Ick, Math!
Ensuring Production Quality

IG3 West
December 10, 2019

Clustify

Panelists

- Bill Dimm, Hot Neuron
- Lilith Bat-Leah, Fronteo
- Cynthia Vasquez, CHAT Consulting
- Grady Glover, The Rodarti Group & Lawrence Bartels
- Tammi Loveland, U.S. DoJ

Clustify

Prevalence

- Percentage having some property (relevant)

Recall

- Percentage of relevant docs found
- Defensibility

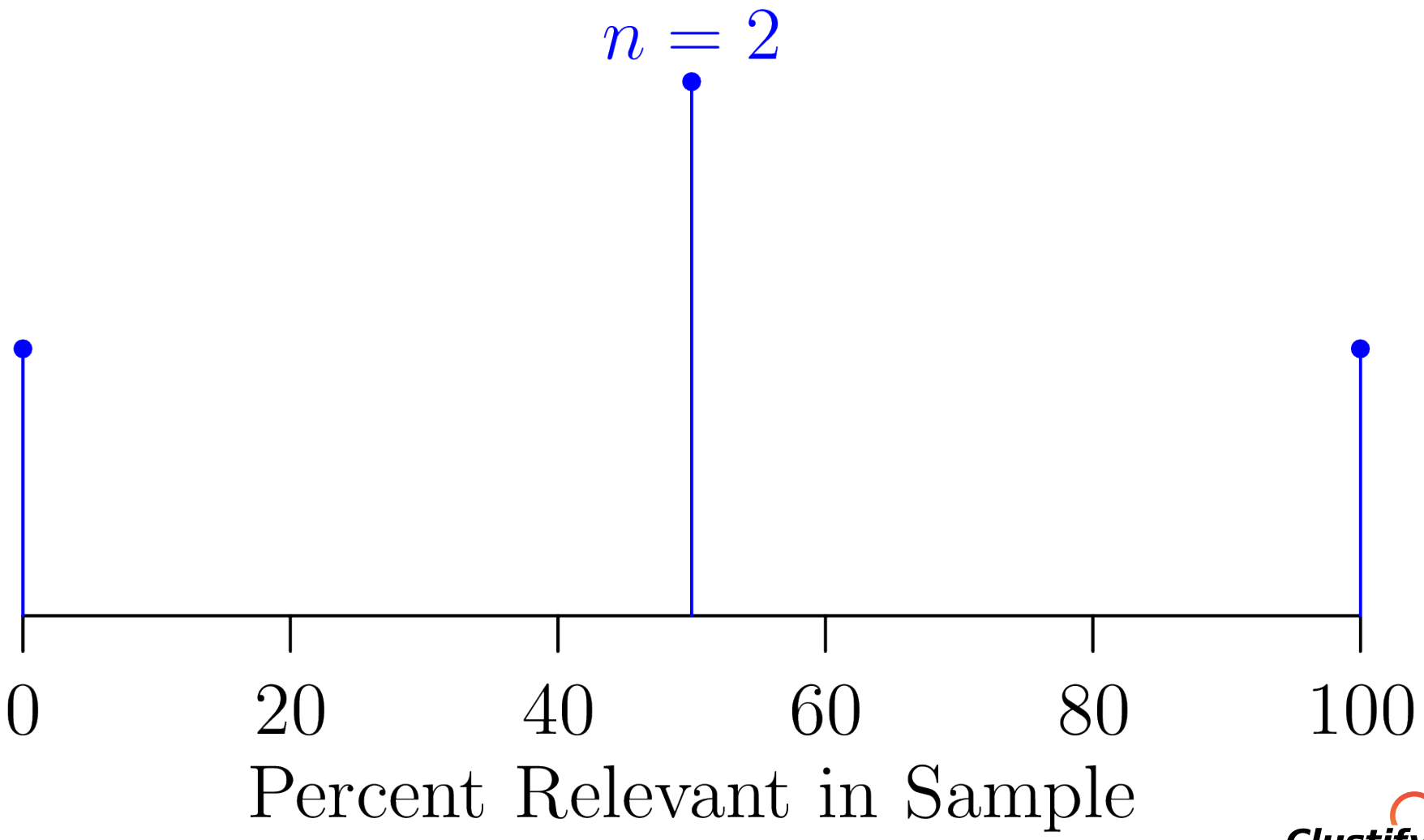
Clustify

Precision

- Percentage of retrieved docs that are relevant
- Cost (review effort)

