Heterogeneity in capture-recapture

James Johndrow
Stanford University

NASEM workshop on human trafficking prevalence estimation
April 8, 2019

Outline

- Review: how capture-recapture estimates are made from multiple incomplete lists of victims
- Heterogeneity: when the probability of capture varies on an individual level
- Problem: poor performance of capture-recapture estimates even when the total population size is idenifiable
- Cause: high risk/variance
- Proposed solution: estimation of the “observable” population size
- Some conclusions

Capture-recapture follows record linkage

List 1

ID	Name	Loc	Date	EntID
1	John Smith	A	11/25	1
2	Jane Wang	B	12/15	2
3	Anna Lopez	A	11/12	3
4	John Smith	A	11/25	4
5	**Alex Brown**	**B**	**12/1**	**5**

List 2

ID	Name	Loc	Date	EntID
50	Anna Lopez	A	11/21	3
51	Jane Wang	B	12/15	2
52	John Smith	A	11/25	4
53	John Smith	A	11/25	1
54	**Emma Green**	**A**	**12/1**	**6**

Output of Record-Linkage

EntID	List1	List 2	…
1	1	1	…
2	1	1	
3	1	1	
4	1	1	
5	1	0	
6	0	1	

Capture-recapture follows record linkage

List 1

ID	Name	Loc	Date	EntID
1	John Smith	A	11/25	1
2	Jane Wang	B	12/15	2
3	Anna Lopez	A	11/12	3
4	John Smith	A	11/25	4
5	**Alex Brown**	**B**	**12/1**	**5**

List 2

ID	Name	Loc	Date	EntID
50	Anna Lopez	A	11/21	3
51	Jane Wang	B	12/15	2
52	John Smith	A	11/25	4
53	John Smith	A	11/25	1
54	**Emma Green**	**A**	**12/1**	**6**

Output of Record-Linkage

EntID	List1	List 2	…
1	1	1	…
2	1	1	
3	1	1	
4	1	1	
5	1	0	
6	0	1	

"capture histories"

$$x_i \in \{0,1\}^T$$

$$x_{it} = \begin{cases} 1 & \text{entity } i \text{ appears on list } t \\ 0 & \text{else} \end{cases}$$

Capture-recapture follows record linkage

List 1

ID	Name	Loc	Date	EntID
1	John Smith	A	11/25	1
2	Jane Wang	B	12/15	2
3	Anna Lopez	A	11/12	3
4	John Smith	A	11/25	4
5	**Alex Brown**	**B**	**12/1**	**5**

List 2

ID	Name	Loc	Date	EntID
50	Anna Lopez	A	11/21	3
51	Jane Wang	B	12/15	2
52	John Smith	A	11/25	4
53	John Smith	A	11/25	1
54	**Emma Green**	**A**	**12/1**	**6**

"capture histories"

$$x_i \in \{0,1\}^T$$

$$x_{it} = \begin{cases} 1 & \text{entity } i \text{ appears on list } t \\ 0 & \text{else} \end{cases}$$

Capture-Recapture

Goal

Estimate the number of individuals with capture history

$$x_i = (0,0,0,\ldots,0)$$

i.e. the number not observed (or “captured”) on any list, or equivalently, the total population size

$$K$$

Sufficient statistics

In general we will have **sufficient statistics**

In the cases I consider here, they are

$$n_s, \quad s = 1, \ldots, T$$

$$n_s \equiv \sum_{i=1}^{n} \mathbf{1}\{i \text{ appears on } s \text{ lists}\}$$

Simple Estimator

Simplest case: suppose that

$$X_{is} \perp\!\!\!\perp X_{it}, \quad s \neq t,$$
$$\mathbb{P}[X_{it} = 1] = q \;\forall\; i, t$$

Simple Estimator

Simplest case: suppose that

$$X_{is} \perp\!\!\!\perp X_{it}, \quad s \neq t,$$
$$\mathbb{P}[X_{it} = 1] = q \ \forall \ i, t$$

Then the **conditional likelihood** of the observed data is

$$L(n_1, \ldots, n_T \mid n, p) = \frac{n!}{\prod_s n_s!} \prod_{s=1}^{T} \frac{\nu_s}{1 - \nu_0}$$
$$\nu_s = \binom{T}{s} q^s (1 - q)^{T-s}$$

