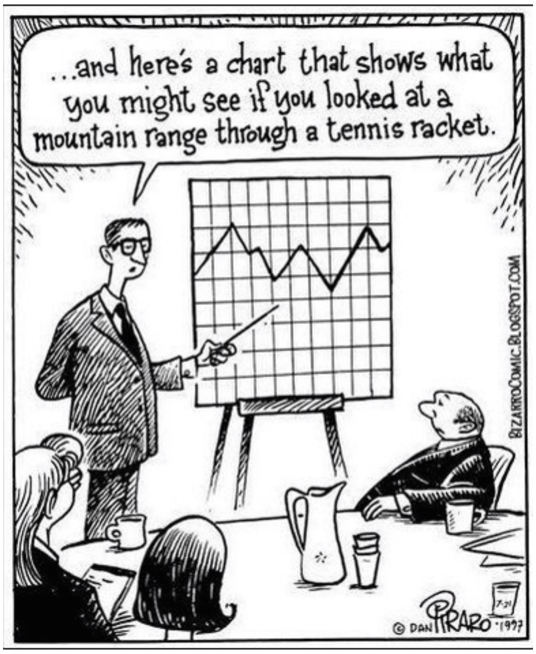
Quantitative Tools - Level 1 - Fall 2015
Louis de Charsonville

Lecture 4

Categorical Variables, Charts and Graphs


...and here's a chart that shows what you might see if you looked at a mountain range through a tennis racket.
BIZARROCOMIC.BLOGSPOT.COM
© DAN PIRARO 1997

1 Reminder

Definition 1.1. Qualitative variable
A qualitative variable (or categorical variable) is a non-numerical variable that categorizes or describes an element of a population. Example : names, hair color, etc. A categorical variable can be :

- ordinal : it can be ordered or ranked.
- nominal : it cannot be ordered

Describing qualitative variables : the most common way to describe a qualitative variable is to count the number of occurrences of each category.

2 Basic graphs

The most common graphs are the following :

- a bar chart (or bar plot, column bar char) is a graph with rectangular bars with lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. The bars can also be stacked.

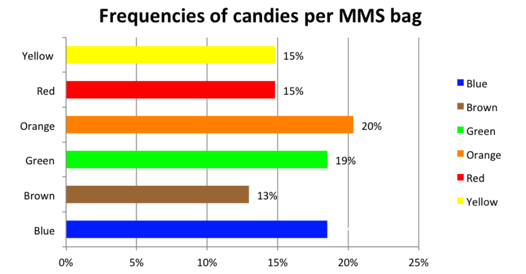


Figure 1: Bar Plot of MMs Frequencies

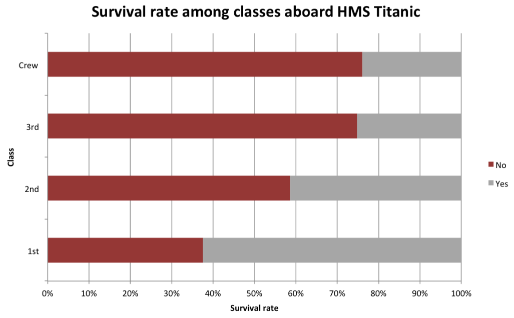


Figure 2: Was it worth paying a first class ticket

- a pie chart : area of the slices are proportional to the relative frequencies.

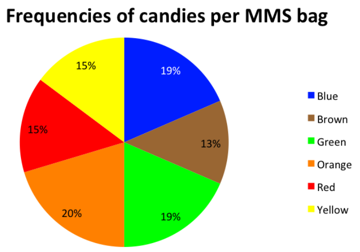


Figure 3: Pie Chart of MMs Frequencies

- a line chart : a chart that display ordered observations linked by a line

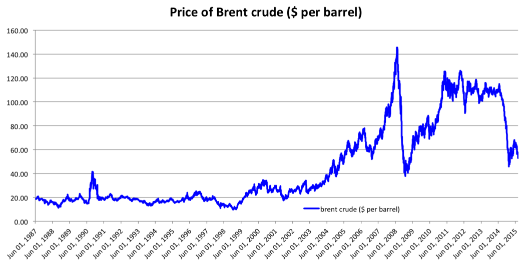


Figure 4: Price of Crude oil

- an histogram : a graphical representation of the distribution of numerical data as tabular frequencies, shown as adjacent rectangles, erected over discrete intervals (bins), with an area (and not the height) equal to the frequency of the observations in the interval.