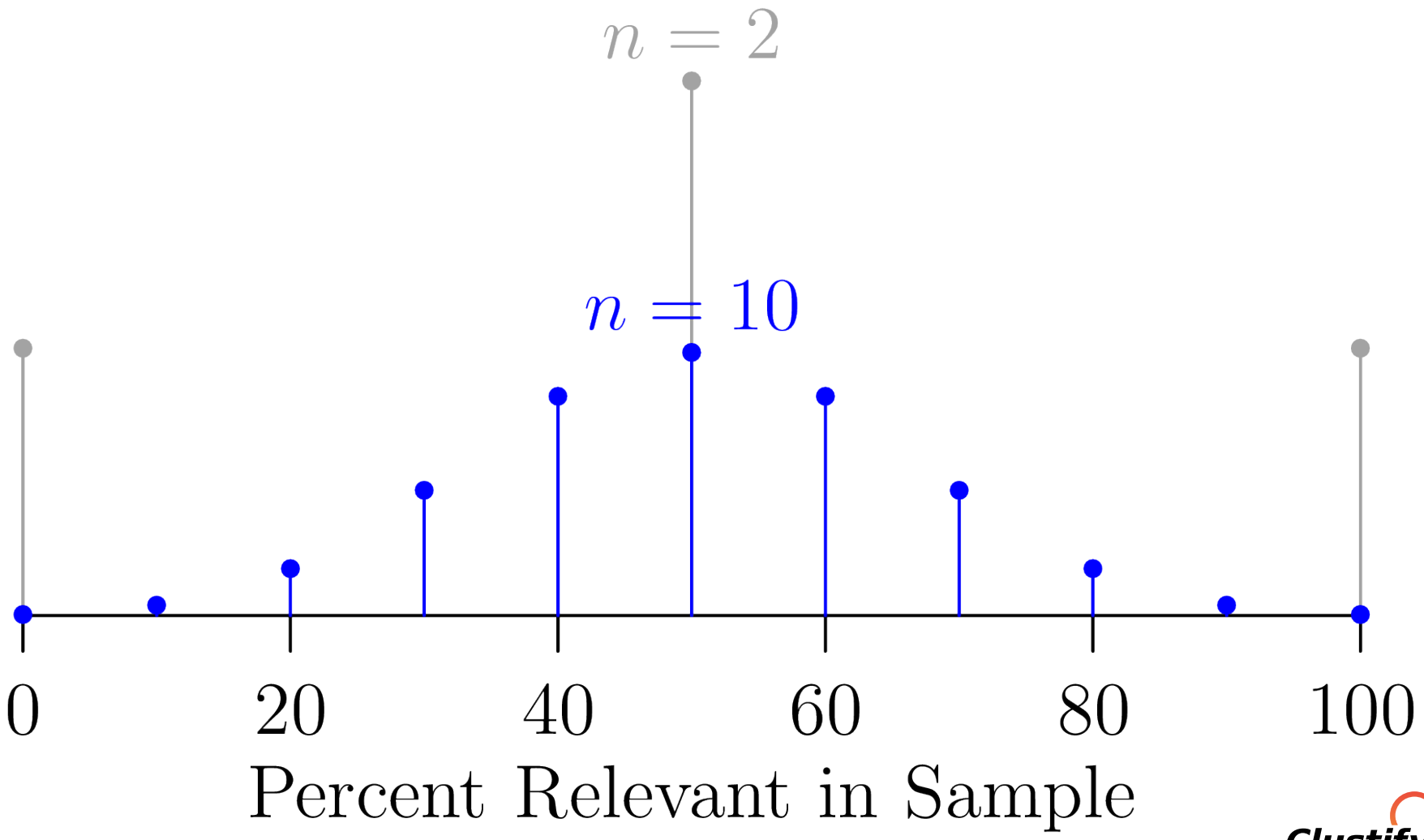
Clustify

Sample 2 Docs



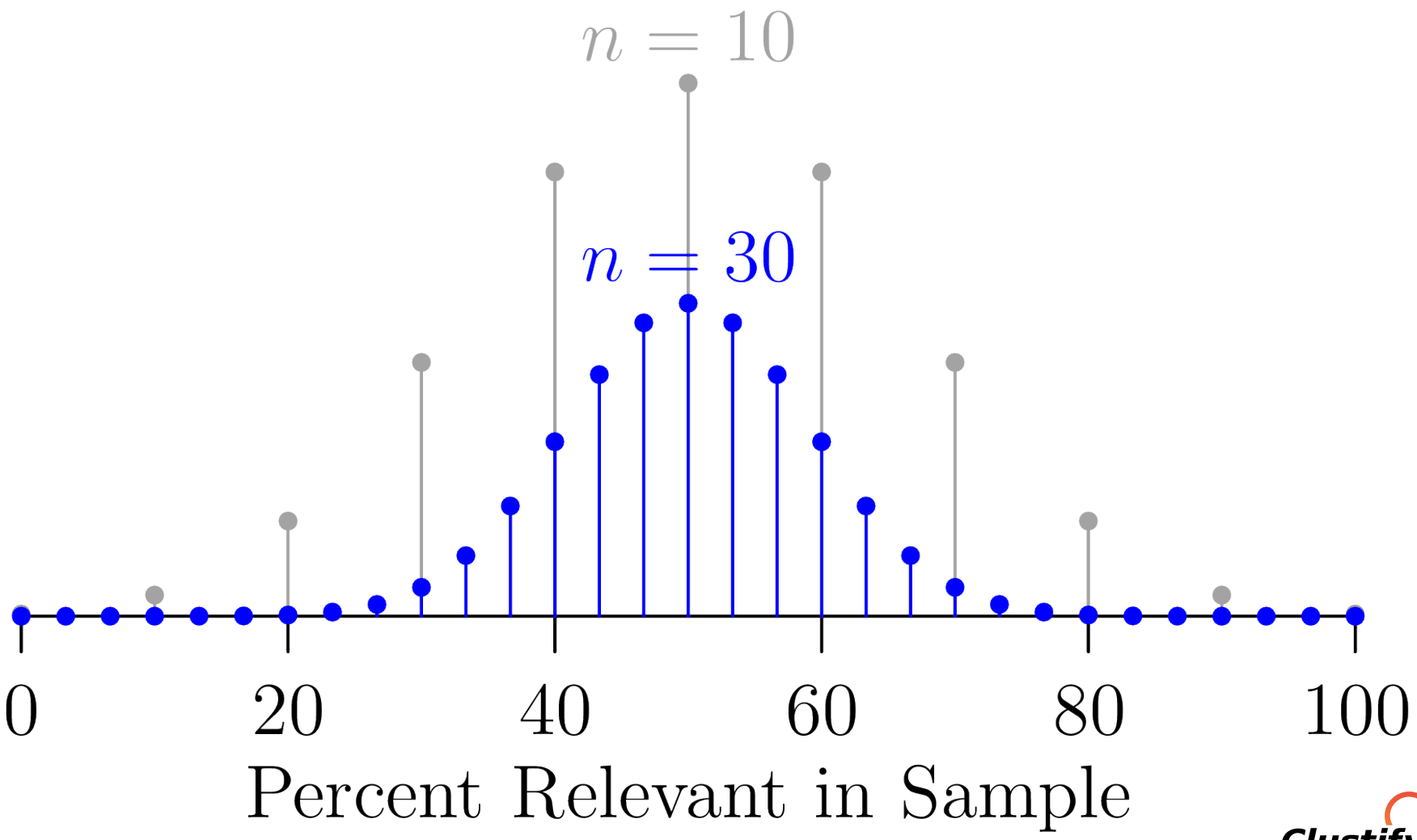
Clustify

Sample 10 Docs



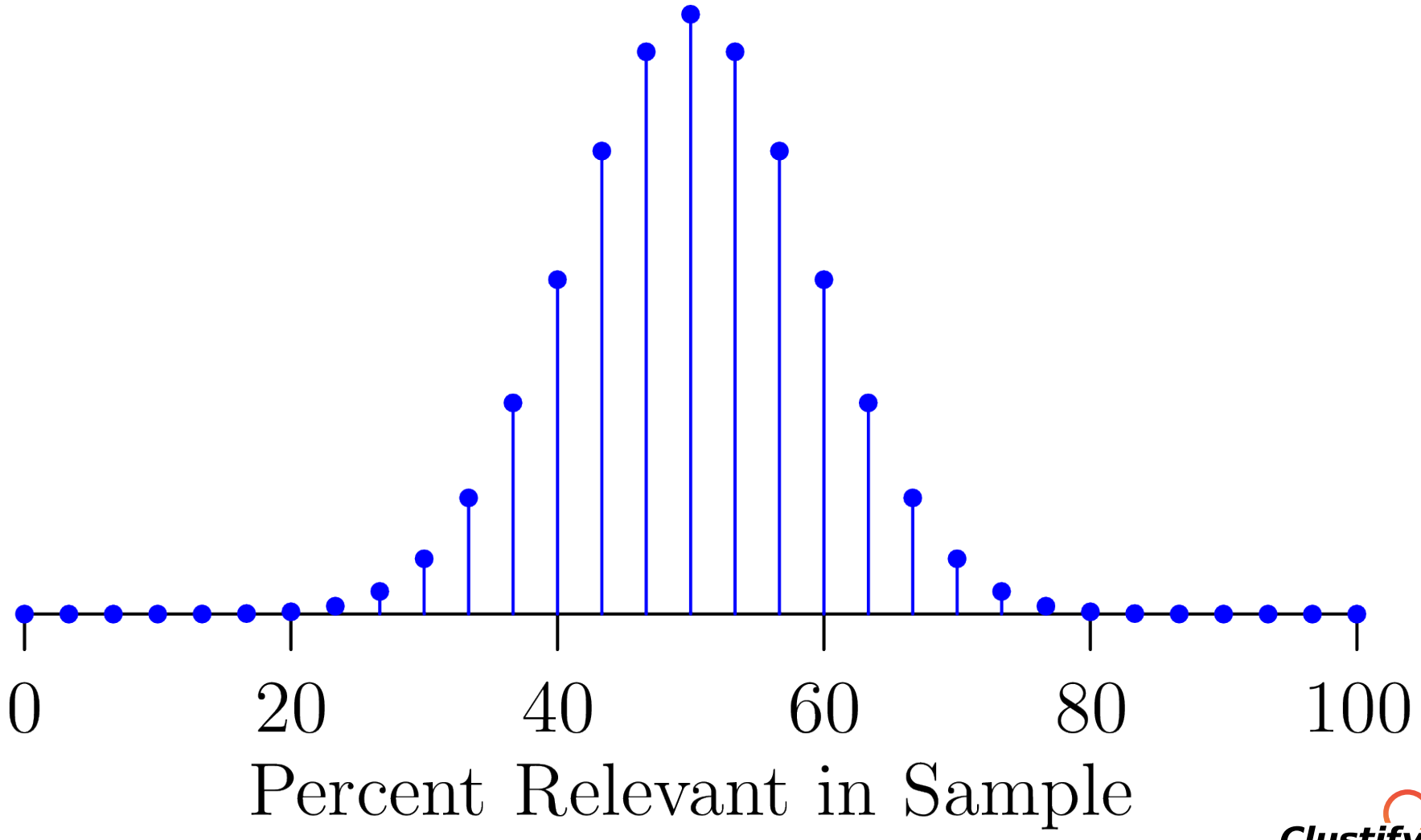
Clustify

Sample 30 Docs



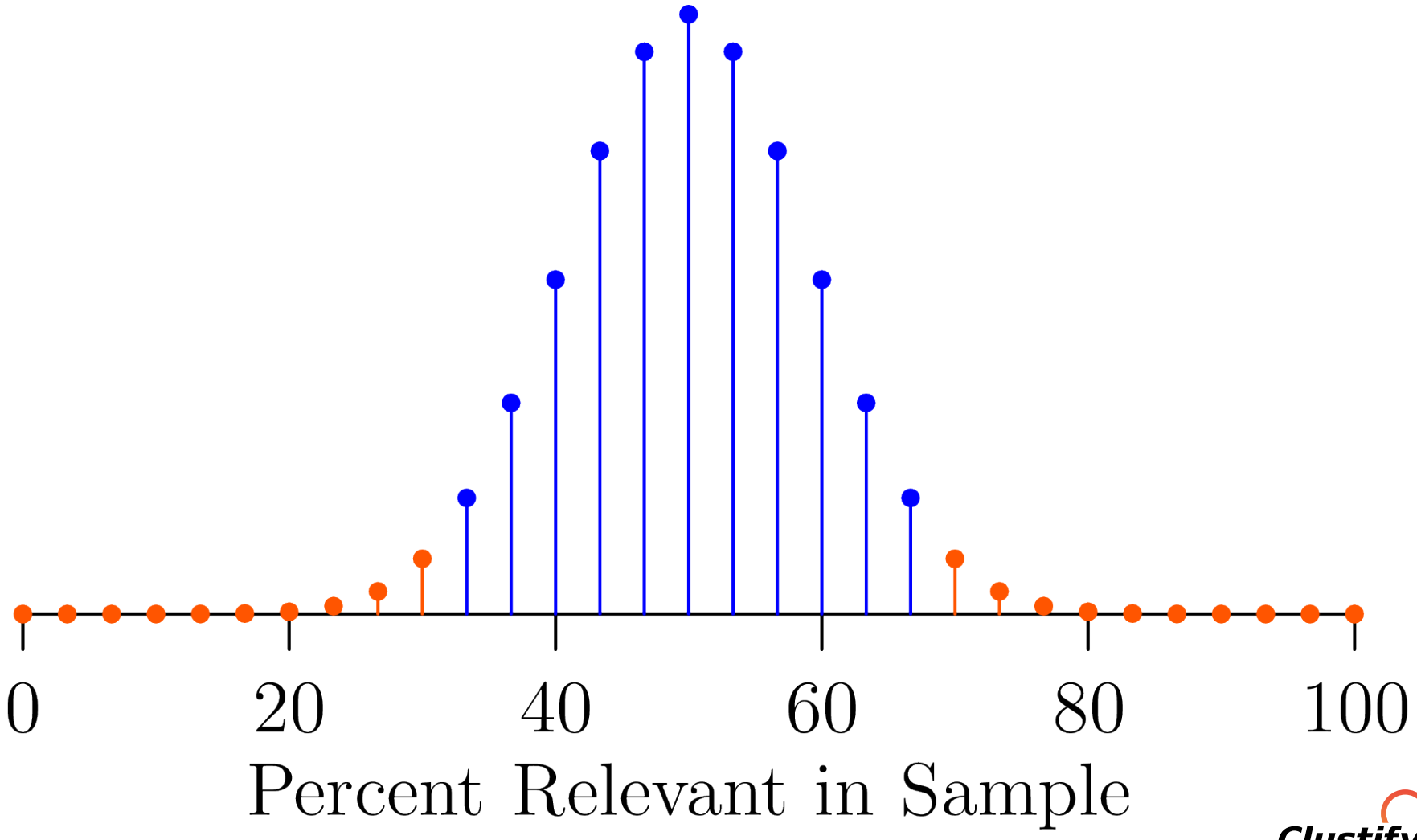
Clustify

Anything Is Possible



Clustify

What Is Plausible?

