Data presentation
and descriptive statistics

Paola Grosso
SNE research group

Today with
Jeroen van der Ham
as “special guest”
Instructions for use

• I do talk fast:
  – Ask me to repeat if something is not clear;
  – I made an effort to keep it ‘interesting’,
    but you are the ‘guinea pigs’...feedback is welcome!

• You will not get a grade:
  – But you will have to do some ‘work’;

• 3 for the price of 2
  – We will start slow and accelerate;
  – We will (ambitiously?) cover lots of material;
  – We will also use more than the standard two hours.
Introduction
Why should you pay attention?

We are going to talk about “Data presentation, analysis and basic statistics”.

Your idea is?
Our motivation

1. An essential component of scientific research;
2. A must-have skill (!) of any master student and researcher (... but useful also in commercial/industry/business settings);
3. It will help to communicate more effectively your results (incidentally, it also means higher grades during RPs).

We want to avoid to hear this from you.
How to conduct a scientific project

- Research your topic
- Make a hypothesis.
- Write down your procedure.
  - Control sample
  - Variables
- Assemble your Materials.
- Conduct the experiment.
- Repeat the experiment.
- **Analyze your results.**
- Draw a Conclusion.

This is our main focus!
Roadmap for today and next week

- Collecting data
- Presenting data
- Descriptive statistics
- A real-life example (Jeroen)

- Basic probability theory
- Probability distributions
- Parameter estimation
- Confidence intervals, limits, significance
- Hypothesis testing
Collecting data

Terminology
Sampling
Data types
Basic terminology

- **Population** = the collection of items under investigation
- **Sample** = a representative subset of the population, used in the experiments
- **Variable** = the attribute that varies in each experiment
- **Observation** = the value of a variable during taken during one of the experiments.
Quick test

Estimate the proportion of a population given a sample.

The FNWI has $N$ students:
you interview $n$ students on whether they use public transport to come to the Science Park; $a$ students answer yes.

Can you estimate the number of students who travel by public transport?
The problem of bias
Sampling

• **Non-probability sampling:**
  
  *some elements of the population have no chance of selection, or where the probability of selection can't be accurately determined.*
  
  – Accidental (or convenience) Sampling;
  – Quota Sampling;
  – Purposive Sampling.

• **Probability sampling:**
  
  *every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined.*
  
  – Simple random sample
  – Systematic random sample
  – Stratified random sample
  – Cluster sample
Variables

Qualitative variables, cannot be assigned a numerical value. Quantitative variables, can be assigned a numerical value.

- **Discrete data**
  values are distinct and separate, i.e. they can be counted

- **Categorical data**
  values can be sorted according to category.

- **Nominal data**
  values can be assigned a code in the form of a number, where the numbers are simply labels

- **Ordinal data**
  values can be ranked or have a rating scale attached

- **Continuous data**
  Values may take on any value within a finite or infinite interval
Quick test

Discrete or continuous?

- The number of suitcases lost by an airline.
- The height of apple trees.
- The number of apples produced.
- The number of green M&M's in a bag.
- The time it takes for a hard disk to fail.
- The production of cauliflower by weight.
Presenting the data

Tables
Charts
Graphs
Frequency tables

- A way to summarize data.
- It records how often each value of the variable occurs.

How you build it?
- Identify lower and upper limits
- Number of classes and width
- Segment data in classes
- Each value should fit in one (and no more) than one class: classes are mutually exclusive

<table>
<thead>
<tr>
<th>Friends</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Percentage (%)</th>
<th>Cumulative (less than)</th>
<th>Cumulative (greater than)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-50</td>
<td>6</td>
<td>6/20</td>
<td>30%</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>51-100</td>
<td>4</td>
<td>4/20</td>
<td>20%</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>101-150</td>
<td>2</td>
<td>2/20</td>
<td>10%</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>151-200</td>
<td>4</td>
<td>4/20</td>
<td>20%</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>201-250</td>
<td>1</td>
<td>1/20</td>
<td>5%</td>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td>251-300</td>
<td>3</td>
<td>3/20</td>
<td>15%</td>
<td>20</td>
<td>3</td>
</tr>
</tbody>
</table>
Of course not everybody is a believer:

"As the Chinese say, 1001 words is worth more than a picture"

John McCartey
Histograms

- The graphical representation of a frequency table;
- Summarizes categorical, nominal and ordinal data;
- Display bar vertically or horizontally, where the area is proportional to the frequency of the observations falling into that class.

Useful when dealing with large data sets;
Show outliers and gaps in the data set;
Building an histogram

Add values

Add title (or caption in document)

Add axis legends
Pie charts

Suitable to represent categorical data;
Used to show percentages;
Areas are proportional to value of category.

Caution:
• You should never use a pie chart to show historical data over time;
• Also do not use for the data in the frequency distribution.
Line charts

Are commonly used to show changes in data over time; Can show trends or changes well.

