Outlier detection and removal

- ~~a statistical viewpoint~~
- ~~a statistician's viewpoint~~
- one statistician's viewpoint

Outlier

Wikipedia : an observation point that is distant from other observations.[1],[2]

"i hate wiki"

[1] : "appears to deviate markedly from other members of the sample in which it occurs"

Grubbs, 1969, Technometrics

[2] : "far removed from the rest of the observations"

Maddala, 1992, Introduction to Econometrics

(1) extreme manifestation : keep

(2) error : investigate

Hypothesis:

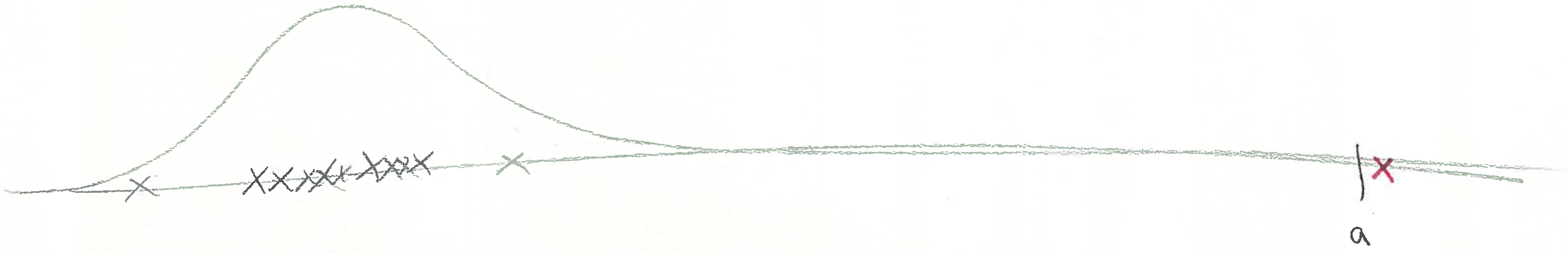
x comes from the same population as the rest of the observations

$P(a < X < b) < \alpha$, $P(X > a)$

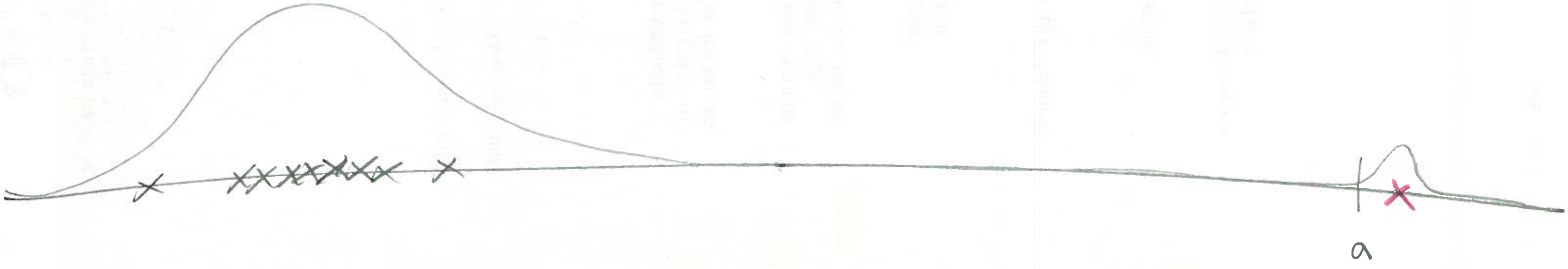
P is unknown

estimate $\hat{P}$ from the data

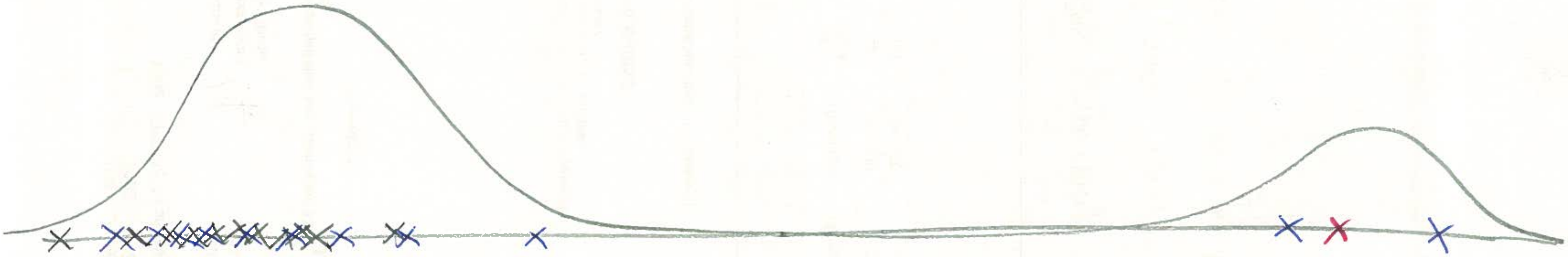
(1) Gaussian distribution : $P(X > a) < \alpha$



(2) Mixture of Gaussian distributions : $P(X > a) > \alpha$



True distribution



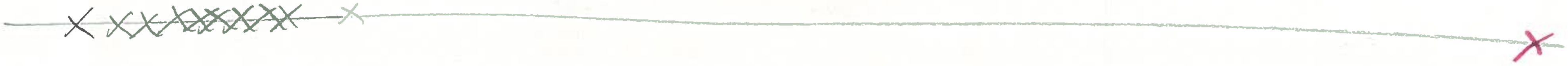
Exploratory data analysis: suggest hypothesis

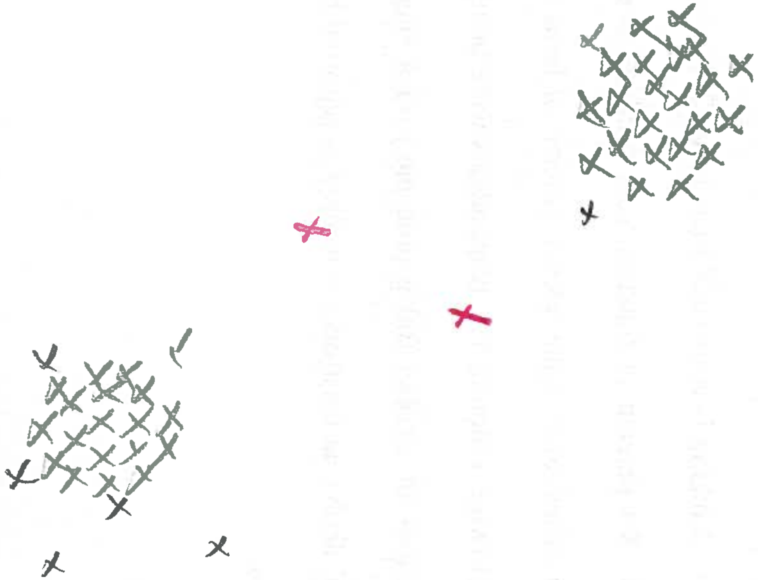
vs

Confirmatory data analysis: test hypothesis

Robustness

median vs mean





increase separability of the two groups

$p > 0.05 \rightarrow p < 0.05$

The curse of dimensionality

Bellman, Dynamic Programming, 1957

The amount of data needed grows exponentially with dimensionality.

Outlier detection

Which assumptions are made ?

Outlier removal

Exploratory data analysis

- extreme caution
- neatly reported, numbers & criteria

Confirmatory data analysis

- just don't
- analysis of censored observations