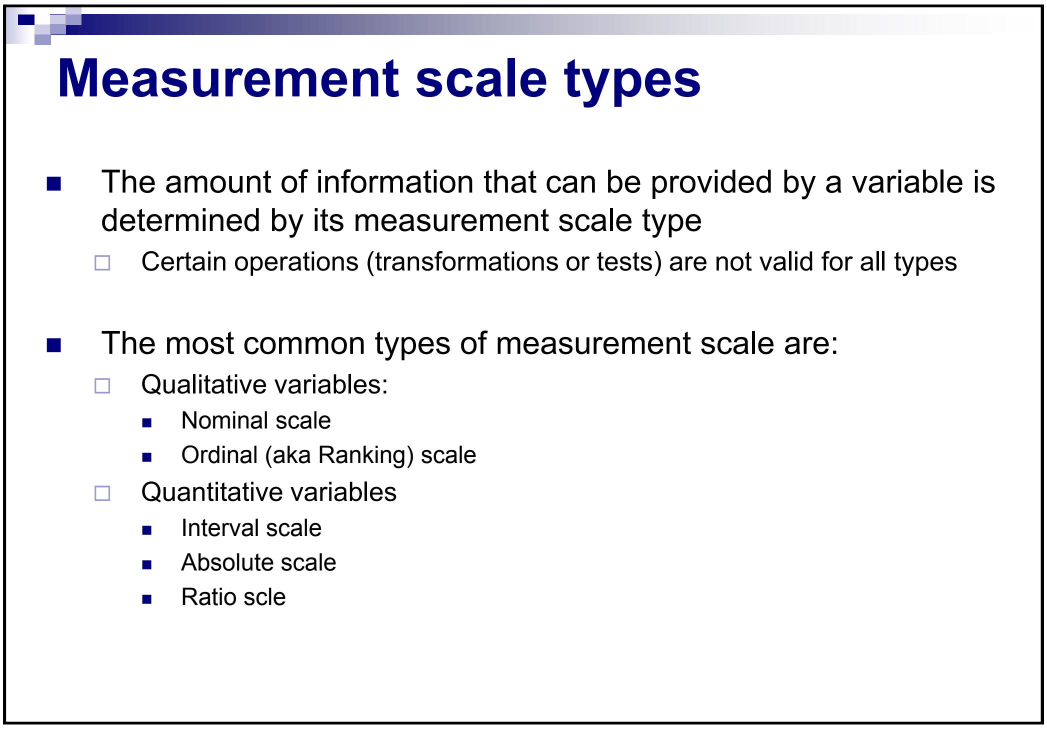
Experimental Research Methodology

– Descriptive Statistics –

Fernando Brito e Abreu (fba@di.fct.unl.pt)
Universidade Nova de Lisboa (http://www.unl.pt)
QUASAR Research Group (http://ctp.di.fct.unl.pt/QUASAR)

