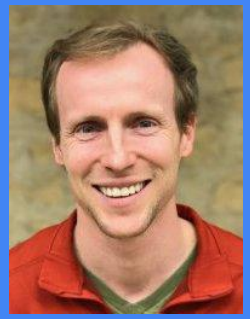
Optimal Confidence Sets for the Multinomial Parameter

Ardhendu Tripathy, Computer Science department



Are these rating distributions different?

1,072 customer reviews

4.2 out of 5 stars

5 star	62%
4 star	18%
3 star	9%
2 star	5%
1 star	6%

690 customer reviews

4.5 out of 5 stars

5 star	74%
4 star	12%
3 star	6%
2 star	4%
1 star	4%

$$\mathbf{p} = (p_5, p_4, p_3, p_2, p_1)$$

Success probability of a binomial

$\Pr(X = 1) = p, \Pr(X = 0) = 1 - p$

Observe n i.i.d. samples $X_1, X_2, \ldots X_n$

$\mathcal{C}_\delta(X_1, \ldots, X_n)$ is a confidence set at level δ if

for all $p \in [0, 1]$, $\Pr(p \notin \mathcal{C}_\delta(X_1. \ldots, X_n)) \leq \delta$

Commonly used sets have approx coverage

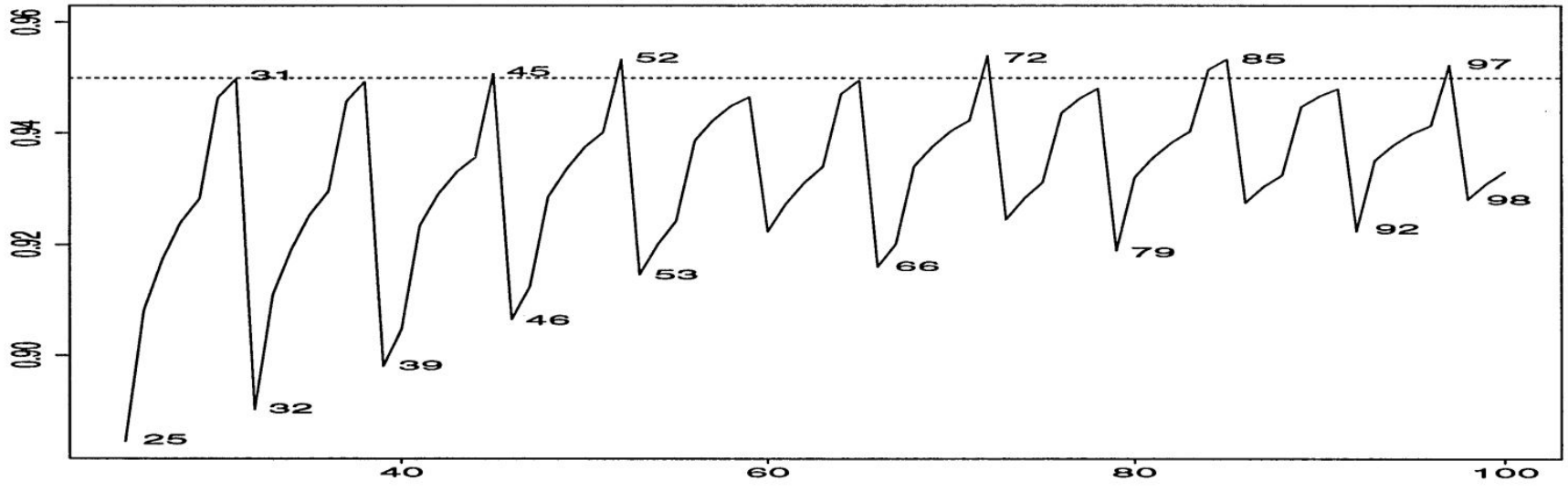


FIG. 1. *Standard interval; oscillation phenomenon for fixed* $p = 0.2$ *and variable* $n = 25$ *to* 100.