Simple Estimator

Procedure for estimating K:

1. Estimate $\hat{q}$ by maximum likelihood on the observed data
2. Make the **Horvitz-Thompson** estimate of K

Estimated probability of appearing on at least one list

$$\hat{K} = \frac{n}{1 - (1 - \hat{q})^T} = \frac{n}{\hat{p}}$$

Most capture-recapture estimators are closely related to this procedure (e.g. log-linear modeling)

Assumptions

This relies on three basic assumptions:

1. No net in/out migration (**closed population)**
2. No individual-level variability in the probability of being captured (**homogeneous capturability)**
3. All lists have the same probability of capture

Assumptions

This relies on three basic assumptions:

1. No net in/out migration (**closed population)**
2. No individual-level variability in the probability of being captured (**homogeneous capturability)**
3. All lists have the same probability of capture

Assumptions

This relies on three basic assumptions:

1. No net in/out migration (**closed population)**
2. No individual-level variability in the probability of being captured (**homogeneous capturability)**
3. All lists have the same probability of capture

Strategies to address violations of this assumption

- Typically dealt with by **stratification** over time and space
- This can result in strata with few observations, requires subject area expertise or statistical testing to define strata
- Alternative: **model** the heterogeneity

Aside: ongoing project on selective inference adaptive stratification using tests of heterogeneity. Ask me if interested!

Capture Heterogeneity

Capture heterogeneity is often studied in the context of a simple model called M_h

$$\mathrm{pr}[X_{it} = 1] = q_i, \quad q_i \sim H$$

Capture Heterogeneity

Capture heterogeneity is often studied in the context of a simple model called M_h

$$\text{pr}[X_{it} = 1] = q_i, \quad q_i \sim H$$

The conditional likelihood is the same, except that now

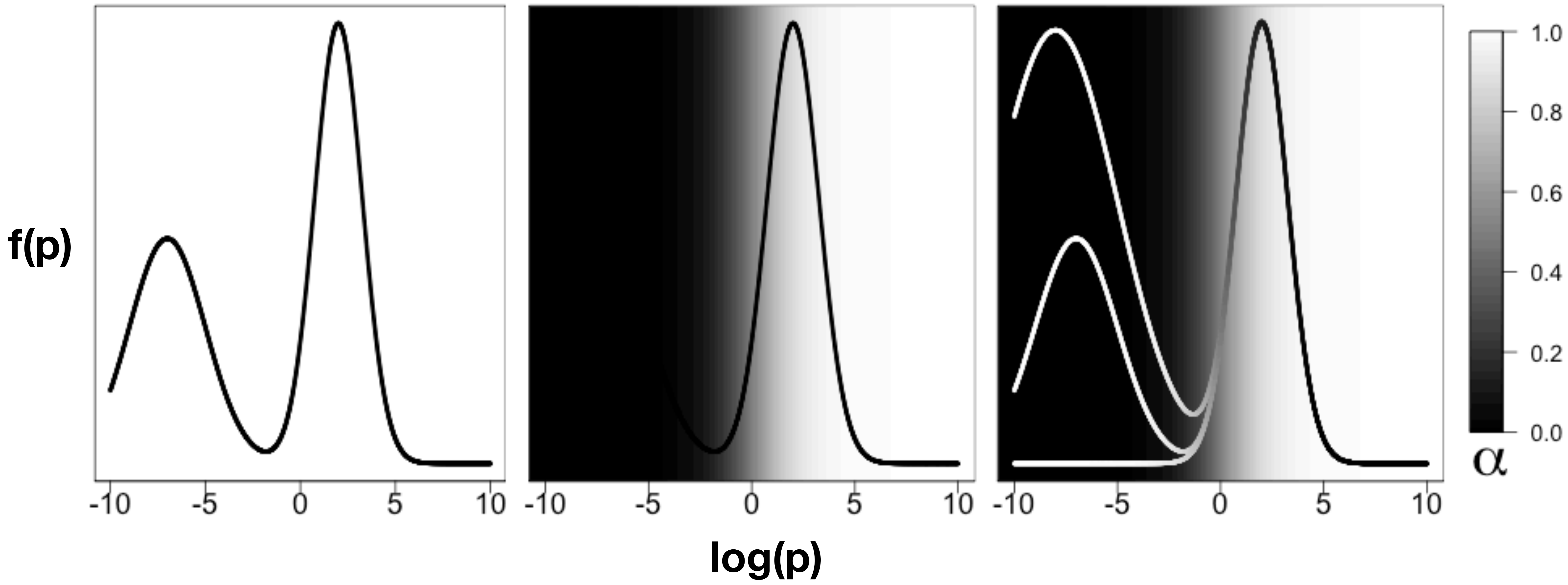
$$\nu_s = \binom{T}{s} \int_0^1 q^s (1-q)^{T-s} H(dq)$$

