lecture 02

ml

maximum likelihood; invariance principle; predictive distribution

review

an example exam question

From the homework:

Simplify the following: $\log\left(\prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{(1-y_i)}\right)$

Let $\ell(\pi) = S\log(\pi) + (n-S)\log\big(1-\pi\big)$ for $0 < \pi < 1$, where S and n (with $0 \le S \le n$) are fixed numerical constants. Find $\frac{d\ell(\pi)}{d\pi}$ and $\frac{d^2\ell(\pi)}{d\pi^2}$.

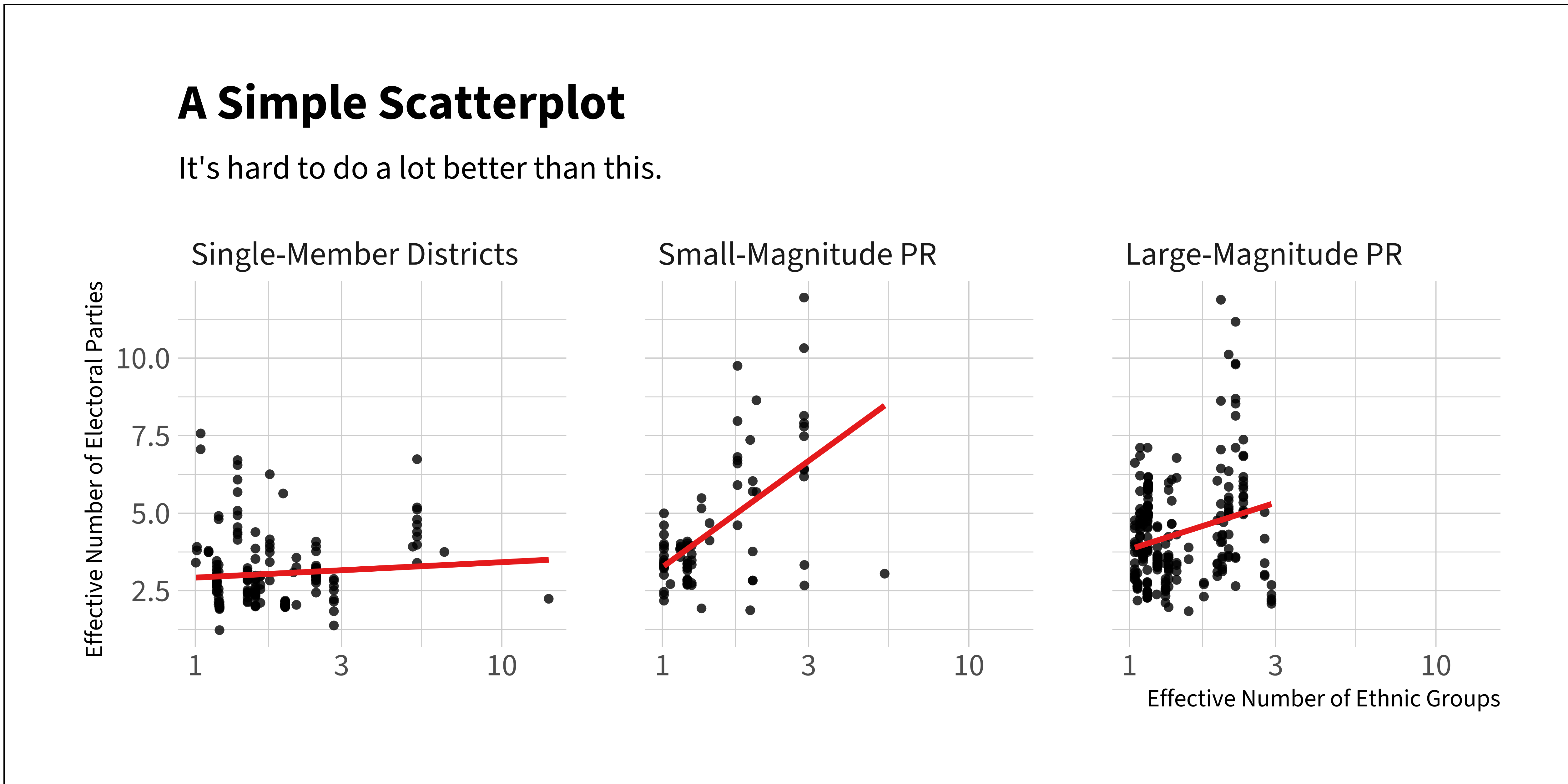
A variant on the exam:

Let $\ell(\pi) = \log\left[\pi^k(1-\pi)^{N-k}\right]$ for $0 < \pi < 1$ and $0 < k < N$. Simplify the right-hand side and find $\dfrac{d\ell(\pi)}{d\pi}$.

I would say this question is "medium-easy" (after all, it's review material). But it's somewhat tedious—I'd give you 5 minutes to work through it.

maximum likelihood

Hypothesis 1: Social heterogeneity increases the number of parties, but only when electoral institutions are sufficiently permissive.[4]



example

the toothpaste cap problem

0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

150 tosses

8 tops

How can we estimate the chance of a top?

Bernoulli

either

1 or **0**

"x" for theory
"y" for applied

$$\Pr(y = 1) = \pi \qquad \Pr(y = 0) = 1 - \pi$$

A.1 Bernoulli Distribution

Table A.1: Summary of the Bernoulli distribution

Property	Description
Name	Bernoulli distribution
Notation	$\text{Bernoulli}(\pi)$
Parameters	$0 \leq \pi \leq 1$: probability of success
Support	$x \in \{0, 1\}$
PMF	$f(x) = \pi^x (1 - \pi)^{1-x}$
CDF	$F(x) = 0$ for $x < 0$; $1 - \pi$ for $0 \leq x < 1$; 1 for $x \geq 1$
Mean	π
Variance	$\pi(1 - \pi)$
Special cases	$\text{Binomial}(n = 1, \pi)$ is $\text{Bernoulli}(\pi)$.

let's try this out!
$f(1) = ?$
$f(0) = ?$

A Principled Way

Assume the observed data (i.e., those 0s and 1s) are draws from a Bernoulli distribution.

What value of π would be

most likely

to generate those data?

0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

This is pretty easy to work out!

Remember, we have

0 0 1 0 0...

$$\begin{array}{ccccccccccc}
\Pr(y_1 = 0) & \times & \Pr(y_2 = 0) & \times & \Pr(y_3 = 1) & \times & \Pr(y_4 = 0) & \times & \Pr(y_5 = 0) & \times & \cdots \\
f(0) & \times & f(0) & \times & f(1) & \times & f(0) & \times & f(0) & \times & \cdots \\
(1-\pi) & \times & (1-\pi) & \times & \pi & \times & (1-\pi) & \times & (1-\pi) & \times & \cdots
\end{array}$$

The number of πs is $k = \sum_{i=1}^{N} y_i$.

The number of $(1-\pi)$ is $k = N - k$.

each 1 contributes a π to the product (e.g., 8)

$$\text{the likelihood: } L(\pi) = \boxed{\pi^k}\boxed{(1-\pi)^{(N-k)}}$$

each 0 contributes a $(1-\pi)$ to the product (e.g., 142)

ASIDE

There's a more general way.

Remember, we have

`y1 y2 y3 y4 y5...`

f(x) can be pmf or pdf!

$$f(y_1) \quad \times \quad f(y_2) \quad \times \quad f(y_3) \quad \times \quad f(y_4) \quad \times \quad f(y_5) \quad \times \quad \cdots$$

$$\text{the likelihood: } L(\pi) = \prod_{i=1}^{N} f(y_i)$$

$$= \prod_{i=1}^{N} \pi^{y_i}(1-\pi)^{1-y_i}$$

Bernoulli model...can use other pmfs or pdfs

$$= \textit{(could simplify as before if we wanted...)}$$

the likelihood: $L(\pi) = \pi^k (1-\pi)^{(N-k)}$

Maximize the likelihood w.r.t. π.

This is the π that would "most likely" generate our data.