Commonly used sets have approx coverage

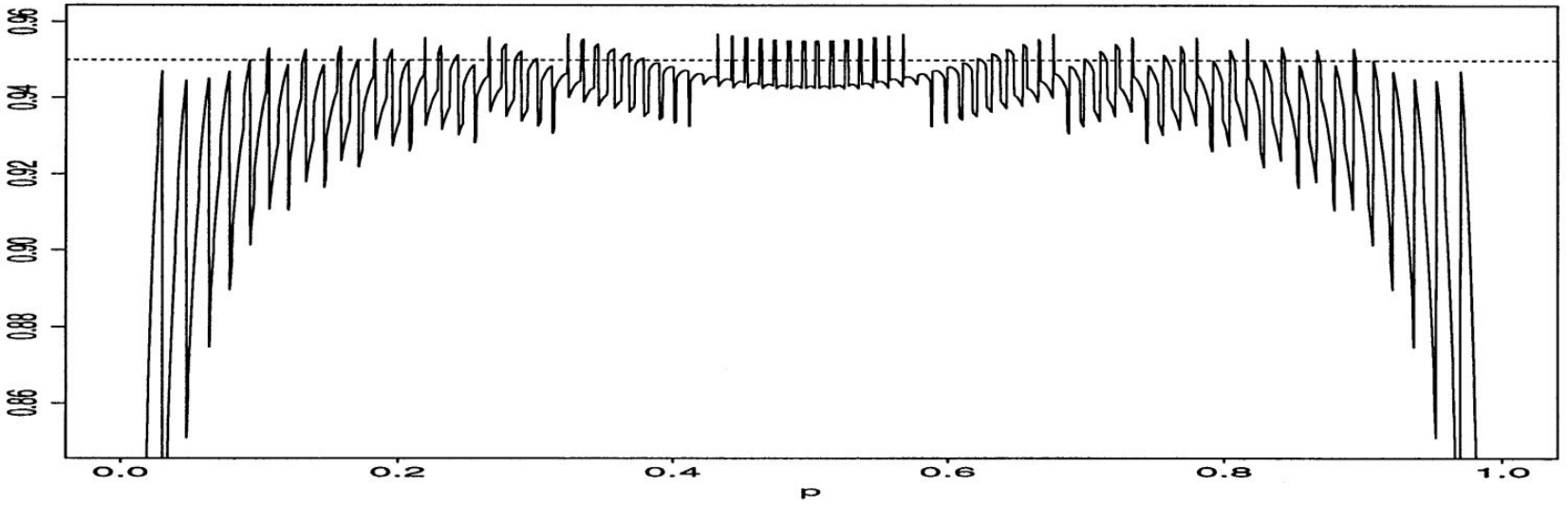


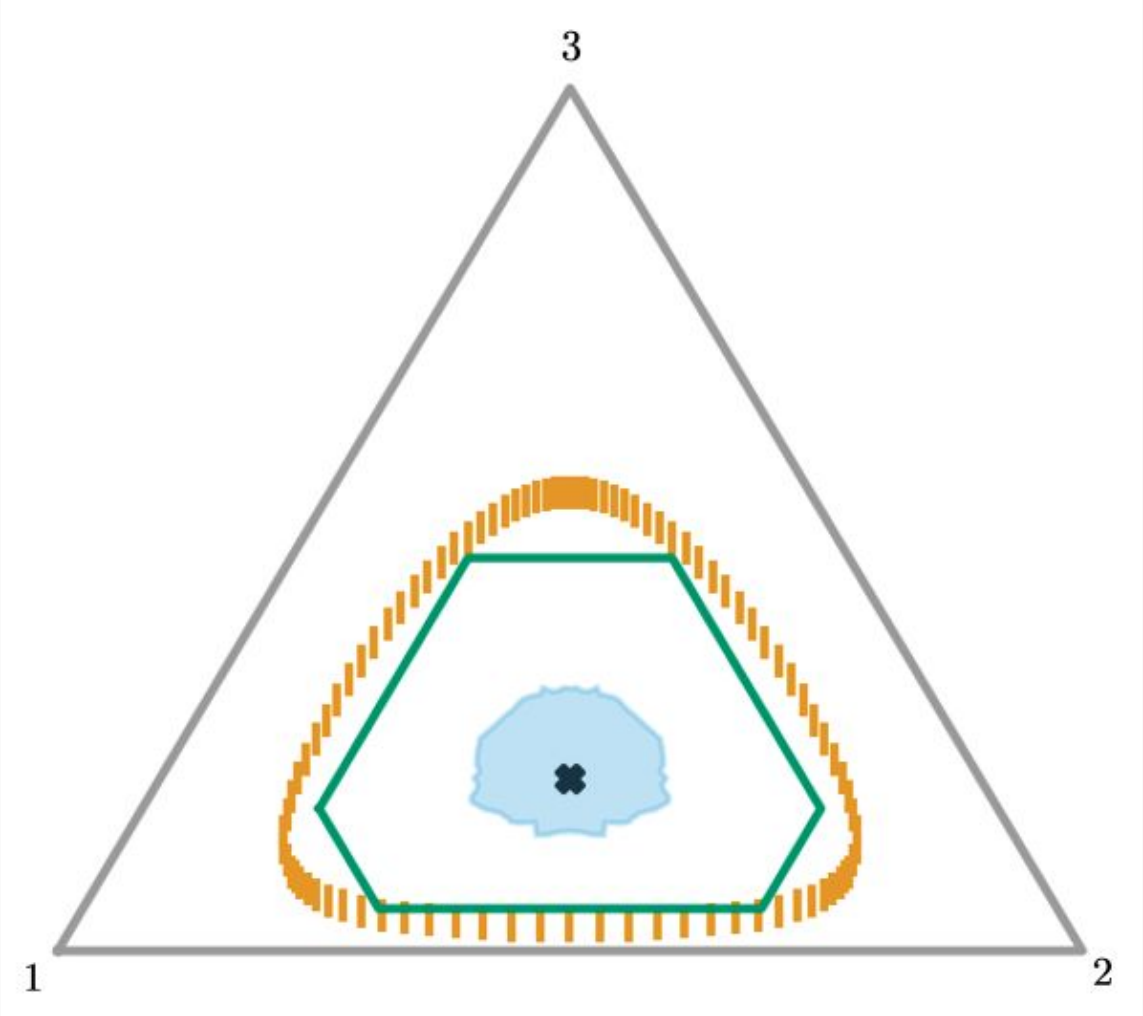
FIG. 3. *Standard interval; oscillation phenomenon for fixed* $n = 100$ *and variable* p.

Level-set method to obtain confidence sets

- Can be applied to data with more than two categories, i.e., multinomial
- Confidence sets have guaranteed coverage
- Confidence sets are “optimal”, i.e., they have smallest expected volume among all confidence sets that have guaranteed coverage
- Obtaining these confidence sets is computationally intensive

Example:
3 categories,
15 samples

- Black cross is empirical average
- Blue is by level-set method
- Green is by union bound over marginal
- Orange is based on Sanov's theorem



Level-set method to obtain confidence sets

- For the given sample size n, compute all possible histograms, called h_i
  - For any p, calculate P(h_i | p) for all h_i
  - Let P(h_1 | p) >= P(h_2 | p) >= P(h_3 | p) >= ....
  - Find smallest i* such that
    - P(h_1 | p) + P(h_2 | p) + … + P(h_i* | p) >= 1 - δ
  - If observed histogram h has index <= i*, p belongs to confidence set
  - Else, p does not belong to confidence set