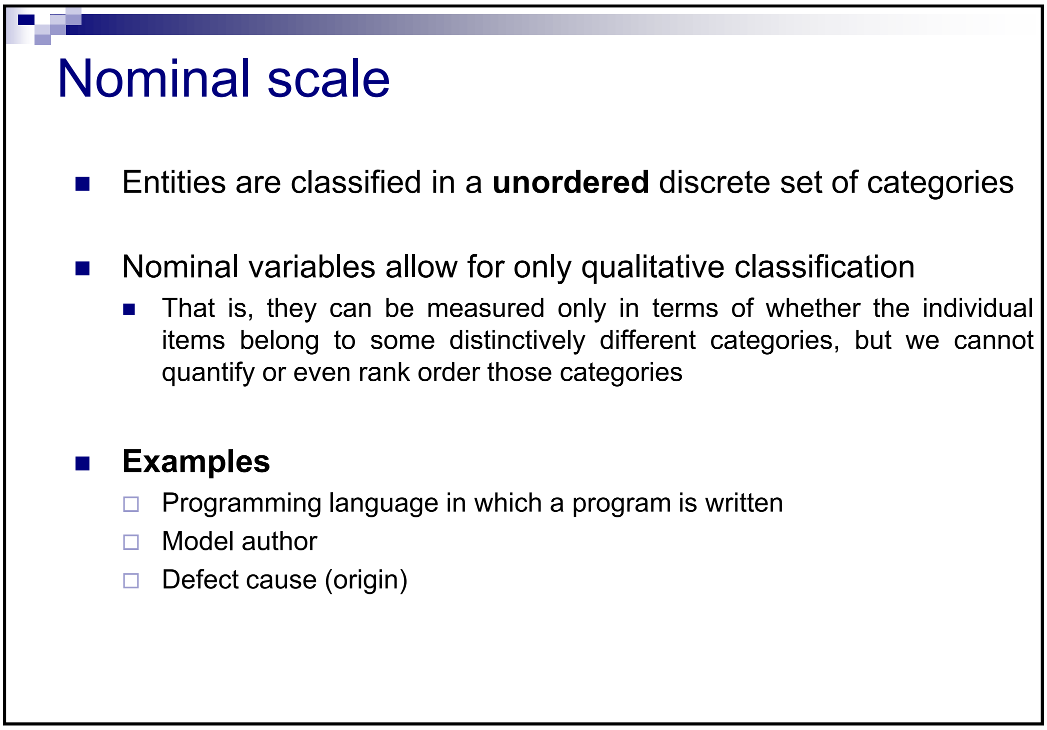
ABSTRACT

- *Scale types*
- *Descriptive statistics*
- *Statistical distribution*
- *Distribution adherence testing*



Measurement scale types

- The amount of information that can be provided by a variable is determined by its measurement scale type
  - Certain operations (transformations or tests) are not valid for all types
- The most common types of measurement scale are:
  - Qualitative variables:
    - Nominal scale
    - Ordinal (aka Ranking) scale
  - Quantitative variables
    - Interval scale
    - Absolute scale
    - Ratio scle



Nominal scale

- Entities are classified in a **unordered** discrete set of categories
- Nominal variables allow for only qualitative classification
  - That is, they can be measured only in terms of whether the individual items belong to some distinctively different categories, but we cannot quantify or even rank order those categories
- **Examples**
  - Programming language in which a program is written
  - Model author
  - Defect cause (origin)

Ordinal scale

- Entities are classified in a **ordered** discrete set of categories
  - Ordinal variables allow us to rank order the items we measure in terms of which has less and which has more of the quality represented by the variable, but still they do not allow us to say "how much more."

- Arithmetic operations are **not** supported for this scale
  - But relational operators are (>, >=, =, !=, <=, <)

- **Examples**
  - Fault impact on operation (Very low, Low, … Very large)
  - Defect removal difficulty (Simple, Normal, Difficult)
  - Execution time categories (Batch, JIT, Interactive, Critical)
  - Scale type (nominal, ordinal, interval)
    - For example, we can say that nominal measurement provides less information than ordinal measurement, but we cannot say "how much less" or how this difference compares to the difference between ordinal and interval scales.

Interval scale

- Values are discrete and ordered like in the ordinal type, but there is a distance among distinct categories
  - The distance between consecutive categories is constant
  - The scale has no absolute zero

- We can **rank the items** measured, as well as quantify and compare the **sizes of differences** between them
  - Example: Instants of system faults occurred in operation
    - A fault occurred on the 5th April is 2 days greater than another on the 3rd April, and the time elapsed between these two instants is half as much as the one between faults occurred in the 10th and 14th April

Absolute scale

- Similar to the interval type but with an absolute zero
  - □ Corresponds to simple counting of entities
- All arithmetic operations are supported

Examples

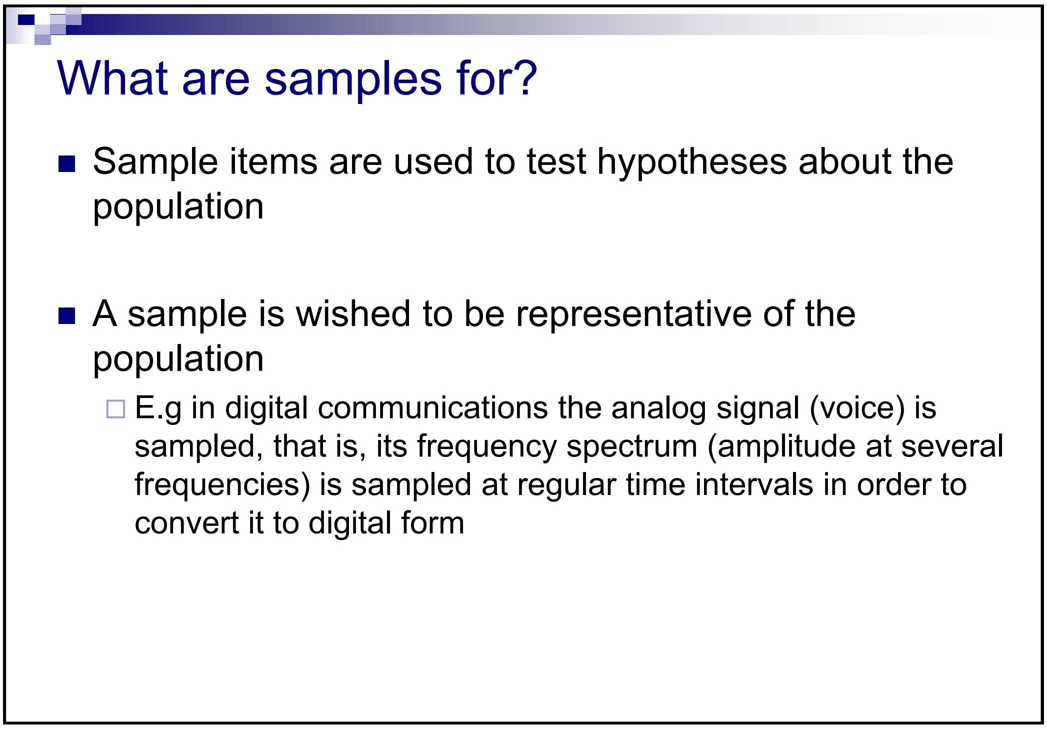
- □ Number of project participants
- □ Number of defects found
- □ Program size (KLOC, number of classes)
  - Example: a program with 800 KLOC is twice as large as one with 400 KLOC