and the corresponding estimate of the total population size is

$\hat{K} = \dfrac{n}{\mathbb{E}_{\hat{F}}(P)}$ where $P_i \sim F$ is the probability that unit *i* is observed on at least one list

Capture Heterogeneity: What We Know

Basic problem: in the presence of capture heterogeneity, K is **not identifiable** without further restrictions*



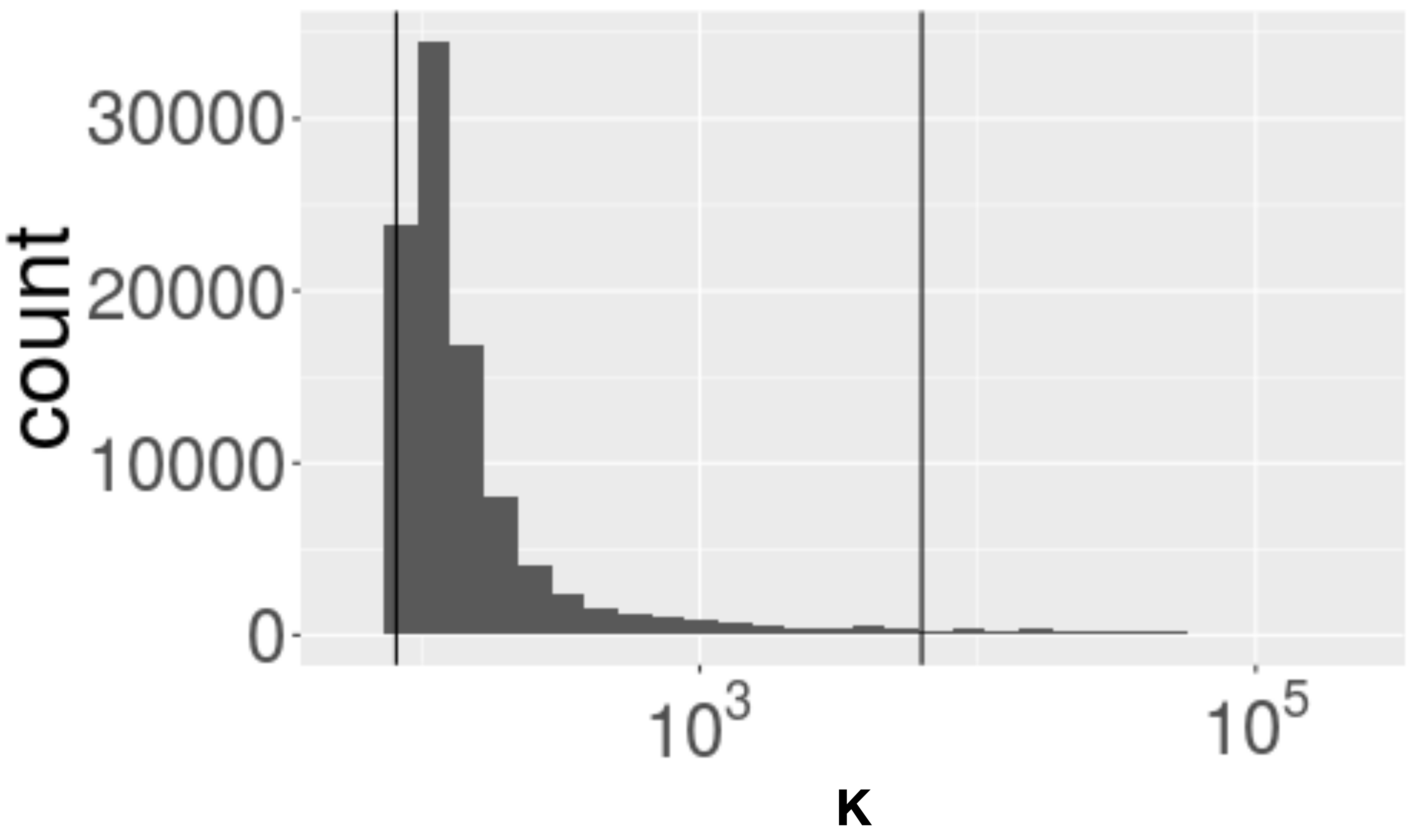
Bottom line: one typically needs to restrict H to be in some family of distributions (cannot nonparametrically estimate H)*