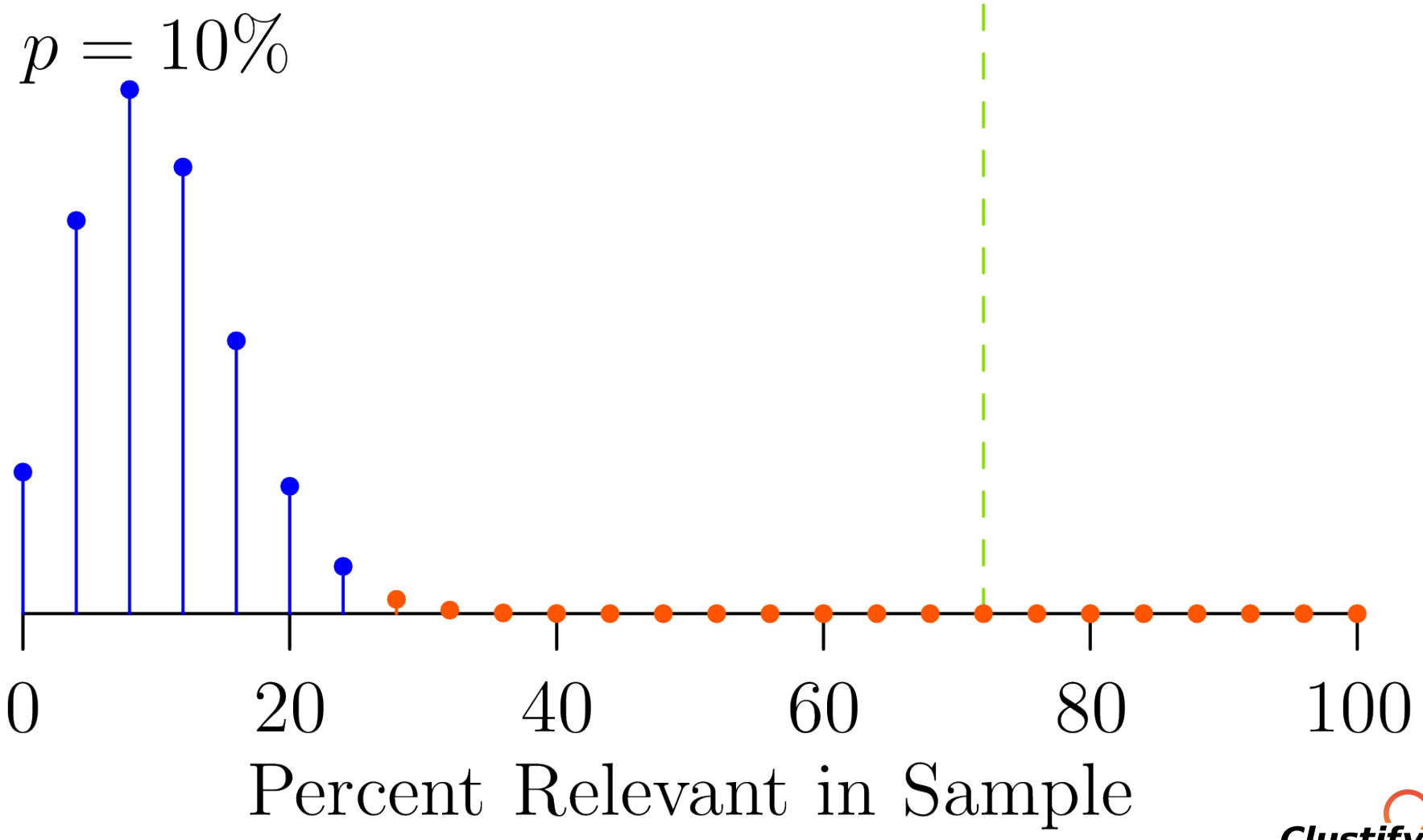


Clustify

Example of Finding a Confidence Interval

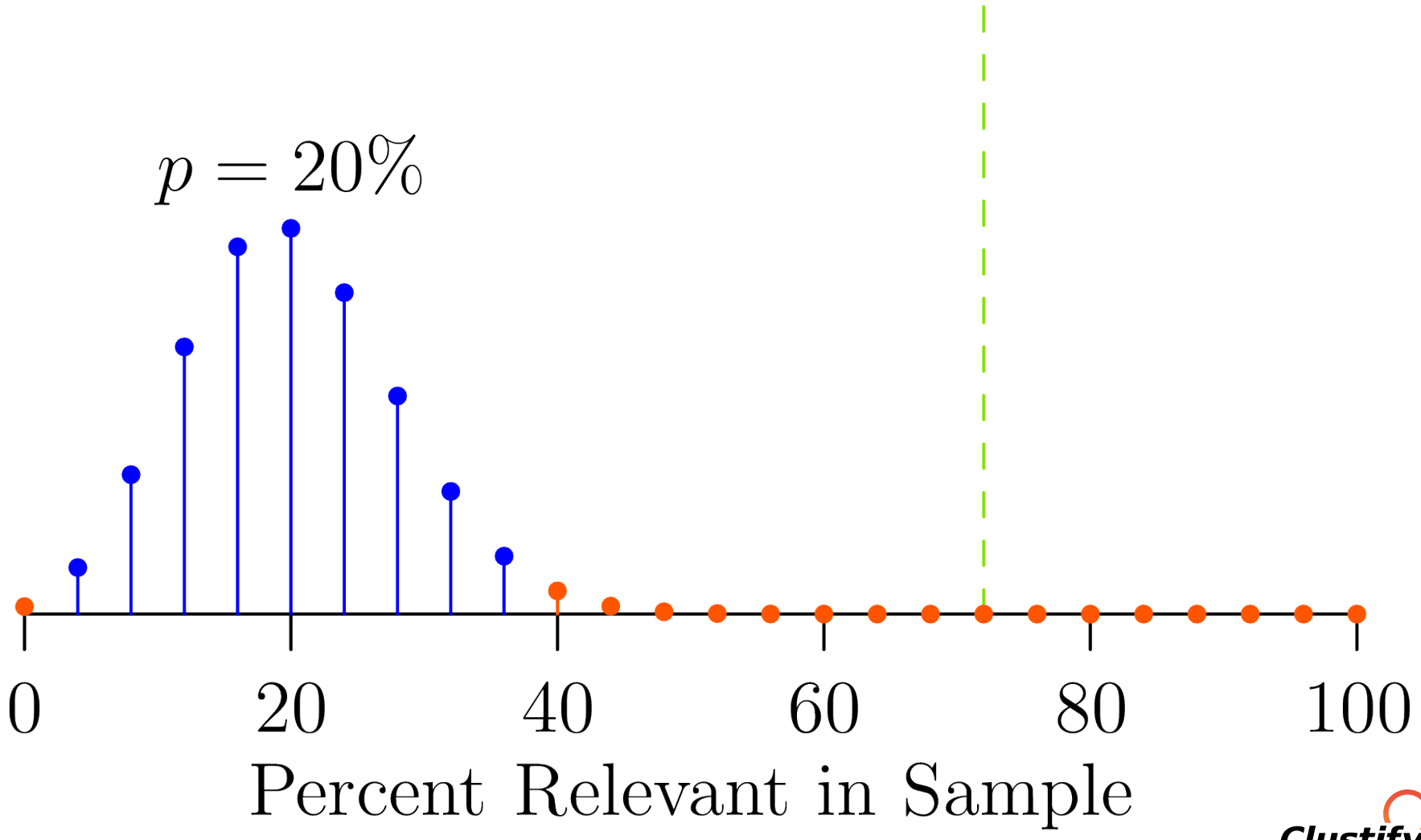
- Sample 25 documents, 18 are relevant
  - 72% of sample is relevant
- What is reasonable prevalence for population?
  - Point estimate is 72%

