numeric Summaries (cont.)
Measure of Association

- Covariance
  - A measure of how two variables vary together
  - If both variables move in the same direction the covariance will be positive
  - If the variables move in different directions the covariance will be negative
  - The problem is that there is no upper limit; the value of the covariance depends on the units of measurement
  - Formula: $$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \overline{y})}{n}$$

Numeric Summaries (cont.)
Measure of Association

- Correlation
  - Equivalent to the covariance for standardised variables
  - Range: $-1 \leq p \leq 1$
    - *0 = no correlation*
    - *1 = a perfect linear positive correlation*
    - *-1 = a perfect linear negative correlation*
  - Unit free
  - Formular: $$\rho = \frac{cov(x,y)}{sd(x)sd(y)}$$

Numeric Summaries (cont.)
Standardised Data

- Useful transformation of data:
  - Subtract the mean and divide by an estimate for the sd:

$$z_i = \frac{x_i - \bar{x}}{sd}$$

- New variable is called standardised or z-score
- Z-score has mean zero and variance (and sd) one

Numeric Summaries (cont.)

Moments about the Mean

R-th moment	$m_r = \frac{\sum_{i=1}^{N}(X_i - \mu_X)^r}{N}$
1. moment (=0)	$m_1 = \frac{\sum_{i=1}^{N}(X_i - \mu_X)^1}{N}$
2. moment (variance)	$m_2 = \frac{\sum_{i=1}^{N}(X_i - \mu_X)^2}{N}$
3. moment (used for skewness)	$m_3 = \frac{\sum_{i=1}^{N}(X_i - \mu_X)^3}{N}$
4. moment (used for kurtosis)	$m_4 = \frac{\sum_{i=1}^{N}(X_i - \mu_X)^4}{N}$

Graphical Techniques
Time-series data

- Line Graph
  - Evolution of a variable over time
  - Informative about trends, seasonal patterns, cycles, etc
- Histogram
  - For both time-series and cross-section data
  - Proportion (or frequency) of observations falling in different classes/ bins
  - Informative on the shape of the distribution
- Scatter Diagram (or XY plot)
  - Informative about the relationship between two variables
  - Complement to the correlation coefficient

Graphical Techniques (cont.)

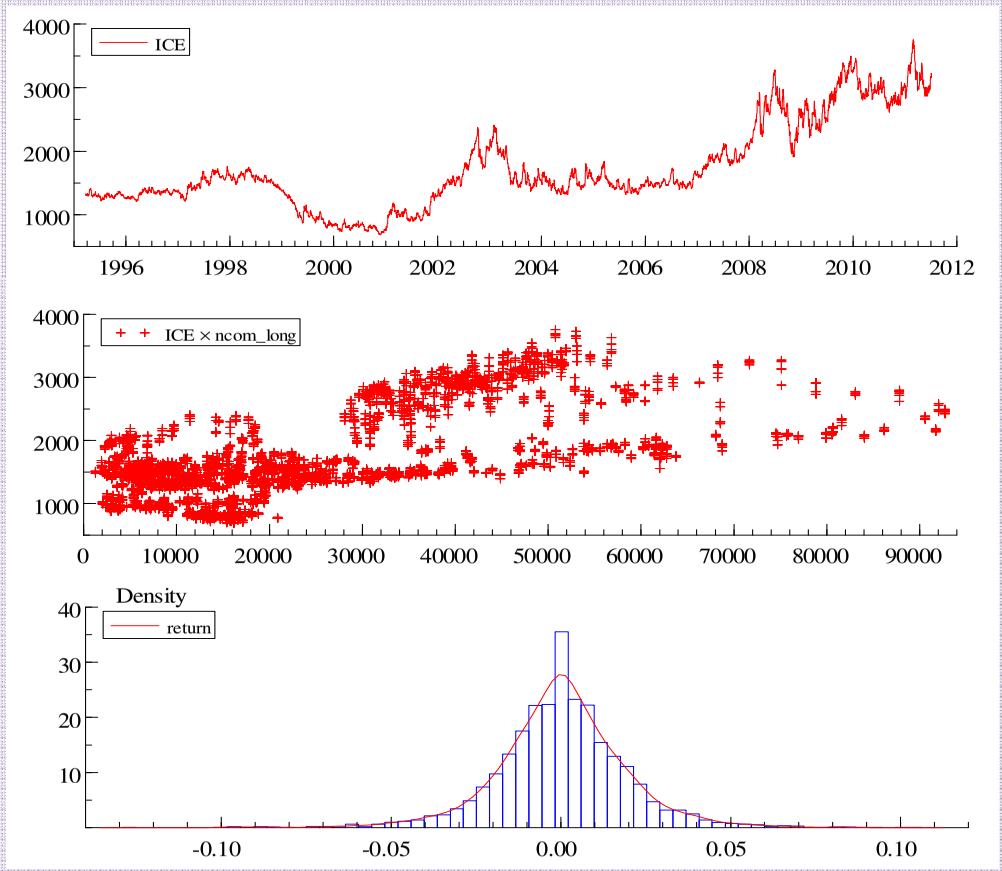
Time-series data

E.g.: ICE cocoa futures price, return & non-commercial traders long positions

Line Graph

Scatter Diagram

Histogram



Graphical Techniques
Cross-section data

- Bar Chart
  - Shows number (frequency) of observations falling in each category
- Histogram
  - Same as bar chart correcting for the width size of each category
- Box Plot
  - Shows the min/max, median, and quartiles on a single diagram