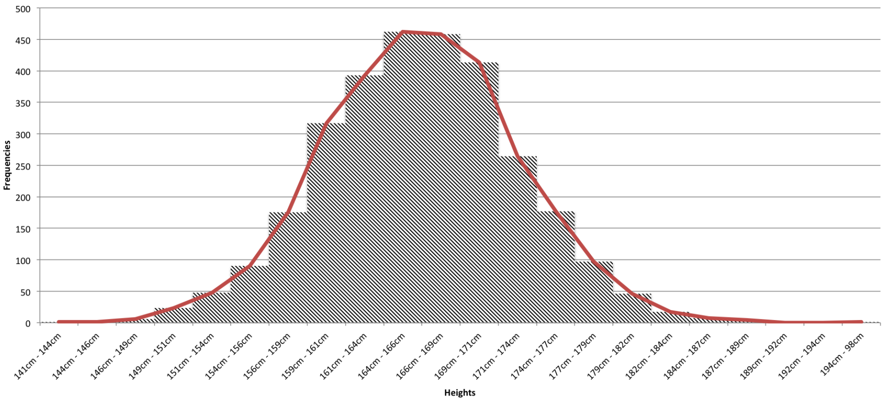


Figure 5: Heights of male criminals over 20 years old undergoing their sentences in the chief prisons of England and Wales in 1902

For **qualitative variables**, the usual graphs are the ones that present frequencies : pie charts and bar plots.

Fallacious graphical representations : Graphical data can show a distorted picture of "real data". Here are a few examples of common misrepresentations

- 3D representations are often misleading due to : (i) perspective which distort dimensions, (ii) the rule chosen for representing frequencies (imagine a graph representing the annual consumption of water in liter of two families, one family is consuming twice than the other, in 2D, the height of the bar for the second family will be twice as big as the bar for the first family. In 3D, it is less clear : do we choose to double the height, the depth, both ?

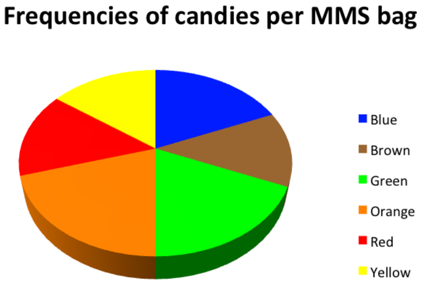


Figure 6: 3D piechart (the blue slice has the same size as the green slice)

- The y-axis should start on zero for a barplots (zero as a baseline for line chart is not always a good option).

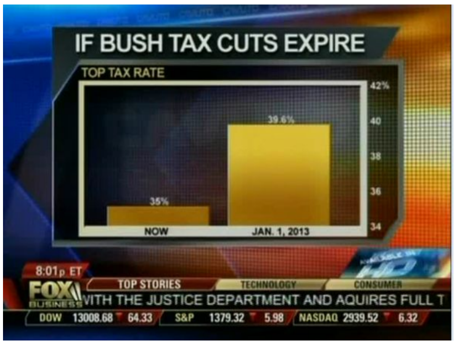


Figure 7: A Drastic Increase In Rate ?

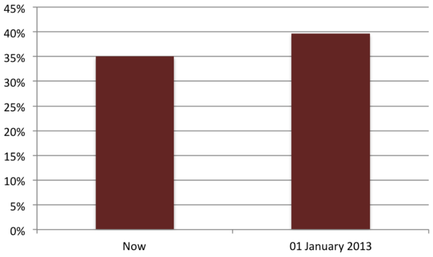


Figure 8: Same graph with a zero baseline for y-axis

Continuous quantitative variables : the distribution of continous quantitative variables is usually represented with an histogram. The entire range of values is divided into a series of intervals. There intervals are

consecutive, adjacent, overlapping and are usually of the same size. A rectangle is erected for each bin, its *area* being proportional to the frequency of cases in the bin. If the bin are the same size, then their height are proportional to the frequency of cases in the bin.
An **ogive** or **cumulative polygon frequency polygon** is a line chart representing the cumulative frequencies.