Plausible Population Prevalence? NO



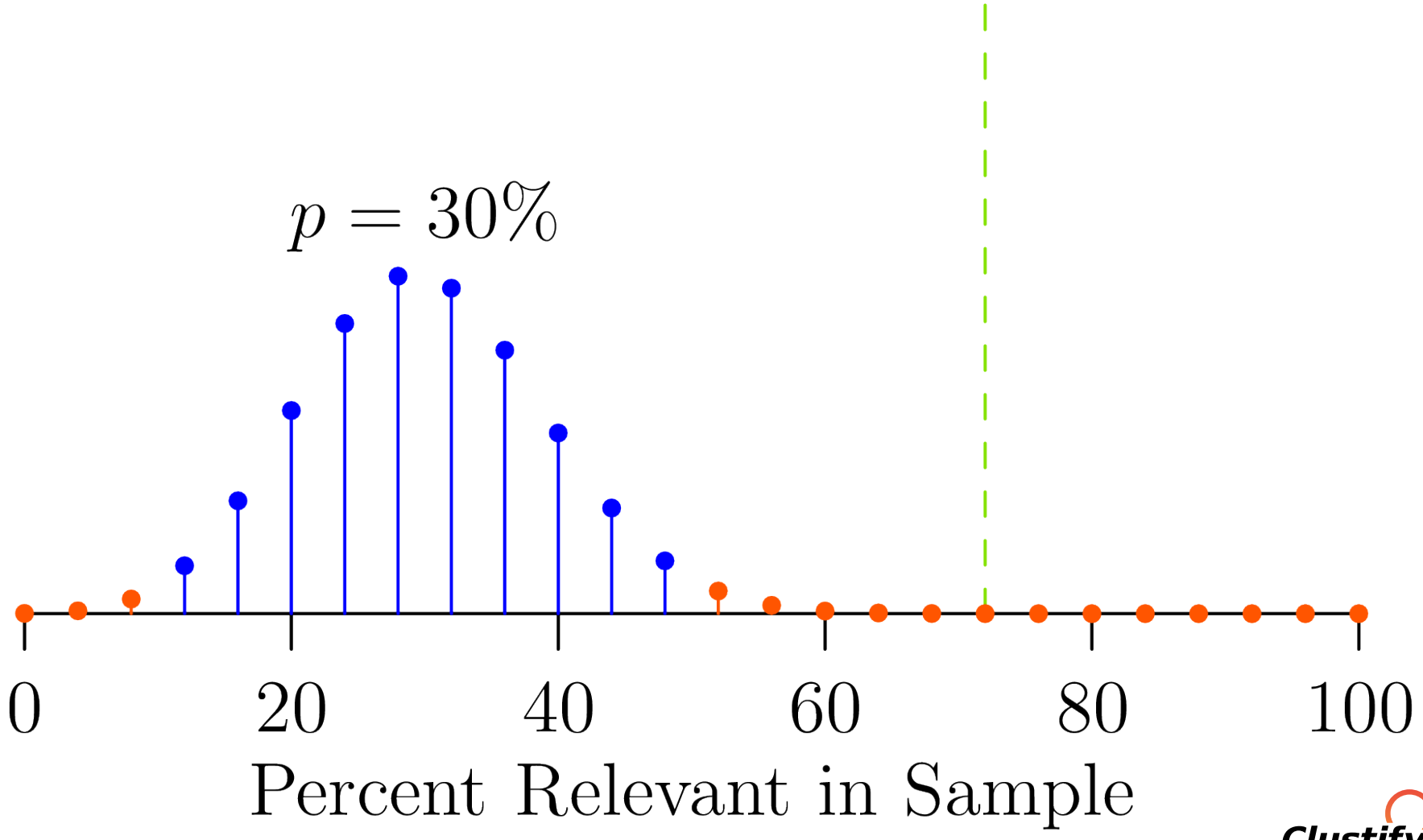
Clustify

Plausible Population Prevalence? NO



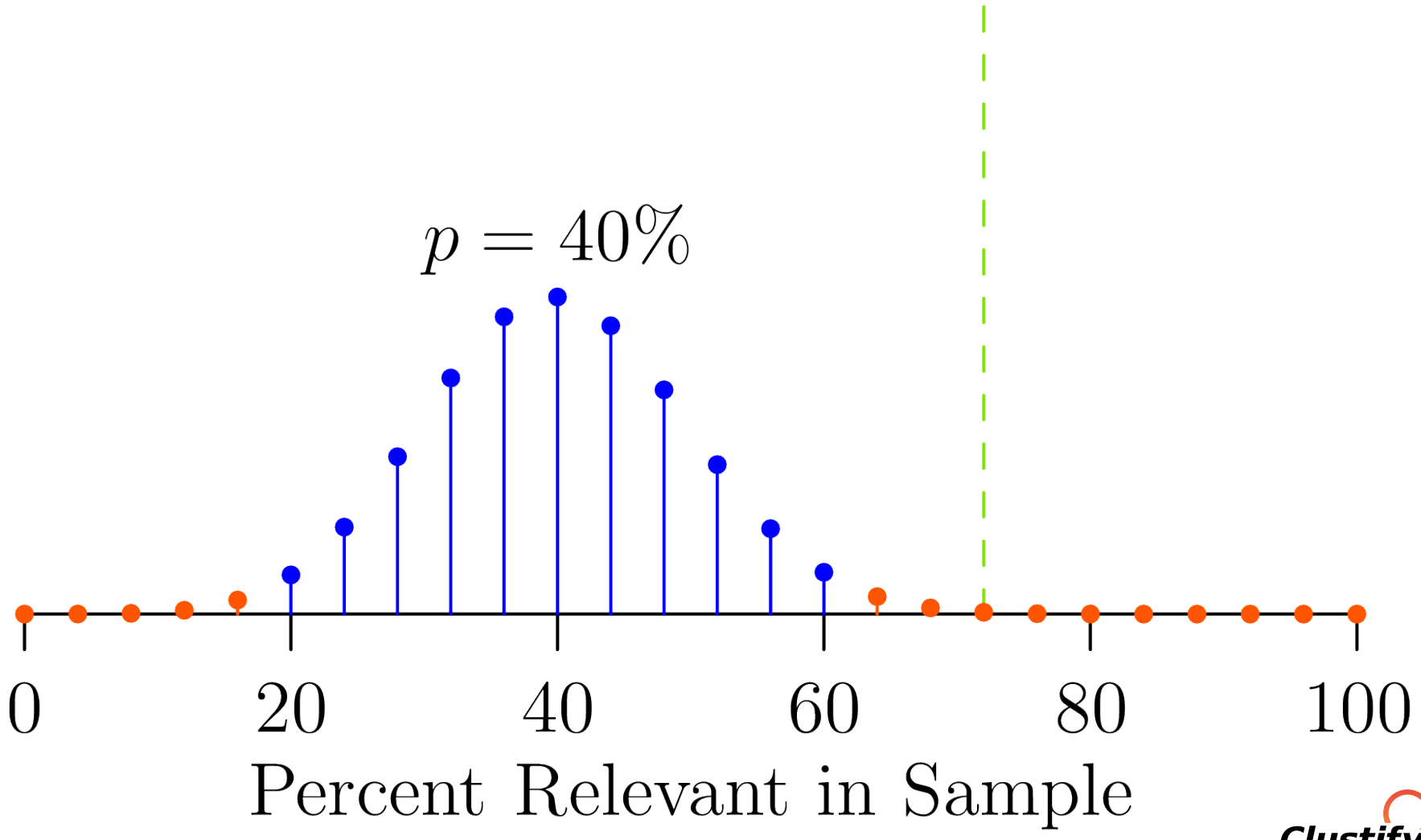
Clustify

Plausible Population Prevalence? NO



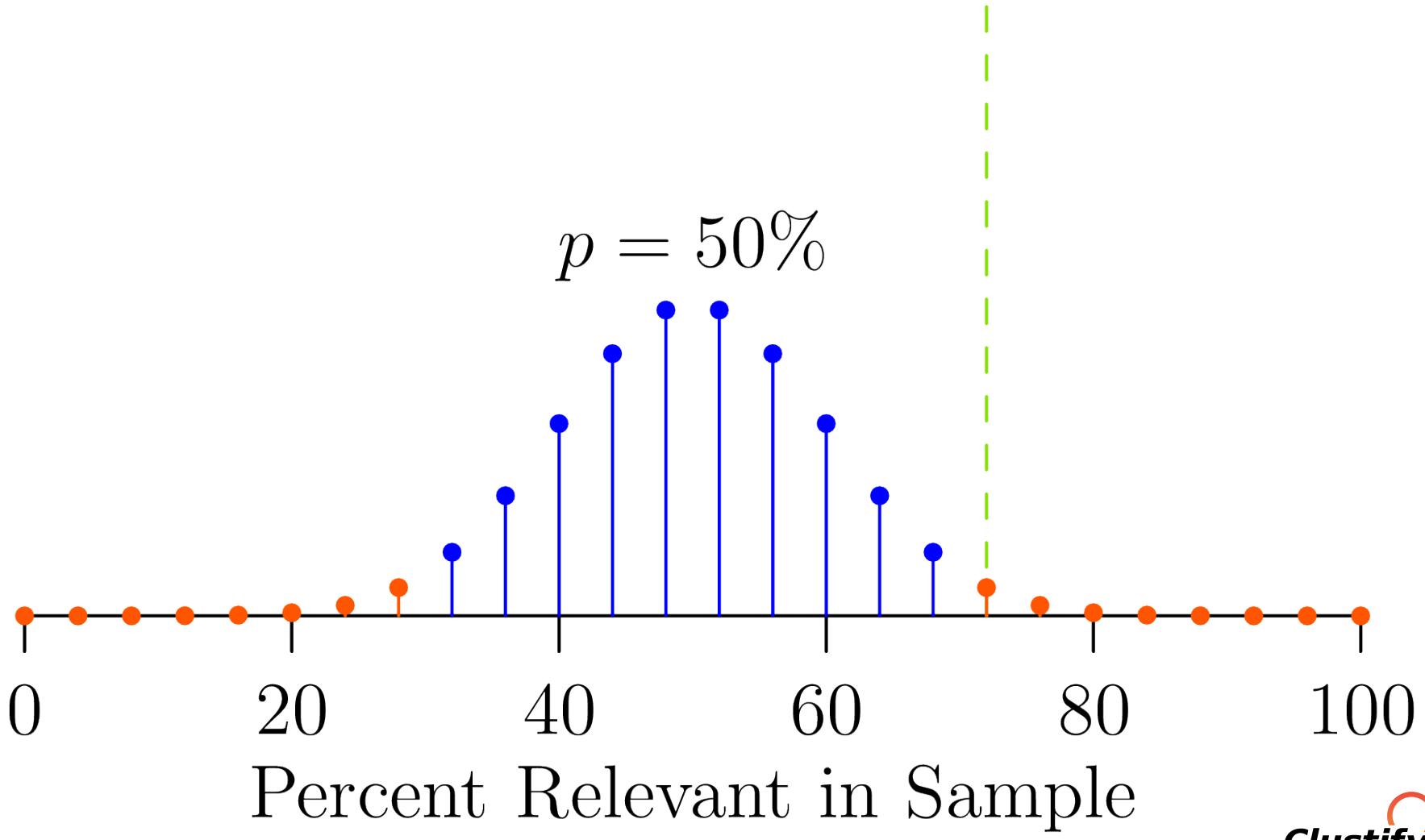
Clustify

Plausible Population Prevalence? NO



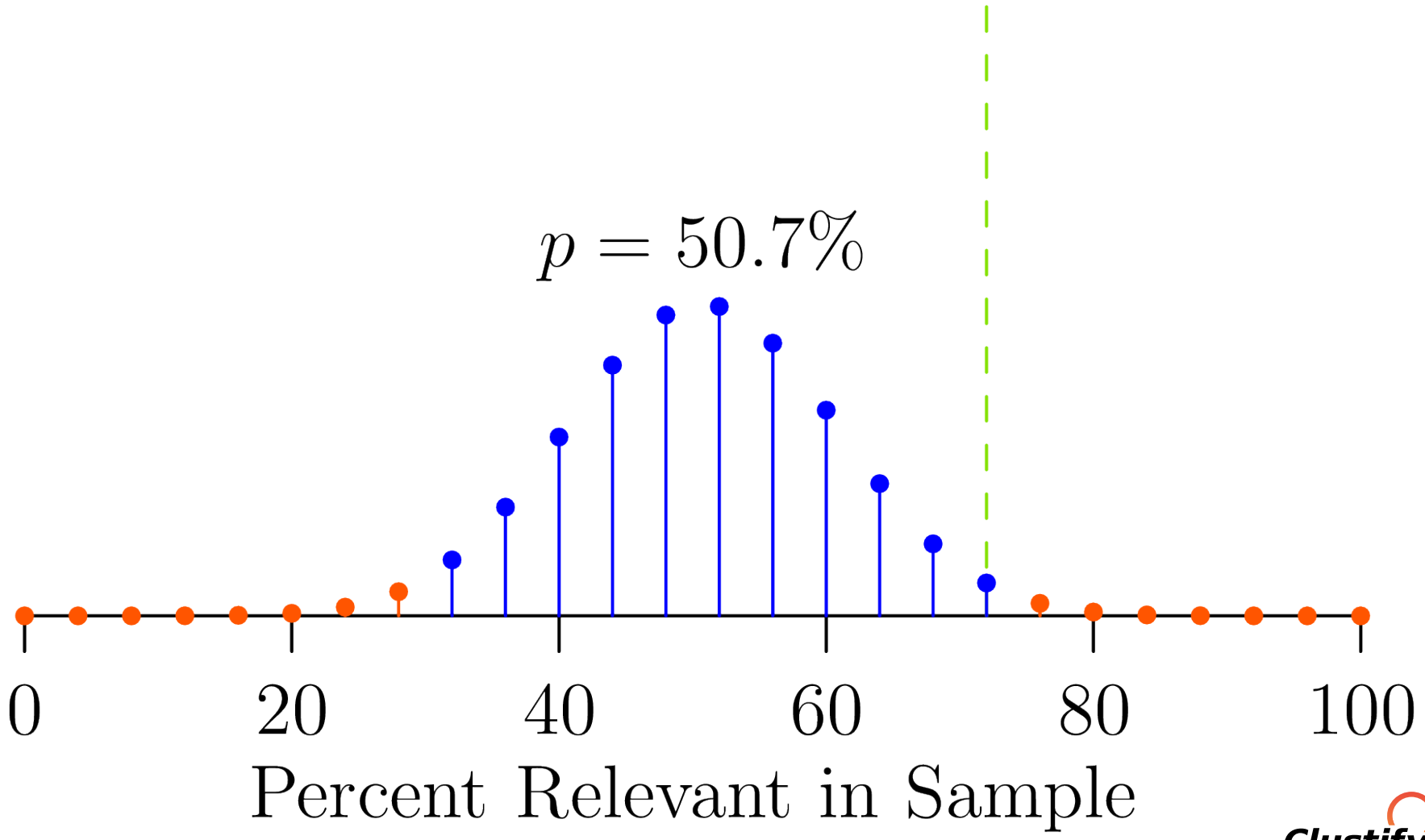
Clustify

Plausible Population Prevalence? NO



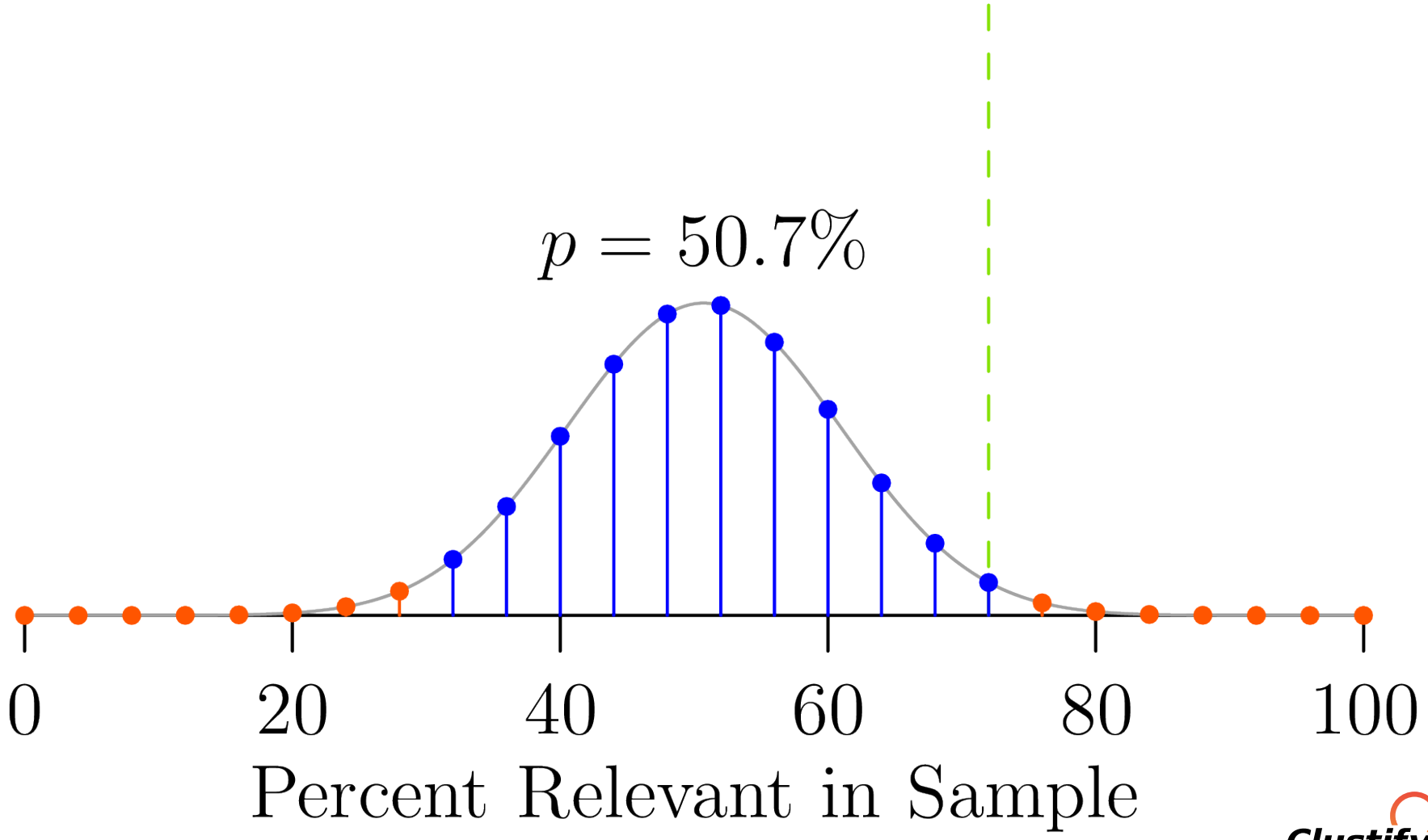
Clustify

Plausible Population Prevalence? YES



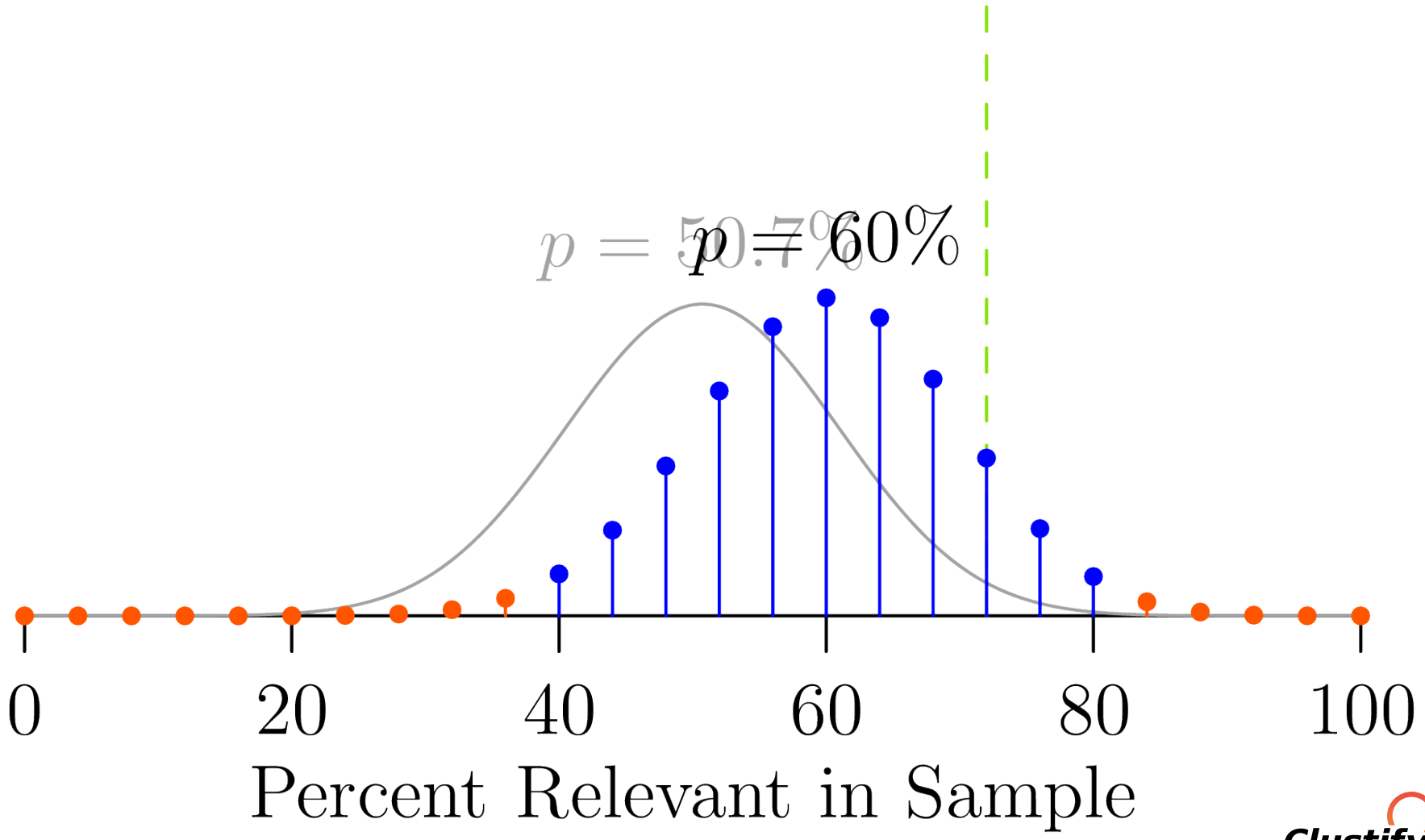
