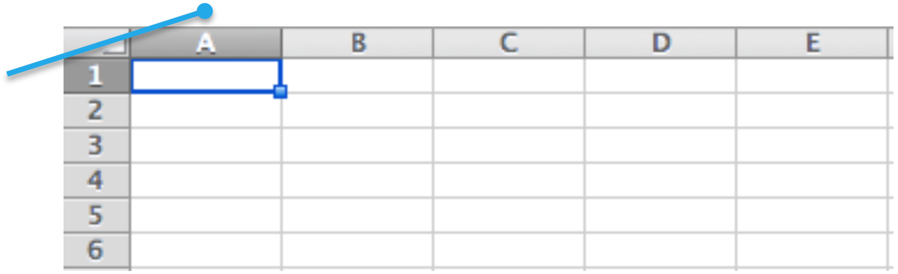
A walk in the park with Probabilites and Stats

it::unimi::sps::webcomm

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.

SP and Psychologhy

- […] people tend to stop searching too soon.
- This may be explained, at least in part, by the cost of evaluating candidates.
- In real world settings, this might suggest that **people do not search enough** whenever they are faced with problems where the decision alternatives are encountered sequentially

Cfr. *https://en.wikipedia.org/wiki/Secretary_problem#Experimental_studies*

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.

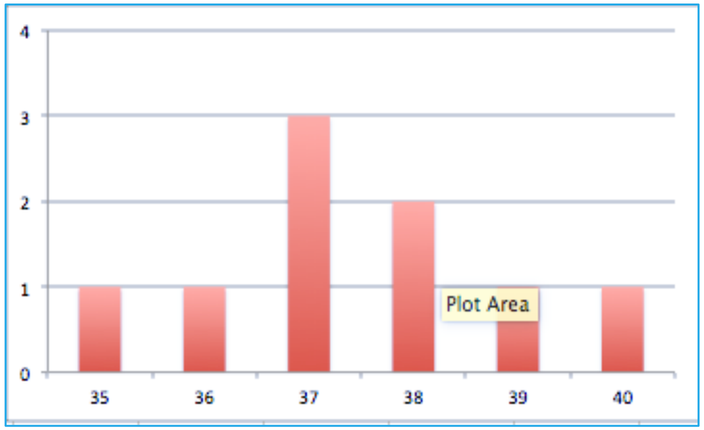
The garden of Probability and Stats

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 various times, collecting a set of *observations*, **a dataset**.
- Analizing observations, we can try to infer some knowledge of the world the data came from.

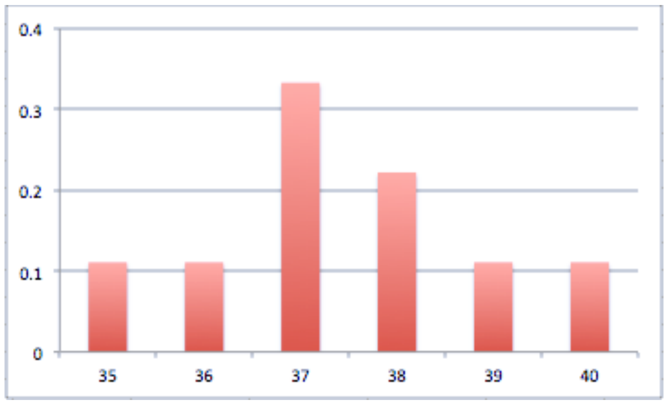
Frequency and frequency histogram

- 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 are needed? The importance of selecting a good population in which make observations.
- **Bias** can deviate 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.
  - What about **"algorithmic bias?"**

Mean vs. Median

- Mean: the simplest average: sum of all values divided by number of observations

  + easy to calculate

  + can be adapted, with math transformations

  - for low # of observations, it tends 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

  - requires an ordering step (expensive to compute)

From small to large: probability

- Informally: ratio of the # of *good* observable values over # of *possible* observable values (*sample space*).
  - 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
- formally, Pr: S → [0..1] (0=impossible, 1=certain) s.t. its integral (sum over S) is 1.

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 $\cong$ 0.3

Epilogue: the 1/e-strategy

- The best know strategy for the secretary problem is "37% rule:"
- Let N be the number of applicants
- Interview the first N/e applicants and fix the threshold score t (e=2.718...)
- Interview the remaining candidates; hire the first whose score > t.
- Pr[X=max] = 1/e = 0.3678…

- What could possibly go wrong???

Final considerations

- The Web is open-domain: hard to fix the sample space (denomitator)
- A phenomenon ('seeing a 6') might have more than one explanation: hard to 'go back' to the original happening
- We try to *maximise the impact of communication* by either
  - Increasing frequencies (numerator)
  - Re-shaping the user base (denominator)
- Better interfaces
- Stastistical tests that allow to estimate impact: **A/B testing**