Ratio scale

- They may correspond to:
  - □ continuous properties (e.g. most characteristics of nature)
  - □ ratios among entities (usually absolute)
- All arithmetic operations are supported
  - □ Most statistical data analysis procedures do not distinguish between the interval, absolute and ratio properties of the measurement scales

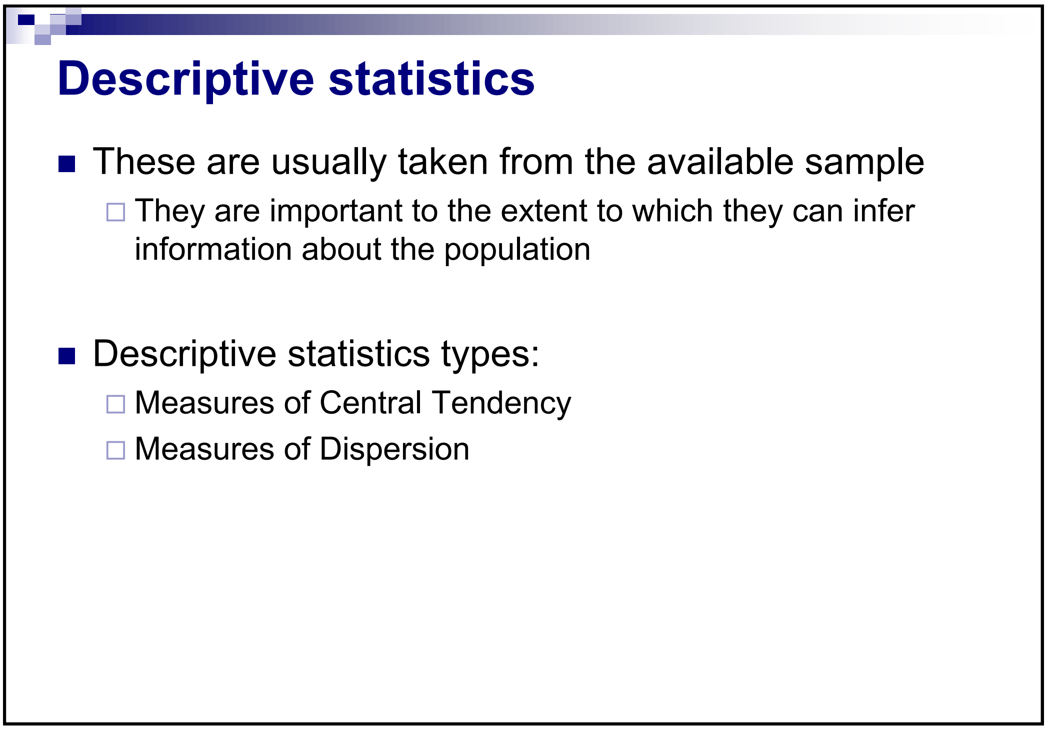
Examples

- □ MTBF, MTTR
- □ Defect density



What are samples for?