**Two known identifiable families: the Beta family and discrete mixtures (with the "correct" number of components)*

Fitting a Simple Model

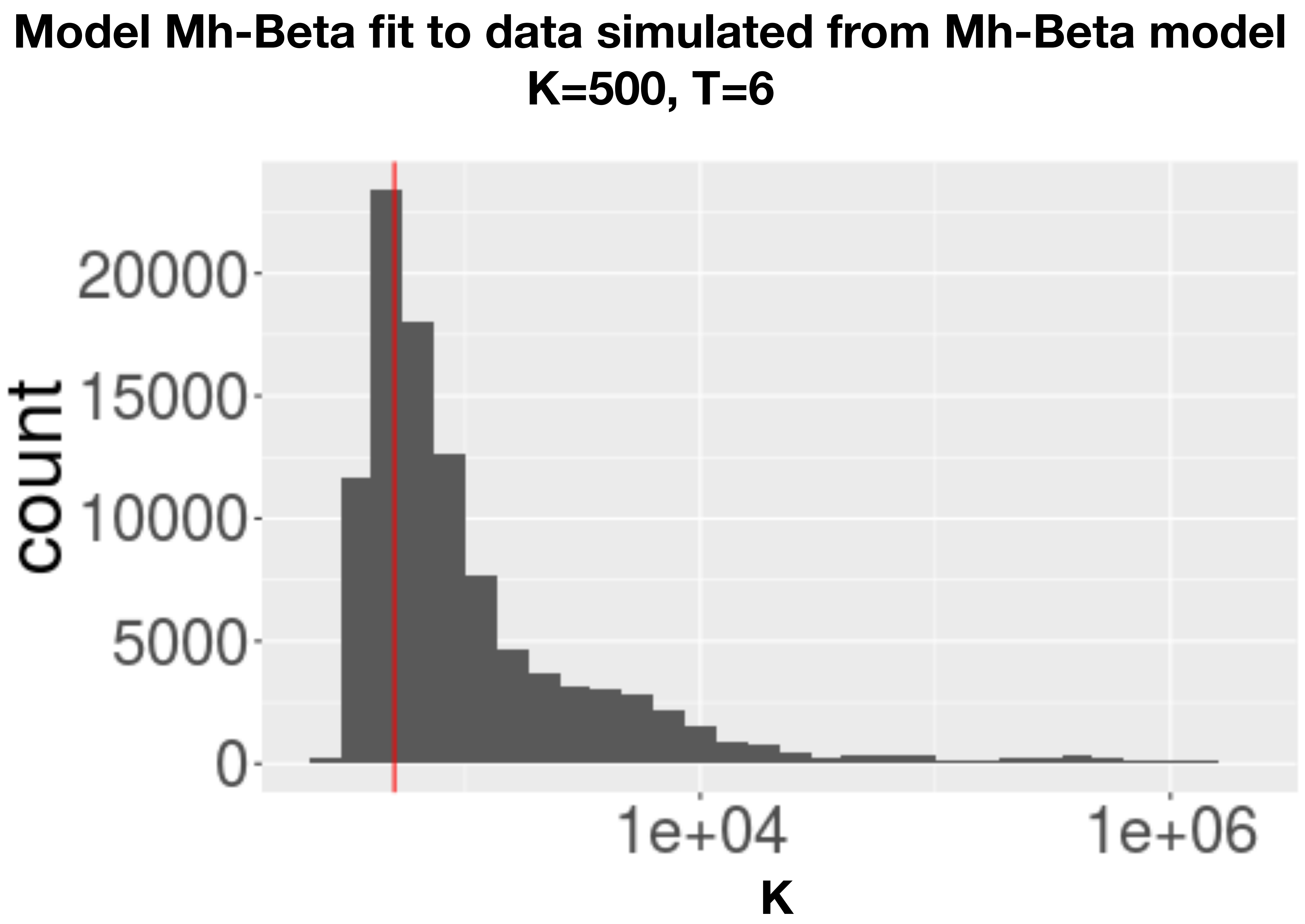
Trying to fit identifiable families to data, we realized that confidence/ credible intervals were still enormous even though *K* was fully identifiable