<table>
<thead>
<tr>
<th>Year</th>
<th>RP2 thesis</th>
<th>Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004/2005</td>
<td>9</td>
<td>17</td>
</tr>
<tr>
<td>2005/2006</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>2006/2007</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>2007/2008</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>2008/2009</td>
<td>10</td>
<td>17</td>
</tr>
</tbody>
</table>
Dependent vs. independent variables

• N.b= the terms are used differently in statistics than in mathematics!

• In statistics, the dependent variable is the event studied and expected to change whenever the independent variable is altered.

• The ultimate goal of every research or scientific analysis is to find relations between variables.
Scatter plots

- Displays values for two variables for a set of data;
- The independent variable is plotted on the horizontal axis, the dependent variable on the vertical axis;
- It allows to determine **correlation**
  - **Positive** (bottom left -> top right)
  - **Negative** (top left -> bottom right)
  - **Null**

with a trend line ‘drawn’ on the data.
... and more

- Arrhenius plot
- Bland-Altman plot
- Bode plot
- Recurrence plot
- Nyquist plot
- Lineweaver–Burk plot
- Star plot
- Shmoo plot
- Stemplot
- Forest plot
- Ternary plot
- Galbraith plot
- Nichols plot
- Funnel plot
- Q-Q plot
- Violin plot
Statistics packages followed by some hands on work
Graphics and statistics tools

Plenty of tools to use to plot and do statistical analysis. Just some you could use:

- gnuplot
- ROOT
- Excel

We will use the open-source statistical computer program R. Make installation yourself;

  $> \text{apt-get install r-base-core}

Run R as:

  $> \text{R}

You find the documentation at:

http://www.r-project.org/
Quick exercise

Create a CSV file with frequency data.

Now in R:

```r
> salaries <- read.csv(file="Path-to-file/Salary.csv")
> salaries
> salaries$Salary
> barplot(salaries$Salary)
> dev.copy(png,'MyBarPlot.png')
> dev.off()
```

Can you improve this barplot?
```
help(barplot)
??plot
```
Descriptive statistics

- Median, mean and mode
- Variance and standard deviation
- Basic concepts of distribution
- Correlation
- Linear regression
Median, mean and mode

To estimate the centre of a set of observations, to convey a ‘one-liner’ information about your measurements, you often talk of average. Let’s be precise.

Given a set of measurements:
\[ \{ x_1, x_2, ..., x_N \} \]

- The **median** is the middle number in the ordered data set; below and above the median there is an equal number of observations.

- The (arithmetic) **mean** is the sum of the observations divided by the number of observations. :

\[
\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i
\]

- The **mode** is the most frequently occurring value in the data set.
Look at the (fictitious!) monthly salary distribution of fresh OS3 graduates:

<table>
<thead>
<tr>
<th>OS3 graduates</th>
<th>Monthly salary (gross in €)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grad 1</td>
<td>1250</td>
</tr>
<tr>
<td>Grad 2</td>
<td>2200</td>
</tr>
<tr>
<td>Grad 3</td>
<td>2345</td>
</tr>
<tr>
<td>Grad 4</td>
<td>6700</td>
</tr>
<tr>
<td>Grad 5</td>
<td>15000</td>
</tr>
<tr>
<td>Grad 6</td>
<td>3300</td>
</tr>
<tr>
<td>Grad 7</td>
<td>2230</td>
</tr>
<tr>
<td>Grad 8</td>
<td>1750</td>
</tr>
<tr>
<td>Grad 9</td>
<td>1900</td>
</tr>
<tr>
<td>Grad 10</td>
<td>1750</td>
</tr>
<tr>
<td>Grad 11</td>
<td>2100</td>
</tr>
<tr>
<td>Grad 12</td>
<td>2050</td>
</tr>
</tbody>
</table>

What is median, mean and mode of this data set?

Can you figure out how to do this in R?

What did you learn?
Outliers

- An outlying observation is an observation that is numerically distant from the rest of the data (for example unusually large or small compared to others)

Causes:
- measurement error
- the population has a heavy-tailed distribution
Symmetry and skewness

- A symmetrical distribution has the same number of values above and below the mean which is represented by the peak of the curve.
- The mean and median in a symmetrical distribution are equal.

Outliers create skewed distributions:
- Positively skewed if the outliers are above the mean: the mean is greater than the median and the mode;
- Negatively skewed if the outliers are below the mean: the mean is smaller than the median and the mode.
Dispersion and variability

The mean represents the ‘central tendency’ of the data set. But alone it does not really gives us an idea of how the data is distributed. We want to have indications of the data variability.

• The range is the difference between the highest and lowest values in a set of data. It is the crudest measure of dispersion.
• The variance $V(x)$ of $x$ expresses how much $x$ is liable to vary from its mean value:

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

$$= \frac{\bar{x}^2 - \bar{x}^2}{x^2 - \bar{x}^2}$$

• The standard deviation is the square root of the variance:

$$s_x = \sqrt{V(x)} = \sqrt{\frac{1}{N} \sum_{i} (x_i - \bar{x})^2} = \sqrt{\bar{x}^2 - \bar{x}^2}$$
Different definitions of the Standard Deviation

\[ s_x = \sqrt{\frac{1}{N} \sum_{i} (x - \bar{x})^2} \] is the S.D. of the data sample

- Presumably our data was taken from a parent distribution which has mean \( \mu \) and S.F. \( \sigma \)

\( \bar{x} \) – mean of our sample

\( s \) – S.D. of our sample

\( \mu \) – mean of our parent dist

\( \sigma \) – S.D. of our parent dist

Beware Notational Confusion!