Clustify

Plausible Population Prevalence? YES



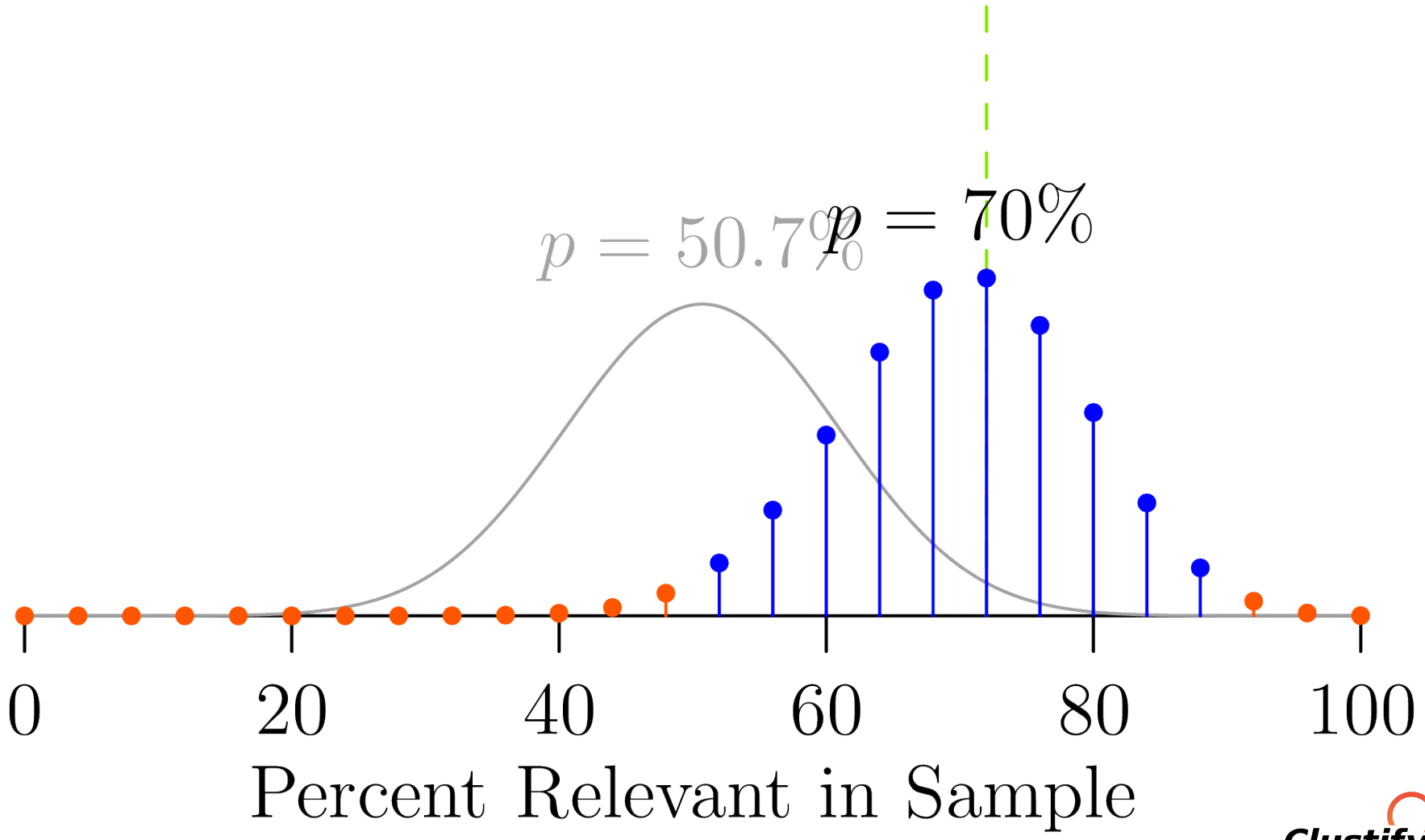
Clustify

Plausible Population Prevalence? YES



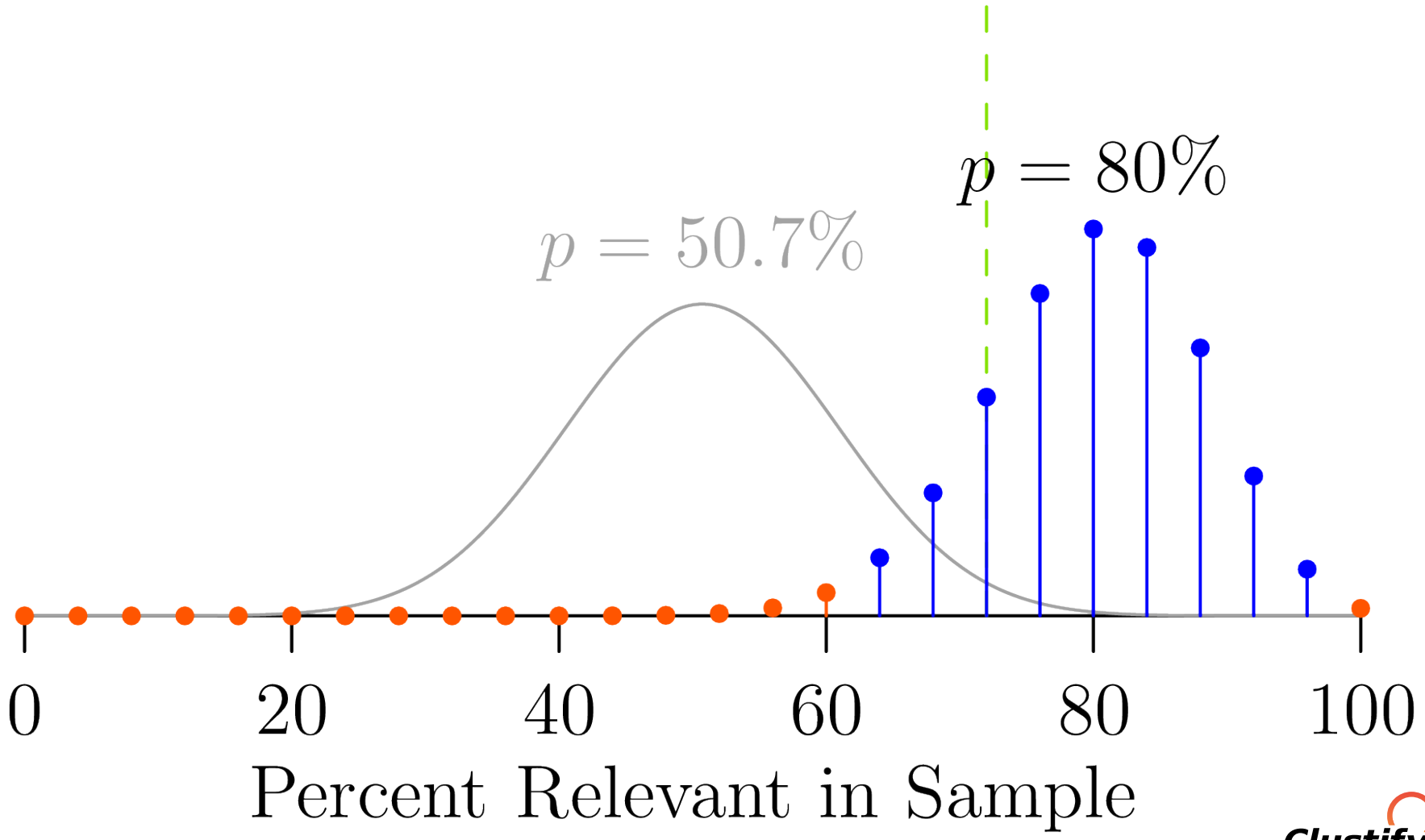
Clustify

Plausible Population Prevalence? YES



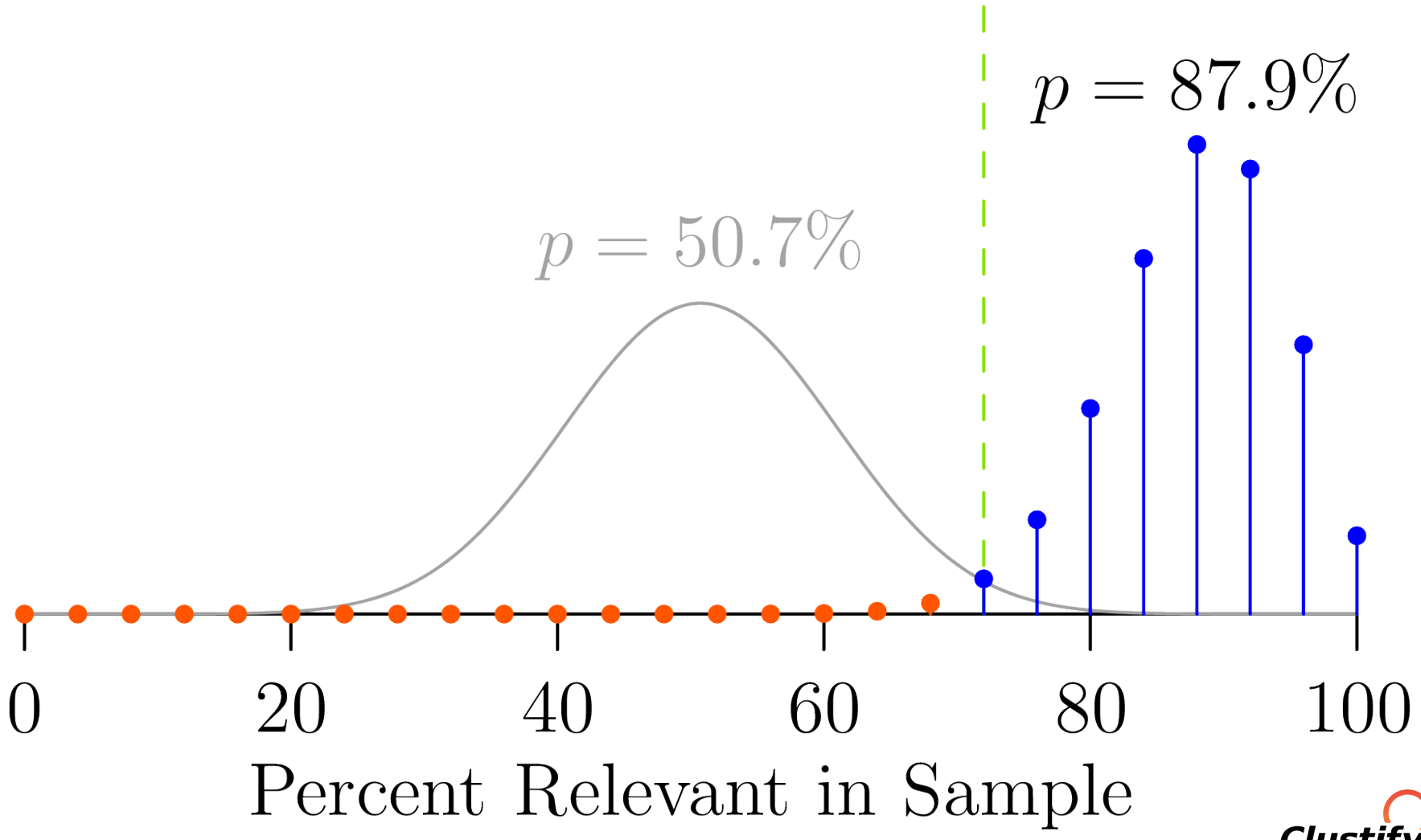
Clustify

Plausible Population Prevalence? YES



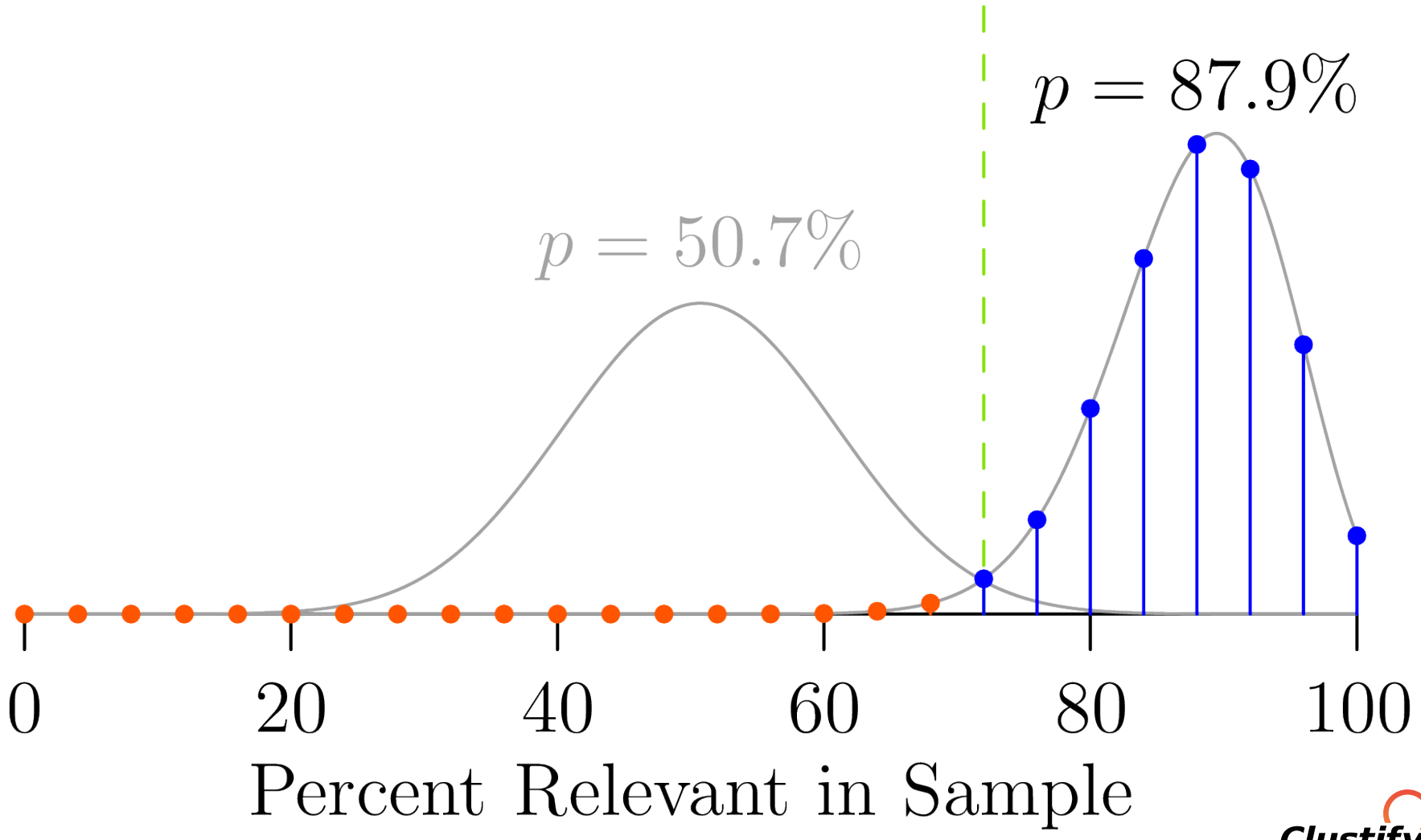
Clustify

Plausible Population Prevalence? YES



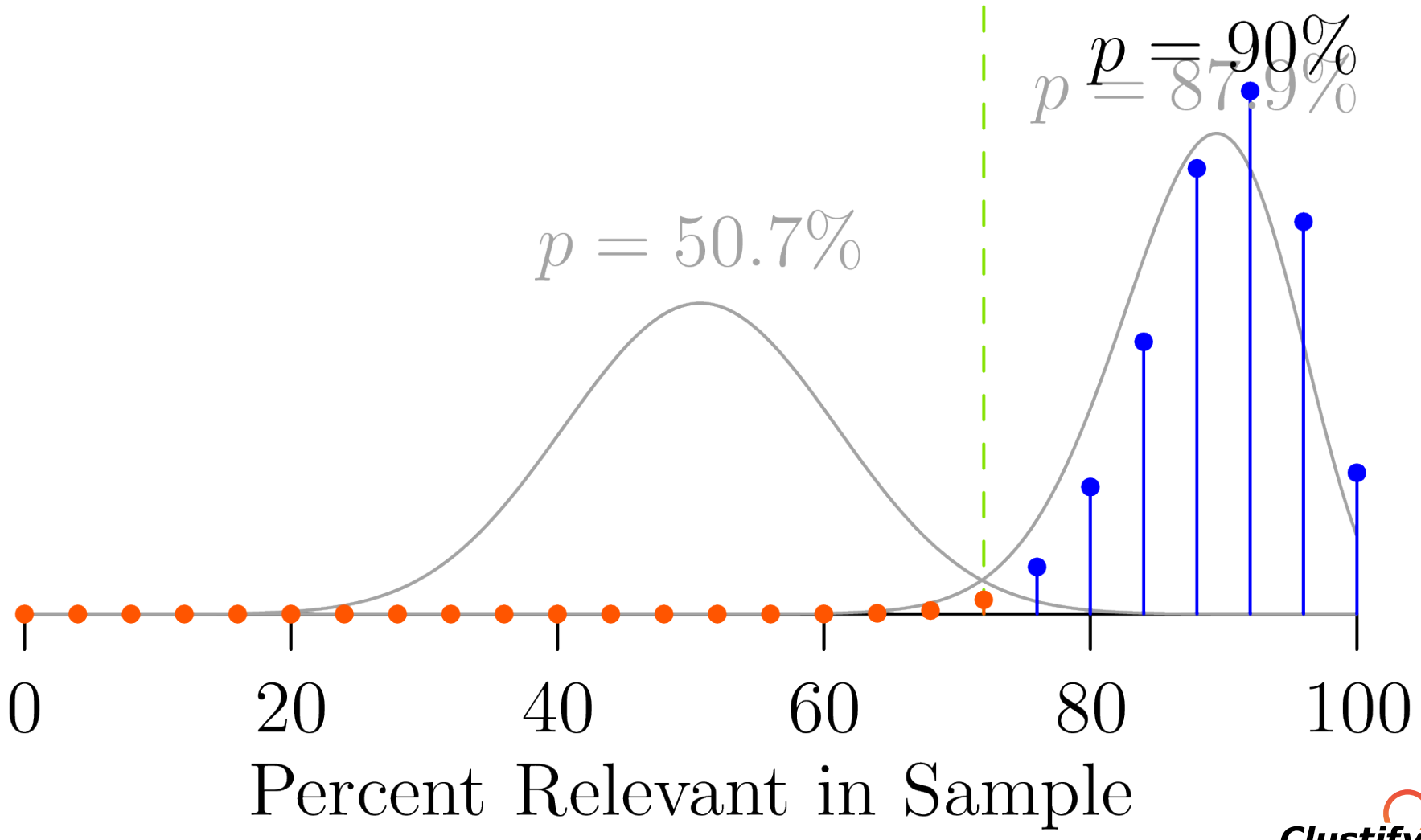
Clustify

Plausible Population Prevalence? YES