- Sample items are used to test hypotheses about the population
- A sample is wished to be representative of the population
  - □ E.g in digital communications the analog signal (voice) is sampled, that is, its frequency spectrum (amplitude at several frequencies) is sampled at regular time intervals in order to convert it to digital form



Descriptive statistics

- These are usually taken from the available sample
  - □ They are important to the extent to which they can infer information about the population
- Descriptive statistics types:
  - □ Measures of Central Tendency
  - □ Measures of Dispersion

Measures of Central Tendency

- **Mode**
  - The most frequent value of a set of values

- **Median**
  - Middle value of an ordered set of values (or the average of the middle two in an even-numbered set)

- **Mean (aka Arithmetic Mean)**
  - An average of n numbers computed by adding some function of the numbers and dividing by some function of n
  - This is probably the most often used descriptive statistic
  - The mean is a particularly informative measure of the "central tendency" of the variable if it is reported along with its **confidence intervals**

- **Geometric mean**
  - This statistic is useful when the measurement scale is not linear; it is computed as:
  - $G = (x_1*x_2*...*x_n)^{1/n}$, where *n* is the sample size.

Confidence Interval for the Mean

- Gives us a range of values around the mean where we expect the "true" (population) mean is located, with a given level of certainty or significance (*p*)

- **Example**
  - Sample mean = 23; *p* = 0.05 confidence interval = [19, 27]
  - **Conclusion**: there is a 95% probability that the population mean is greater than 19 and lower than 27

- The calculation of confidence intervals is based on the assumption that the variable is **normally distributed** in the population
  - The estimate may not be valid if this assumption is not met, unless the sample size is large, say *n*=100 or more

- The width of the confidence interval depends on
  - **Value of p** – smaller *p*-level leads to wider confidence intervals thereby increasing the "certainty" of the estimate, and vice versa
  - **Sample size** – The larger the sample size, the more reliable its mean (smaller interval)
  - **Variation of data values** – The larger the variation, the less reliable the mean (larger interval)

Measures of Dispersion

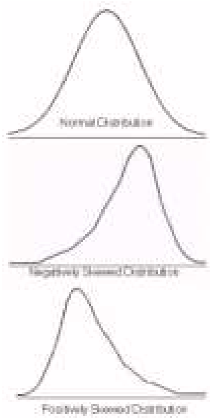
- **Quartile**
  - Definition: any of three points that divide an ordered distribution into four parts, each containing one quarter of the total cases
  - The 2nd quartile is de **median**
- **Decile**
  - Definition: any of nine points that divide an ordered distribution into equal intervals, where each interval contains one-tenth of the total cases
  - The 5th decile is de **median**
- **Percentile**
  - Definition: any of the 99 numbered points that divide an ordered distribution into 100 parts, each of which contains one-hundredth of the total cases
  - The **median** is the 50 percentile

Measures of Dispersion (around the mean)

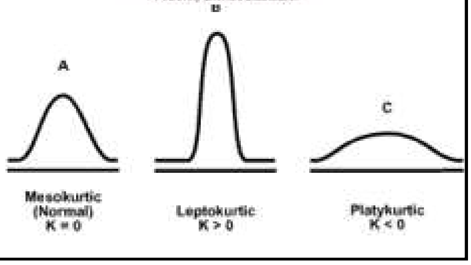
- **Interval of variation (aka range)**
  - Difference between maximum and minimum values of the variable within the considered cases
- **Variance range**
  - Is the sum of squared deviations from the mean divided by one less than the number of cases
  - Is measured in units that are the square of those of the variable itself (the square of the standard deviation)
- **Standard deviation**
  - Is the square root of the variance
  - Is measured in the same units as the variable itself
  - A measure of dispersion around the mean. In a normal distribution, 68% of the cases fall within one standard deviation of the mean and 95% of the cases fall within two standard deviations
    - Example, if the mean age is 45, with a standard deviation of 10, 95% of the cases would be between 25 and 65 in a normal distribution

Measures of Dispersion

- ***Skewness***
  - □ Measures the deviation of the distribution from symmetry



- ***Kurtosis***
  - □ Measures the "peakedness" of the distribution relative to the normal distribution



Statistics for each scale

Scale Type	**Measure of Central Tendency**	**Measure of Dispersion**	**Measure of Dependency (correl. coef.)**
Nominal	Mode	Frequency (as in histograms)	
Ordinal	Median, Percentile	Interval of variation	Spearman, Kendall
Interval & Absolute	Mean	Standard deviation, variance range	Pearson
Ratio	Geometric mean	Coefficient of variation	

Statistical distribution

- The distribution shape tells us the frequency of values from different ranges of the variable

- **distribution function:**
  - □ A distribution function (also known as the probability distribution function) of a continuous random variable X is a mathematical relation that gives for each number x, the probability that the value of X is less than or equal to x.
  - □ For example, a distribution function of height gives, for each possible value of height, the probability that the height is less than or equal to that value.
  - □ For discrete random variables, the distribution function is often given as the probability associated with each possible discrete value of the random variable; for instance, the distribution function for a fair coin is that the probability of heads is 0.5 and the probability of tails is 0.5.

