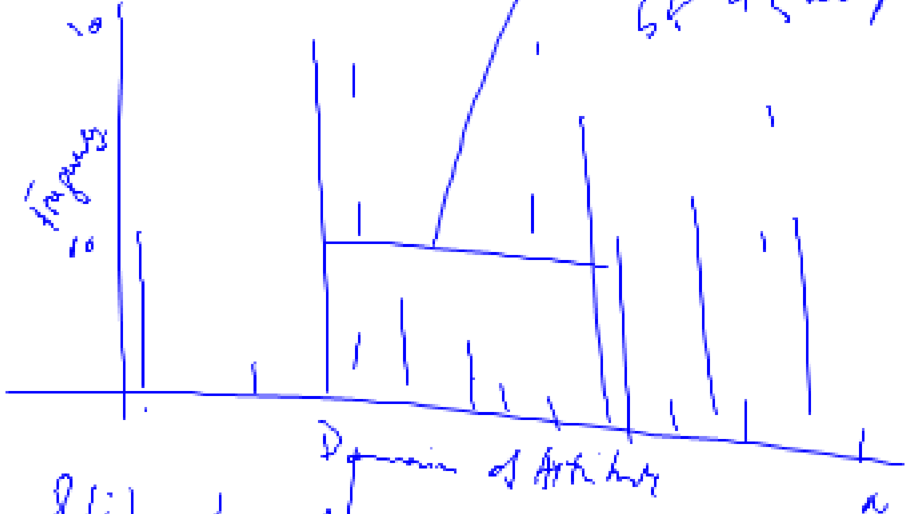
Lecture 23

Histogram

Approx all frequencies by avg. freq.

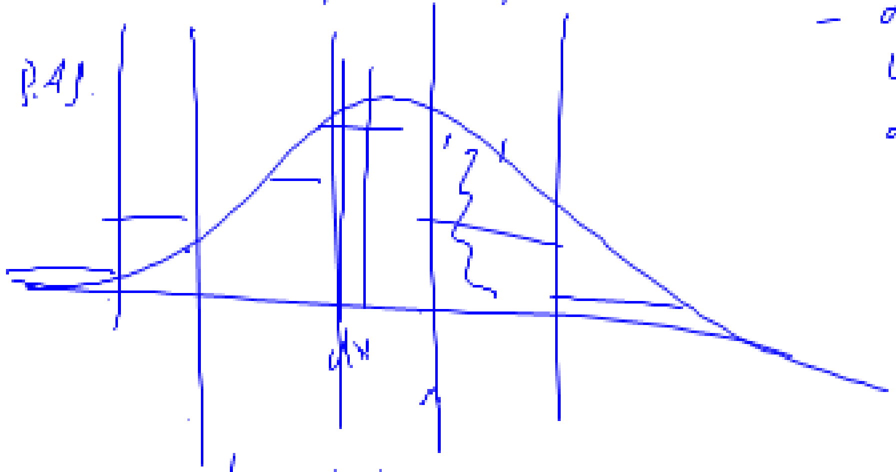


$f(i)$ - true freq.

$\hat{f}(i) = f_{avg,i}$ if $i \in b$

History

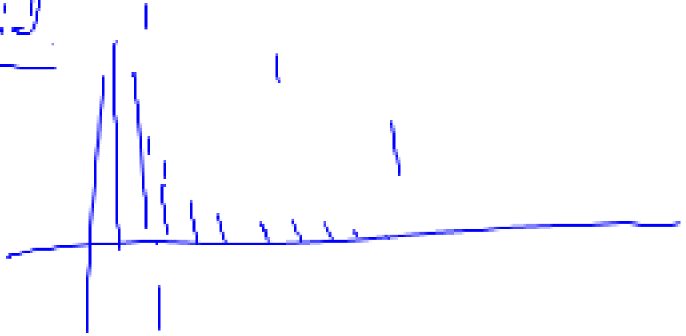
- Statistics
  - as nonparametric approximation of densities
- $N(\mu, \sigma)$



when cdf cannot be computed in required form

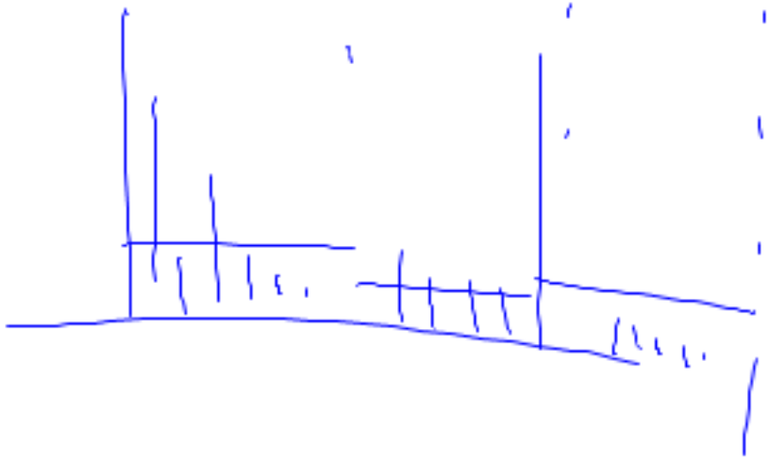
- developing theory to say how good histogram approx is.

Binomial



Why deal with histograms?

- Space



Assumption made by histograms

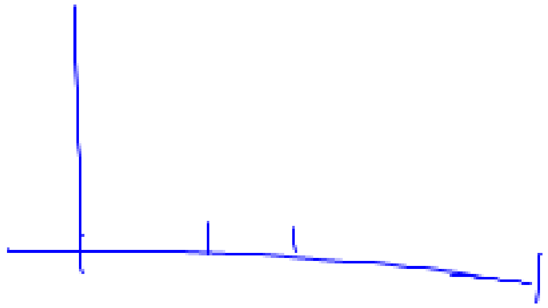
- There is an order on the attribute values
- frequency, in the vicinity, changes smoothly

↑ usually true in statistics

Histograms in statistics are used as function approximations for p.d.f.

Types of histogram

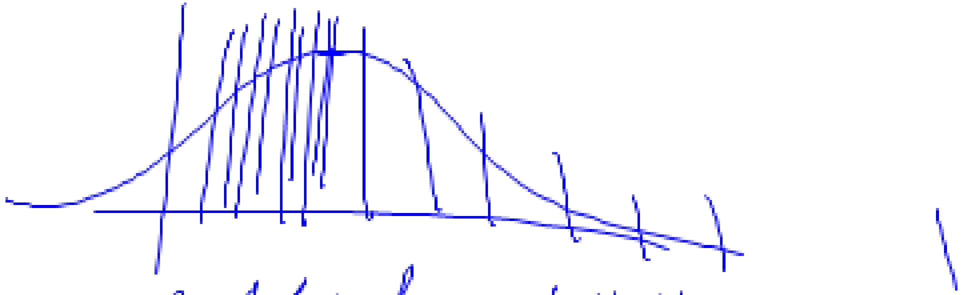
• Equi-width



10 buckets of equal size

• Equi-depth histogram

- ensure the number of tups in a bucket is the same



- good choice for selectivity estimation

c.f.

Size of Join Computation

R, S (join on A)

$$|R \bowtie_A S| = \sum_i f_i \, g_i$$

$\tilde{f}_i$, $\tilde{g}_i$ approx

$$|R \bowtie_A S| \approx \sum_{b \in B} \tilde{f}_b \cdot \tilde{g}_b \cdot |b|$$