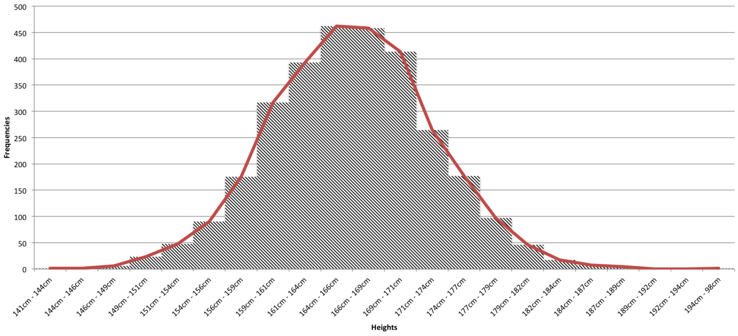


Figure 9: Heights of male criminals over 20 years old

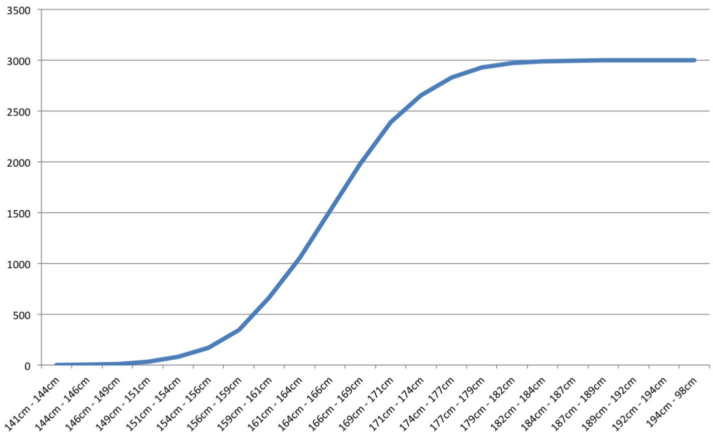


Figure 10: Ogive

Histograms are also used for representing the age structure of a country in a **population pyramid**.

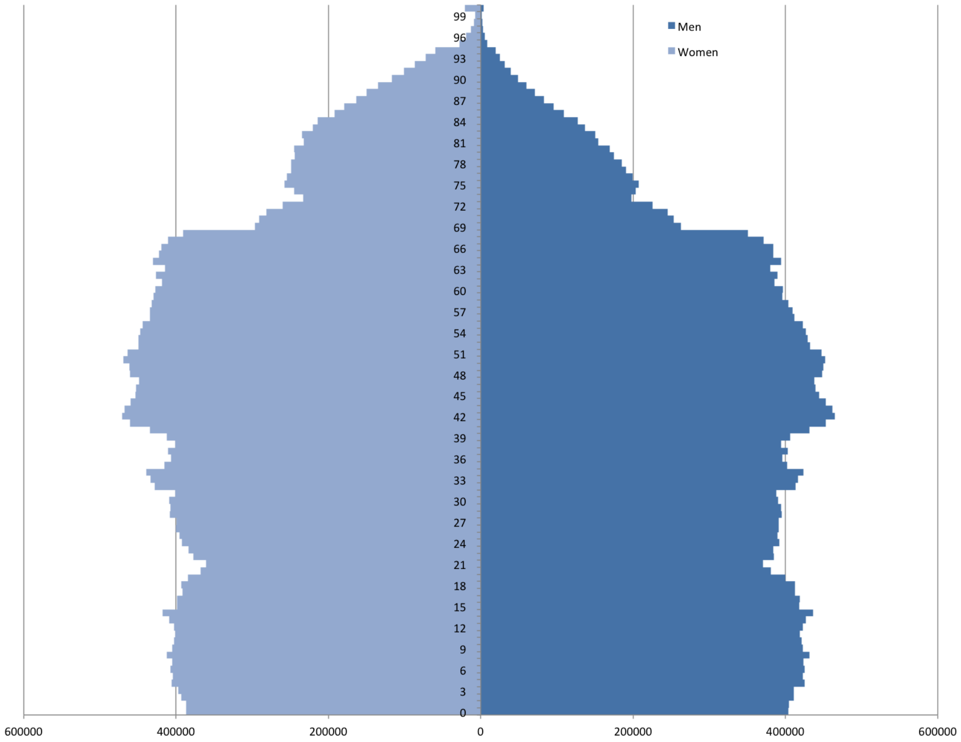


Figure 11: Age Structure in France in 2015

* * *