Evaluating Recommenders with Distributions

RecSys Workshop on Perspectives on Evaluation

Michael Ekstrand, *Boise State University / People and Information Research Team*
Ben Carterette, *Spotify*
Fernando Diaz, *Google*
(all authors contributed equally and are randomly ordered)

Typical Evaluation

1. Generate ranking for each user
2. Compute metric for each ranking
3. Take average
4. Perform significance test*

* Or don’t, see Ihemelandu & Ekstrand’s paper

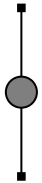
What's Wrong?

- Only considers one perspective (usually)
- Only considers one metric (usually)
- Treats users as interchangeable
- Collapses varied experience into one measurement
- Collapses time into one snapshot

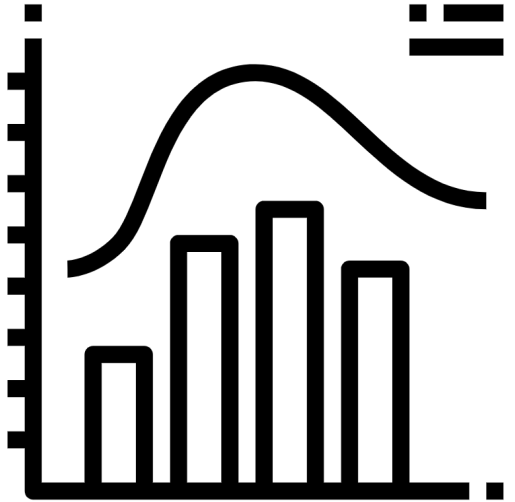
There's no such thing as an "average user"!

Consider Distributions

Point Estimate



Interval Estimate



Distribution

Sources of Uncertainty

Distributions are over some uncertainty or randomness.

Sources:

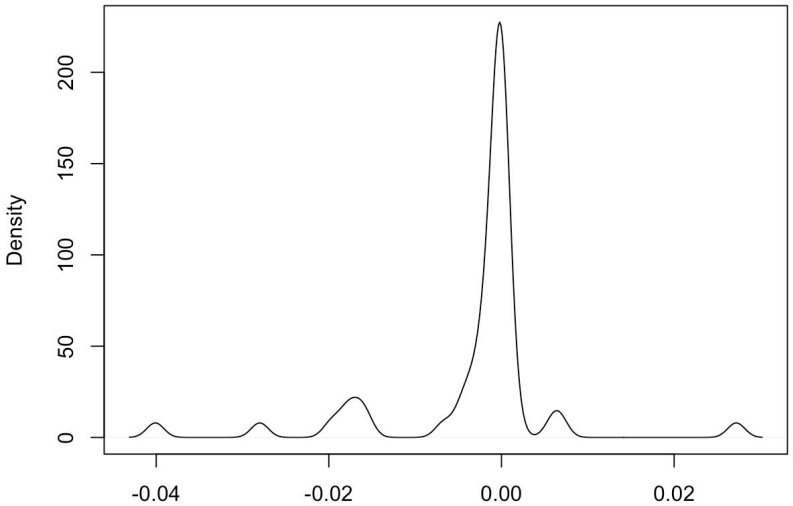
- Data annotation/labels
- Data splitting
- Arrival of information requests, items
- Uncertainty in modeling users/intents/relevance
- Stochastic ranking/retrieval policies (e.g. explore-exploit)

Distribution of Utility

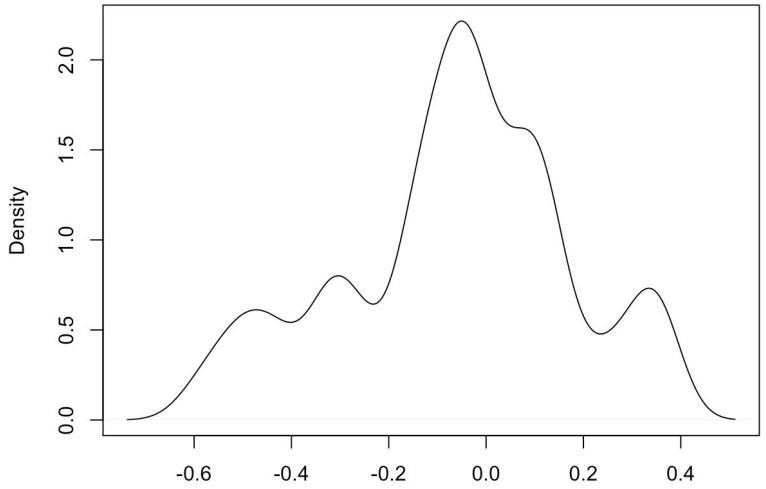
- what does the distribution of effectiveness look like across a population of ***consumers*** with diverse backgrounds and needs?
- what does the distribution of effectiveness look like across a population of ***providers*** with diverse backgrounds and needs?

Distribution of Differences

The distribution of differences can tell us a lot that a mean difference or a p-value cannot.



p = 0.03



p = 0.06

Difference in Distributions

We have a distribution of e.g. utility. Now what?

Compare to **current system** (e.g. in A/B test)

- Does new system distribution better meet goals?
- Distribution replacement for independent two-sample tests

Compare to **ideal target** (example: Expected Exposure)

- Compare graphically or numerically (e.g. K-L divergence)
- Which system-under-test is closest to ideal?

Distribution over Repeated Runs

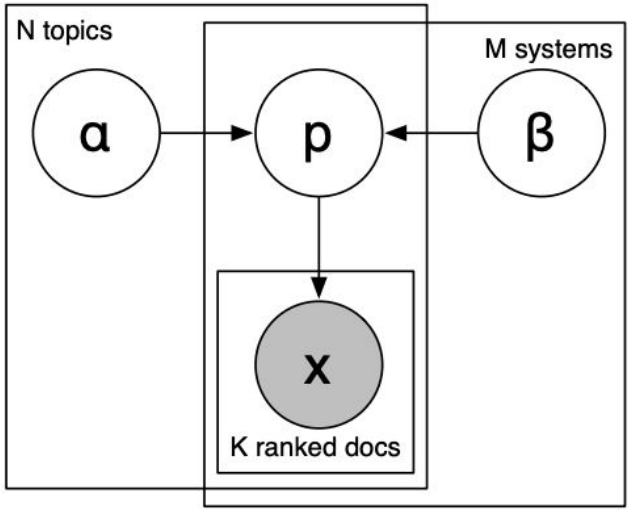
- simulation-based evaluation includes a variety of randomness in users, items, dynamics, and algorithms
- what is the distribution of impact over multiple runs of the simulation?
- what does this mean in unsimulated environments (e.g. production systems)?

Graphical Models and Posterior Distributions

Graphical models provide a way to explicitly model factors related to evaluation.

In this model, “topic” (user request) and “system” (search/recommendation algorithm) influence a coin flip that determines observed result relevance x.

Posterior distributions of alpha and beta can be used to understand differences between requests and differences between systems.



Tools

- Graphical inspection (plot the distributions!)
- Multiple statistics
- Distribution-oriented metrics (e.g. EE loss)
- Confidence measures (for system scores, feedback, etc.)
- Monte Carlo simulations

Lots of research needed!

Open Research

- How do we quantify relevant uncertainty into distributions?
- How do we leverage uncertainty for better systems?
- What statistics are useful for capturing distributional impact?
- How do we think about and present distributional results in ways that support effective decision-making?

Outcomes

- No user or producer left behind
- Understanding who benefits from new work
- Holistic understanding of system behavior and impact

Make recommendation good for everyone it affects.

Photo by Jon Tyson on Unsplash