Uniform distribution

- This distribution takes two parameters, a and b (a<=b)
  - □ A uniform variate takes values between these two parameters with equal probability

- The density function is flat between a and b

- **Example**: dice or coin toss result

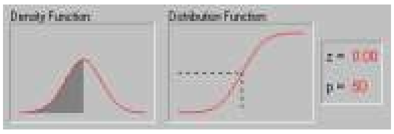
Poisson distribution

Is defined as:

- f(x) = (x * e-)/x!
  for x = 0, 1, 2, ..,   0 <
- Where:
  - (lambda) is the expected value of *x* (the mean)
  - (e) is the base of the natural logarithm, sometimes called Euler's e (2.71...)
- Has an asymmetric curve
- It is often used to represent arrival events
  - e.g. in models encompassing network simulation or queues in general

Normal (Gaussian) distribution

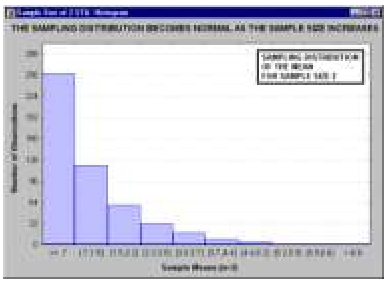
- Normal or Gaussian
  - Continuous symmetric bell curve
  - The distribution is uniquely determined by its mean and standard variance



- A characteristic property of the Normal distribution is that 68% of all of its observations fall within a range of ±1 standard deviation from the mean, and a range of ±2 standard deviations includes 95% of the cases
  - In other words, in a Normal distribution, observations that have a standardized value of less than -2 or more than +2 have a relative frequency of 5% or less
  - Standardized value means that a value is expressed in terms of its difference from the mean, divided by the standard deviation

Normal (Gaussian) distribution

- It has been noted empirically that many measurement variables have distributions that are at least approximately normal
  - Even when a distribution is non-normal, the distribution of the mean of many independent observations from the same distribution becomes arbitrarily close to a normal distribution as the number of observations grows large

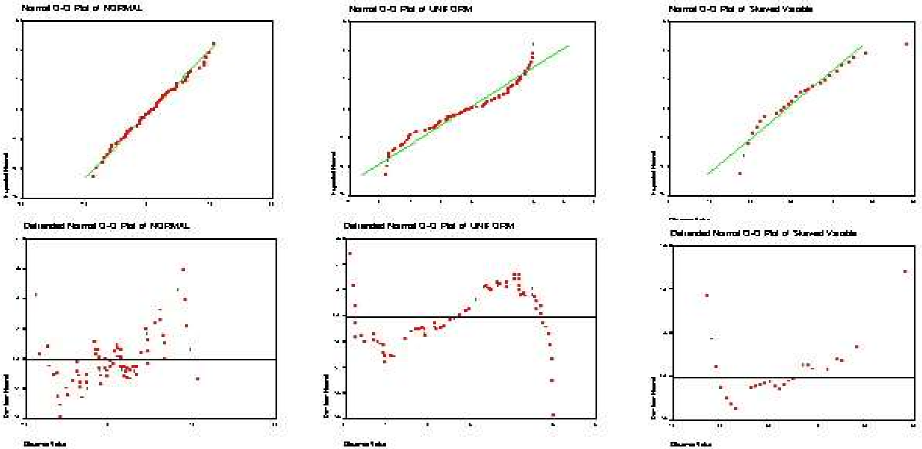


Central limit theorem

- As the sample size (of samples used to create the sampling distribution of the mean) increases, the shape of the sampling distribution **becomes normal**
  - **Note**: for n=30, the shape of that distribution is "almost" perfectly normal

Assessing Normal distribution

- Plots
  - Q-Q (Quantile-Quantile) plots
  - P-P Plots


Normal Q-Q Plot of NORMAL
Normal Q-Q Plot of UNIFORM
Normal Q-Q Plot of Skewed Variable
Detrended Normal Q-Q Plot of NORMAL
Detrended Normal Q-Q Plot of UNIFORM
Detrended Normal Q-Q Plot of Skewed Variable
Expected Normal
Dev from Normal
Observed Value