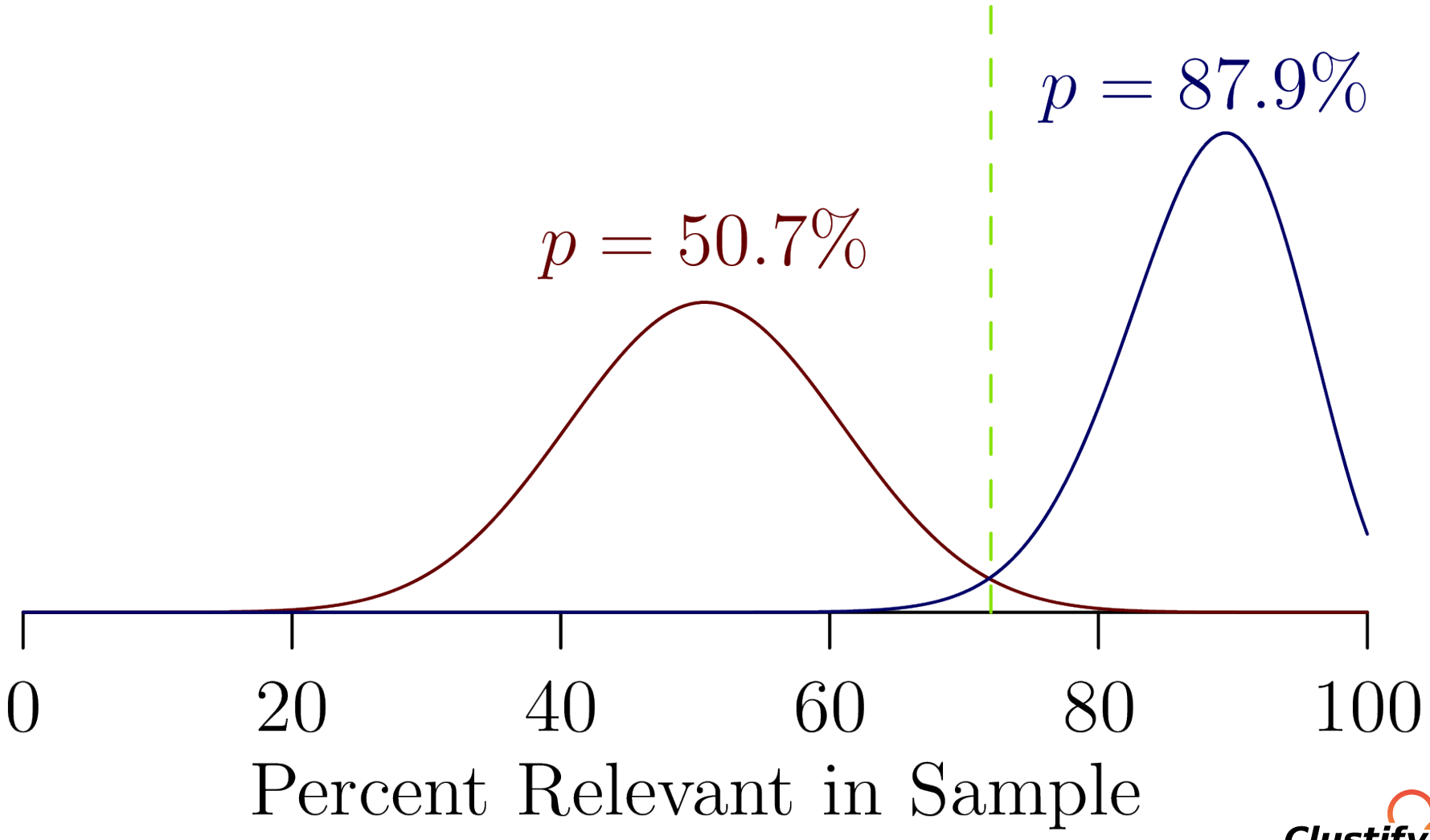
Clustify

Plausible Population Prevalence? NO



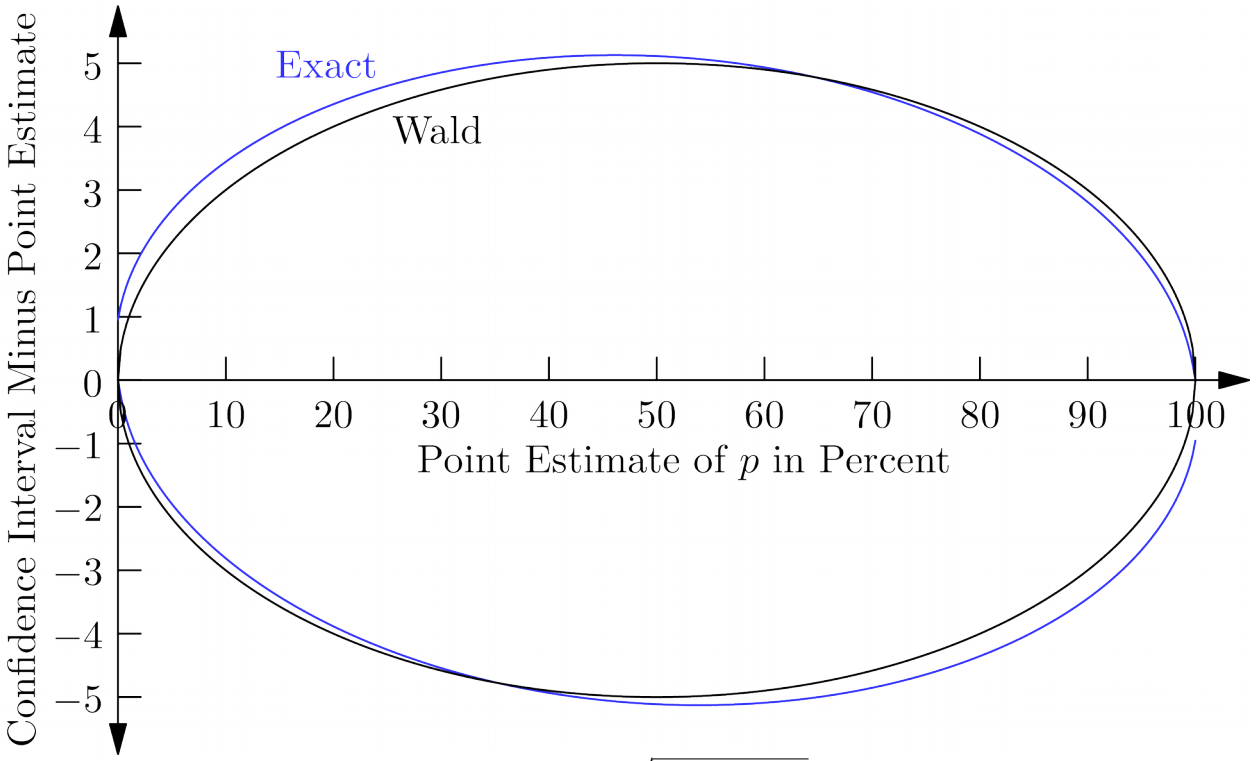
Clustify

Confidence Interval



Clustify

Wald Approximation



$$\delta p = 1.96\sqrt{\frac{p(1-p)}{S}}$$

$$\delta p_{\max} = \frac{0.98}{\sqrt{S}}$$

$$S = \frac{0.96}{\delta p_{\max}^2}$$

Clustify

Standard Sample Sizes, 95% Confidence

$\delta\rho_{\max}$	S
1%	9,600
2%	2,400
3%	1,067
4%	600
5%	384
7%	196
10%	96
15%	43

Example

- Sample 10,000 Docs From Full Population
  - 100 are relevant