Model Mh-Beta fit to snowshoe hare data (n=77, T=6)



Fitting a Simple Model

Trying to fit identifiable families to data, we realized that confidence/ credible intervals were still enormous even though *K* was fully identifiable



A Simple Idea

Basic idea: huge variance is possibly caused by the fitted distribution placing mass near zero

Intuition: when the population consists mostly of individuals who are **nearly invisible** to the sampling design, our uncertainty about K explodes

Recall that in presence of heterogeneity, the Horvitz-Thompson estimator becomes

$$\hat{K} = \frac{n}{\mathbb{E}_{\hat{F}}(P)}$$ where $P_i \sim F$ is the probability that unit *i* is observed **on at least one list**

A Simple Idea

Basic idea: huge variance is possibly caused by the fitted distribution placing mass near zero

Intuition: when the population consists mostly of individuals who are **nearly invisible** to the sampling design, our uncertainty about K explodes

Recall that in presence of heterogeneity, the Horvitz-Thompson estimator becomes

$$\hat{K} = \frac{n}{\mathbb{E}_{\hat{F}}(P)}$$ where $P_i \sim F$ is the probability that unit *i* is observed **on at least one list**

Note: the results assume that we **observe** the capture probabilities; therefore they are **optimistic** about how difficult the problem is

A Simple Idea

Basic idea: huge variance is possibly caused by the fitted distribution placing mass near zero

Intuition: when the population consists mostly of individuals who are **nearly invisible** to the sampling design, our uncertainty about K explodes

Recall that in presence of heterogeneity, the Horvitz-Thompson estimator becomes

$$\hat{K} = \frac{n}{\mathbb{E}_{\hat{F}}(P)}$$ where $P_i \sim F$ is the probability that unit *i* is observed **on at least one list**

Even if we knew F

$$\mathbb{E}\left\{\frac{n}{\mathbb{E}_F(P)}\right\} = \frac{K\mathbb{E}_F(P)}{\mathbb{E}_F(P)} = K$$

$$\mathrm{var}\left\{\frac{n}{\mathbb{E}_F(P)}\right\} = K\frac{1-\mathbb{E}_F(P)}{\mathbb{E}_F(P)} \propto K\frac{1}{\mathbb{E}_F(P)}.$$

Possible Solution: Changing the Inferential Objective

Possible solution: Estimate the population that is **minimally visible** to our sampling mechanism

We call this

$$K_\alpha = \sum_{i=1}^{N} \mathbf{1}\{P_i > \alpha\}$$

Possible Solution: Changing the Inferential Objective

Possible solution: Estimate the population that is **minimally visible** to our sampling mechanism

We call this

$$K_\alpha = \sum_{i=1}^{N} \mathbf{1}\{P_i > \alpha\}$$

Why it might work:

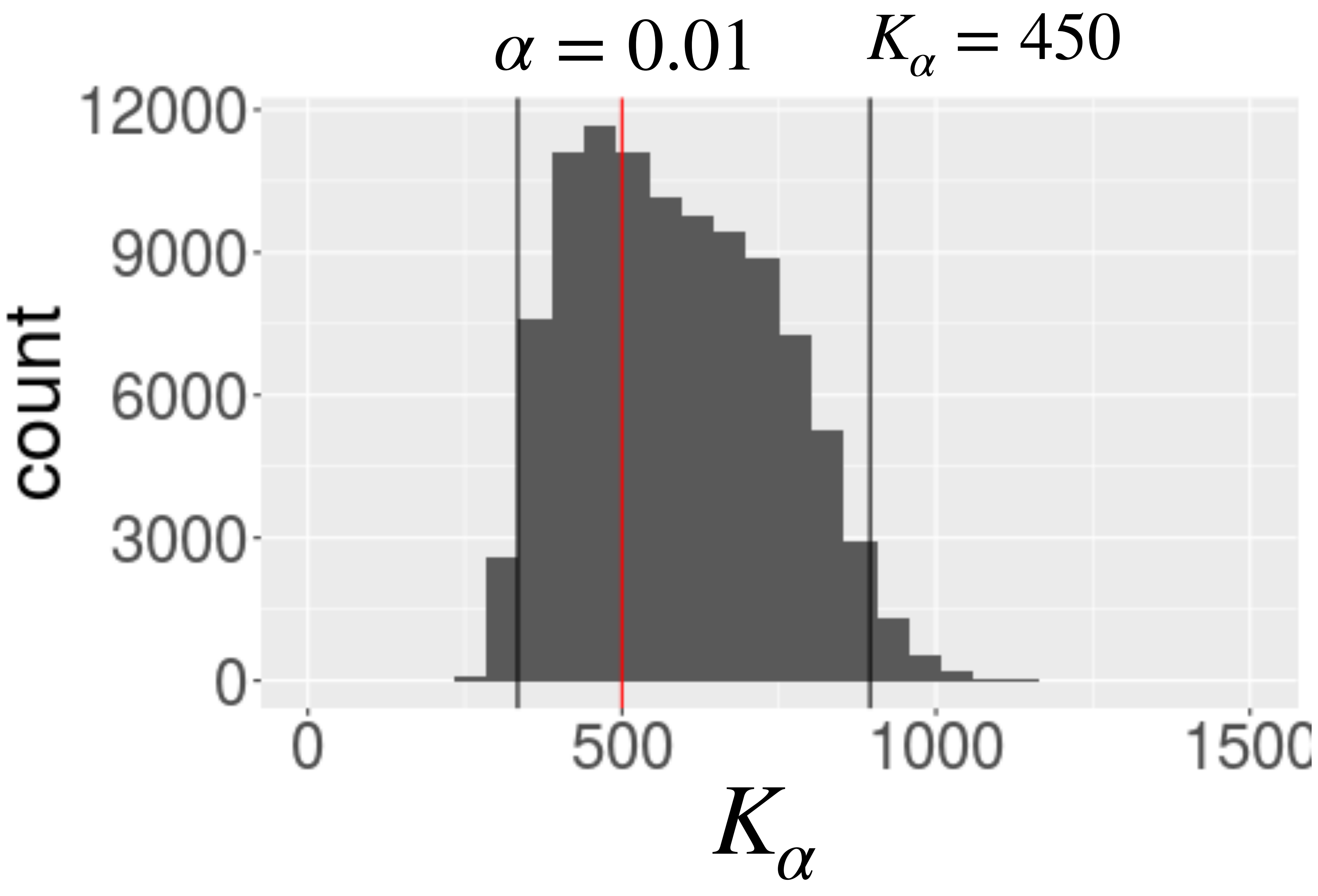
$$\hat{K}_\alpha = \frac{n_\alpha}{\mathbb{E}_{\hat{F}}(P \mid P > \alpha)} = \frac{n_\alpha}{\mathbb{E}_{\hat{F}_\alpha}(P)}, \text{ where } n_\alpha = \sum_{i=1}^{N} \max_t X_{it} \mathbf{1}\{P_i > \alpha\}$$

and $\mathbb{E}_{\hat{F}_\alpha}(P) > \alpha$ so the variance cannot get too big (at least when F is known)