Sep.06 2010 - Slide 34
Different definitions of the Standard Deviation

- Which definition of $\sigma$ you use, $s_{\text{data}}$ or $\sigma_{\text{parent}}$, is matter of preference, but be clear which one you mean!

- In addition, you can get an unbiased estimate of $\sigma_{\text{parent}}$ from a given data sample using

$$
\hat{\sigma}_{\text{parent}} = \sqrt{\frac{1}{N-1} \sum_{i} (x - \bar{x})^2} = s_{\text{data}} \sqrt{\frac{N}{N-1}}
$$

\[ s_{\text{data}} = \sqrt{\frac{1}{N} \sum_{i} (x - \bar{x})^2} \]
Quartiles and percentiles

**Quartiles:**
$Q_1$, $Q_2$ and $Q_3$ divide the sample of observations into four groups:
- 25% of data points $\leq Q_1$;
- 50% of data points $\leq Q_2$; ($Q_2$ is the median);
- 75% of data points $\leq Q_3$.

The semi-inter-quartile range ($SIQR$), or quartile deviation, is:

$$SIQR = \frac{Q_3 - Q_1}{2}$$

The **5-number summary**: (min_value, $Q_1$, $Q_2$, $Q_3$ and max_value)

**Percentiles:**
The values that divide the data sample in 100 equal parts.
Box and whisker plot

It uses the 5-number summary.
Correlation and regression
Correlation offers a predictive relationship that can be exploited in practice; it determines the extent to which values of the two variables are "proportional" to each other.

Proportional means linearly related; that is, the correlation is high if it can be "summarized" by a straight line (sloped upwards or downwards);

This line is called the regression line or least squares line, because it is determined such that the sum of the squared distances of all the data points from the line is the lowest possible.
Covariance and Pearson’s correlation factor

• Given 2 variables \( x, y \) and a dataset consisting of pairs of numbers:

\[
\{(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N)\}
\]

Dependencies between \( x \) and \( y \) are described by the sample covariance:

\[
\text{cov}(x, y) = \frac{1}{N} \sum_{i} (x_i - \bar{x})(y_i - \bar{y})
\]

\[
= (x - \bar{x})(y - \bar{y})
\]

\[
= xy - \bar{x} \bar{y}
\]

(has dimension \( D(x)D(y) \))

The sample correlation coefficient is defined as:

\[
r = \frac{\text{cov}(x, y)}{s_x s_y} \in [-1, +1] \quad \text{(is dimensionless)}
\]
Visualization of correlation

$r = 0$

$r = 0.1$

$r = 0.5$

$r = -0.7$

$r = -0.9$

$r = 0.99$
Correlation & covariance in >2 variables

- Concept of covariance, correlation is easily extended to arbitrary number of variables

\[
\text{cov}(x_{(i)}, x_{(j)}) = x_{(i)} x_{(j)} - \bar{x}_{(i)} \bar{x}_{(j)}
\]

- so that \( V_{ij} = \text{cov}(x_{(i)}, x_{(j)}) \) takes the form of a \( n \times n \) symmetric matrix

- This is called the *covariance matrix*, or *error matrix*

- Similarly the correlation matrix becomes

\[
\rho_{ij} = \frac{\text{cov}(x_{(i)}, x_{(j)})}{\sigma_{(i)} \sigma_{(j)}}
\]

\[
V_{ij} = \rho_{ij} \sigma_i \sigma_j
\]
Quick test

Create a CSV file with frequency data. Read the file into the R memory in variable `obesity`. Run the following commands:

```
attach(obesity)
plot(Weight, Food_consumption)
cor(Weight, Food_consumption)
cor(obesity)
cor.test(Weight, Food_consumption)
```

What have you learned?
Careful with correlation coefficients!

- Correlation does not imply cause
- Correlation is a measure of linear relation only
- Misleading influence of a third variable
- Spurious correlation of a part with the whole
- Combination of unlike population
- Inference to an unlike population
Least-square regression

The goal is to fit a line to \((x_i, y_i)\):

\[ y_i = a + bx_i + \varepsilon_i \]

such that the vertical *distances* \(\varepsilon_i\) (the error on \(y_i\)) are minimized.

The resulting equation and coefficients are:

\[
\hat{y} = a + bx
\]

\[
b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\text{cov}(x,y)}{s_x^2} = \frac{r s_y}{s_x}
\]

\[
a = \bar{y} - b\bar{x}
\]

Note, the correlation coefficient here.
Quick test

From the example before in R:

```r
> pairs(obesity)
> fit <- lm(Food_Consumption~Weight)
> fit
> summary(fit)
> plot(Weight,Food_consumption,pch=16)
> abline(lm(Food_consumption~Weight),col='red')
```
Jeroen van der Ham:
An end-to-end statistical analysis
See you next week...