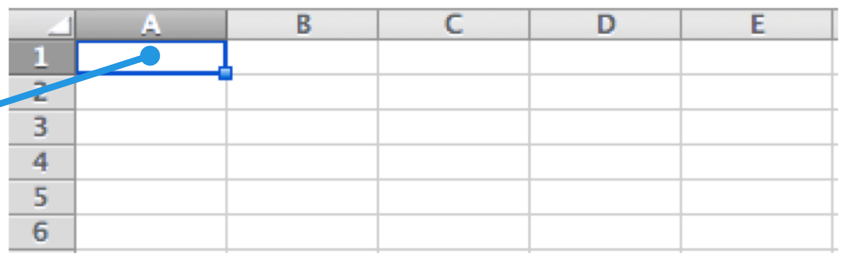
Knowledge and Stats

Knowledge

Data presentation: Spreadsheet

- A spreadsheet is a collection of data orginized as row of cells:

The cell "A1"



- Each cell can contains a value or a "way" to determines its value, a *function*.
- Functions create *relations* between cells.
- Collecting data create *questions* and the problem to find *answers*

Functions with complete knowledge

- The function Max() returns the max value in a set of given values.
- The input set on a spreadsheet it is well defined and clear; we can provide the exact (optimal) solution for the *problem Max*

Functions with incomplete Knowledge

- Sometime on the real world it is not possible to collect the whole data set:
  - Data set too big, ex: *the average age of the world population*.
  - Data set extension unknown because hidden into a too big population: *The number of games owned by Italian owners of a Commodore 64 console.*
  - Lack of time for task execution: *Find the best candidate by deep interview for a job*
- These are problems with *incomplete Knowledge*

The *secretary* problem

- An administrator wants to hire the best secretary out of n rankable applicants for a position.
- The applicants are interviewed one by one in random order.
- During the interview, the administrator can rank the applicant among all applicants interviewed so far, but is unaware of the quality of yet unseen applicants.
- **A decision about each particular applicant is to be made immediately after the interview. Once rejected, an applicant cannot be recalled**.

What is the best stopping strategy?

The *secretary* problem (contd)

- Why the secretary problem is meaningful abstraction for web communications:
- data is flowing, cannot be easily saved, there's non finite domain to refer to.

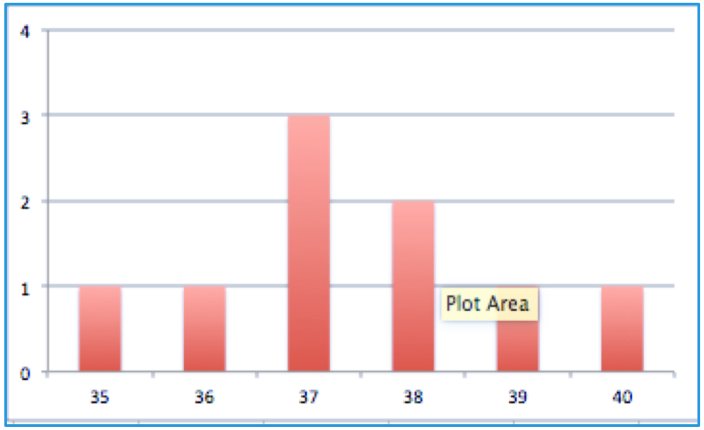
A walk in the garden with Stat and Prob

The source of Knowledge

- A sensor/probe returns one of a finite set of possible values
  - Thermometer: A number into 34.5÷43.5 with step of 0.1.
  - Dice: 1,2,3,4,5,6
  - Political ballot: one of two candidates
- We can repeat measurement variouse times, collecting a set of *observations*, a dataset.
- Analizing observations, we can try to infer some knowledge of the world the data came from.

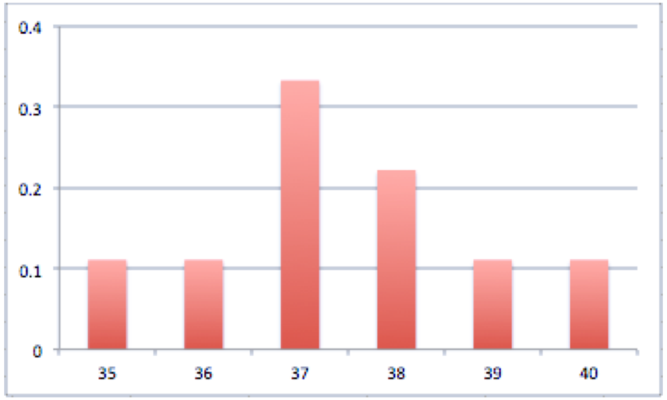
Frequency and frequency histogram

- Frequency: How many times a particular value happened in my observations?
- Frequency histogram: How my frequency are spread among my observations?
  - given this observations:{37,35,36,37,37,38,40,38,39}
  - Fr(35)=1, Fr(36)=1, Fr(37)=3
    Fr(38)=2, Fr(39)=1, Fr(40)=1
  - FrHist(35÷40)={1,1,3,2,1,1}



… toward Knowledge

- Frequency normalization: reformat histogram in order to *hide* the dataset size, and *try to generalize:*
  - given this observations: {37,35,36,37,37,38,40,38,39}
  - #observation = 9
  - NormlizedFr(35)=1/9, NormFr(36)=1/9, Norm Fr(37)=3/9, NormFr(38)=2/9, NormFr(39)=1/9, NormFr(40)=1/9
  - NormalizedFrHist(35÷40)
    ={1/9, 1/9, 3/9, 2/9, 1/9, 1/9}



The important of having multiple observations

- Many observations you made, more your observations are near to the reality (*the law of large numbers*)
- How many observations I have to do? The importance of selecting a good population in which make observations.
- Bias can mistify data:
  - I tend to use thermometer when I'm sick so my average temperature from that observations dont represent my *real* avarege temperature.
  - Usually young people dont reply to the home phone; interviews with this chanel tend to reach more adults.

Average vs Median

- Average: the sum of all values divided by number of observations

  + easy to calculates

  + can be manipulates with a lot of math transformations

  - for low number of observation, it tend to be biased by outliers
- Median: the observation in the middle, i.e. ordering observation by value, it is the observation value who have the same number of observation before and after itself

  + less sensible to outliers respect Average

  - require an ordering step (expensive to calculate)

.. from the little to the big…

- Probability: informally, the number of *good* observable values ratio the number of *possible* observable values.
  - Dice:
    - possible observable values: {1,2,3,4,5,6}
    - Probability of "5": 1/6
  - Coin:
    - possible observable values: {"head","tail"}
    - Probability of "head": ½
- Probability: more formally, a number from 0 (impossible) and 1 (always true) thet express the expected frequency of happening of a particular event.

Exercise

The Probability of seeing a 'six' when throwing two dice:

- possible observable values:
  <1,1>, <1,2>, <1,3>, <1,4>, <1,5>, <1,6>
  <2,1>, <2,2>, <2,3>, <2,4>, <2,5>, <2,6>
  <3,1>, <3,2>, <3,3>, <3,4>, <3,5>, <3,6>
  <4,1>, <4,2>, <4,3>, <4,4>, <4,5>, <4,6>
  <5,1>, <5,2>, <5,3>, <5,4>, <5,5>, <5,6>
  <6,1>, <6,2>, <6,3>, <6,4>, <6,5>, <6,6>
- good observable values:
  <6,1>, <6,2>, <6,3>, <6,4>, <6,5>, <6,6>, <1,6>, <2,6>, <3,6>, <4,6>, <5,6>
- Pr("seeing a 6") = 11/36 ≅ 0.3