Empirically, It Works

Model Mh-Beta fit to data simulated from Mh-Beta model
K=500, T=6

Posterior samples of K_α



Theoretical Risk Bounds

The same thing happens when we estimate *F*
For example, if *F* is a Beta(1,b) distribution and we estimate b by maximum likelihood, the asymptotic risk goes to infinity at a linear rate in b:

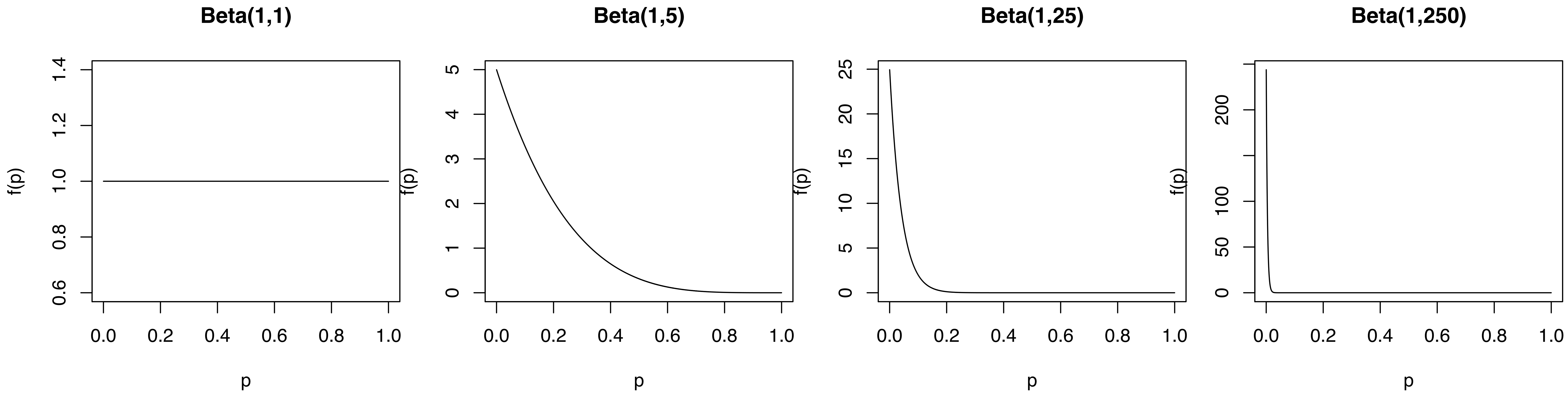
$$\text{Risk}(\hat{K}) = \frac{b^2(1+b)}{b^2+(b+1)^2}K + Kb$$

$$\propto Kb$$

Theoretical Risk Bounds

The same thing happens when we estimate *F*
For example, if *F* is a Beta(1,b) distribution and we estimate b by maximum likelihood, the asymptotic risk goes to infinity at a linear rate in b:

$$\text{Risk}(\hat{K}) = \frac{b^2(1+b)}{b^2 + (b+1)^2}K + Kb$$
$$\propto Kb$$



Theoretical Risk Bounds

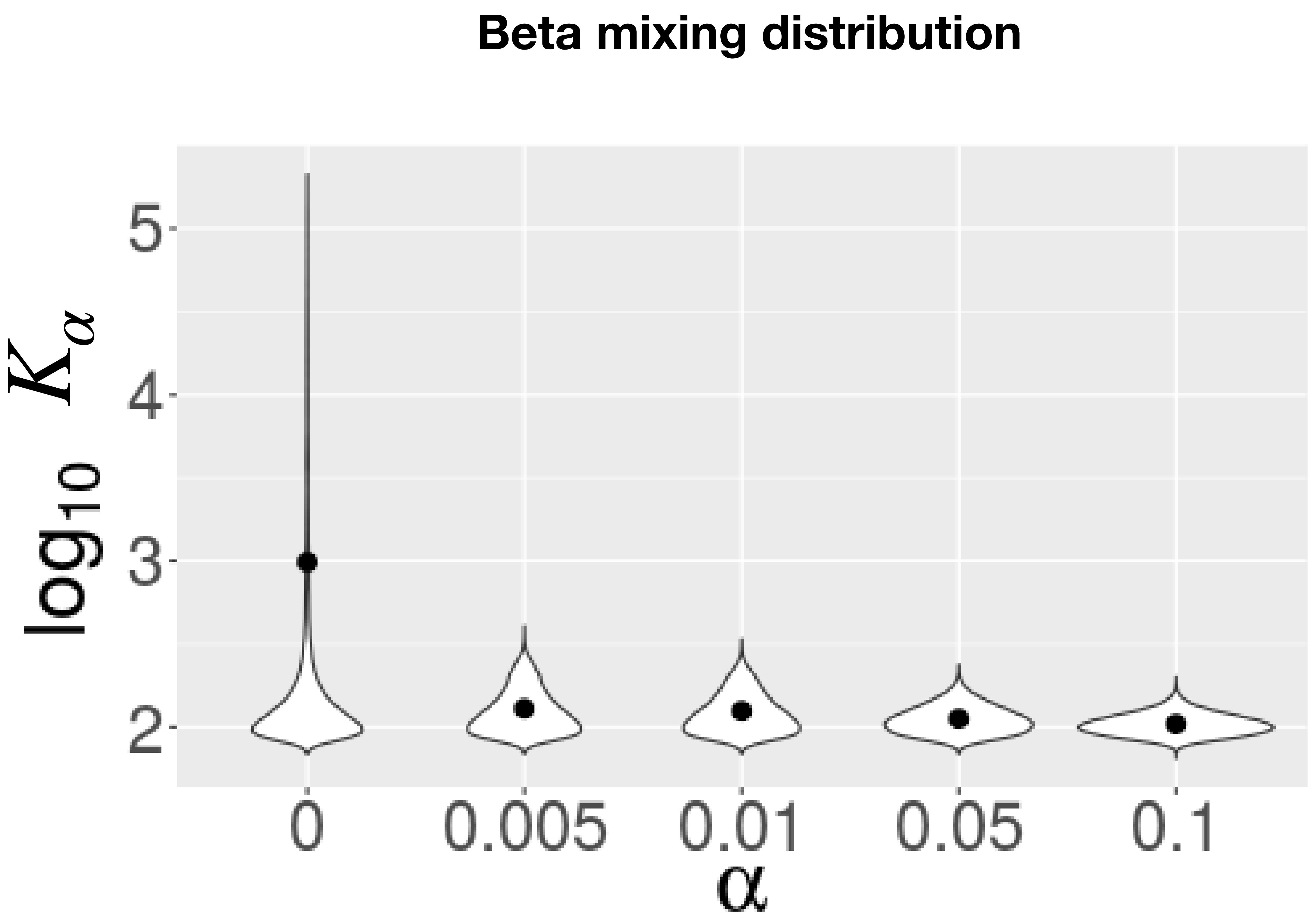
On the other hand, the risk for K_α remains bounded

If *F* is a Beta(1,b) distribution and we estimate b by maximum likelihood, the asymptotic risk for K_α is bounded as a function of b

$$\textbf{Risk}(\hat{K}_\alpha) \leq \frac{b^2(1+b)}{b^2+(b+1)^2}\frac{K(1-\alpha)^{b+2}}{(1+b\alpha)^3} + \frac{K(1-\alpha)^b}{(1+\alpha b)}\left\{b+1-(1-\alpha)^b(b\alpha+1)\right\}$$

$$\propto K(1-\alpha)^b$$

Back to the Hares



Estimates of K_α are much more plausible, **even as estimates of K.**

Takeaways

- caveat that estimates only pertained to people with non-zero probability of capture replaced with **minimal visibility**
- Give estimates with **practically useful** intervals.
- $\hat{K}_\alpha$ is also useful as a **biased estimator** of *K*.
  - Two possible scenarios: (1) it's not actually that biased or (2) a huge proportion of the population is invisible to the sampling design and we couldn't have estimated them anyway.
- Practical: try many mixing distributions
- New/ongoing project: selective inference for adaptive stratification based on tests for heterogeneity

Collaborators

Thanks!

Questions?

See:

Johndrow, J. E., K. Lum, and D. Manrique-Vallier. “Low risk population size estimates in the presence of capture heterogeneity." *Biometrika* (forthcoming)