  - 200 match search query
  - 80 are relevant and match search query

Quantity	Point Est	Sample Size	Worst Case CI	Wald CI
Prevalence	1%	10,000	±1%	±0.2%
Precision	40%	200	±7%	±6.8%
Recall	80%	100	±10%	±7.8%

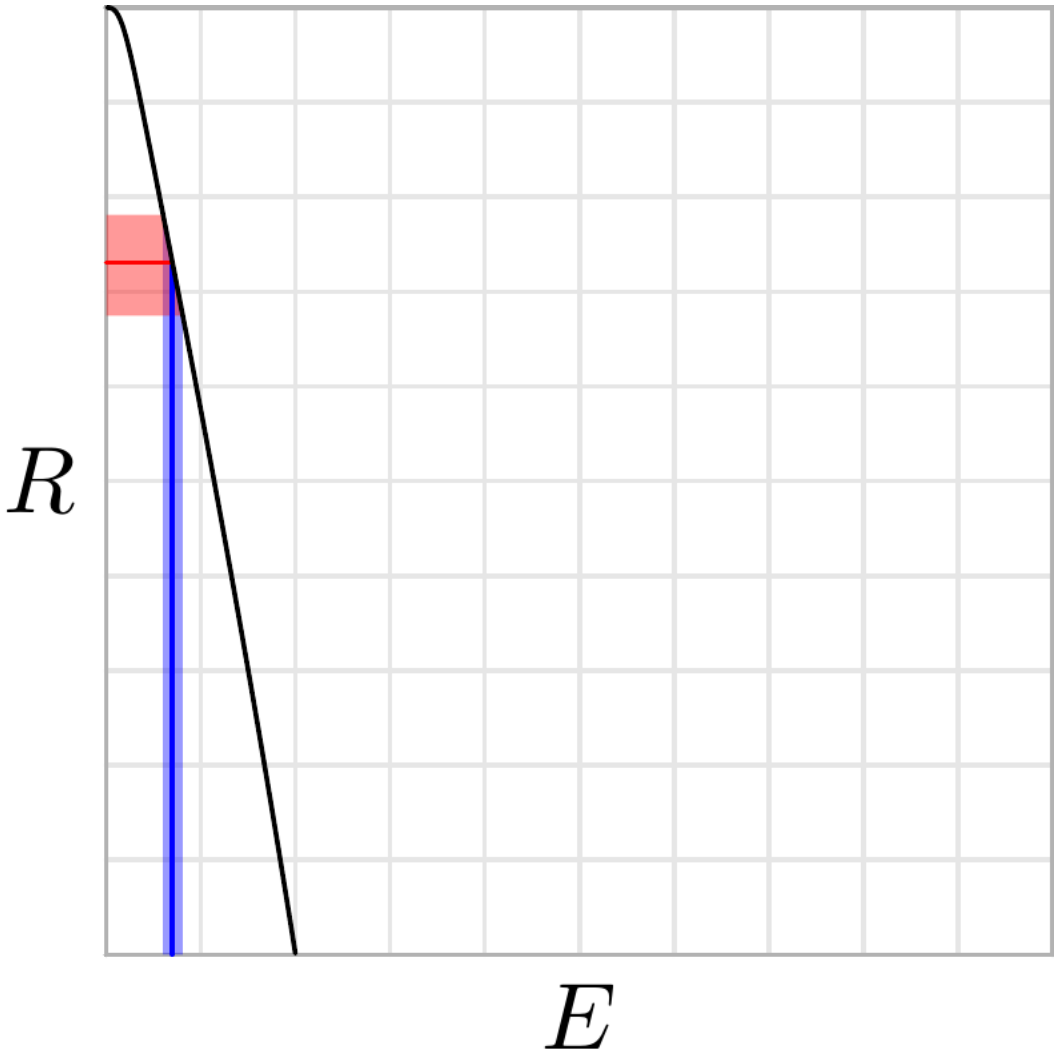
Clustify

Elusion Sampling

- Synonyms:
  - Null Set
  - Elusion Set
  - Discard Set
  - Negatives
- Problems
  - Bias (no fix)
  - Sample Size

Clustify

Recall From Elusion



Clustify

Elusion Sampling

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + (N-n)E}$$

$$\delta R = 1.96\sqrt{\frac{E(1-E)}{S}}\frac{(N-n)\mathrm{TP}}{[\mathrm{TP} + (N-n)E]^2}$$

$$\delta R_{\max} = \frac{0.98}{\sqrt{S}}\frac{9}{8\sqrt{3}}\sqrt{\frac{N-n}{\mathrm{TP}}}$$

$$S = \frac{0.96}{\delta R_{\max}^2}\frac{27}{64}\frac{N-n}{\mathrm{TP}}$$

Clustify

Elusion vs. Direct Method

R	$S_{\text{Elusion}}/S_{\text{Direct}}$
0.90	0.47
0.80	0.53
0.70	0.60
0.60	0.70
0.50	0.84
0.40	1.05
0.30	1.41
0.20	2.11
0.10	4.22

Lies, Damn Lies, and Statistics

- Inclusive Emails
- Search Terms

Clustify