MLE Recipe

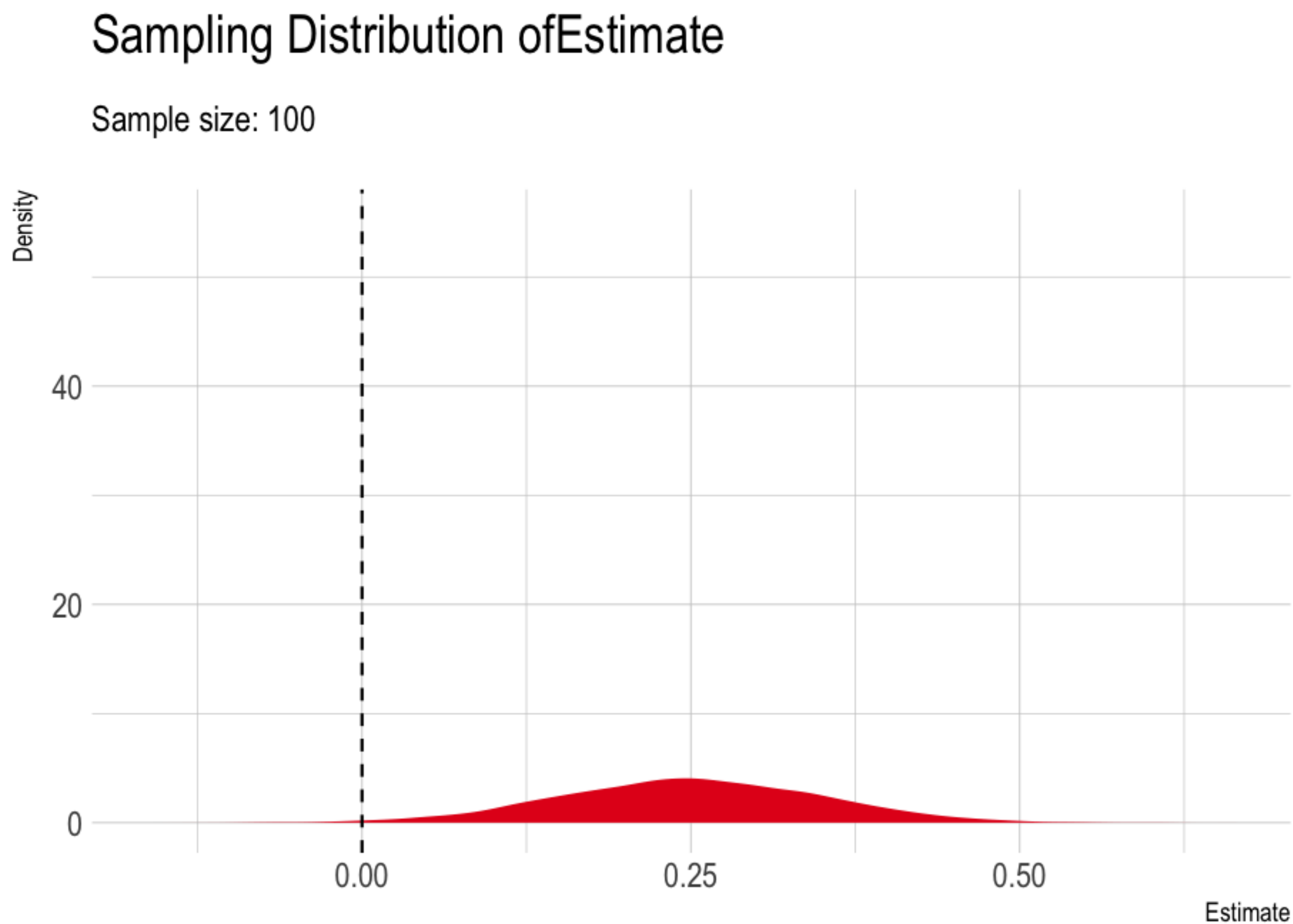
1. Write down the likelihood
2. Take log of the likelihood.
3. Simplify the log-likelihood.
4. Take derivative of the log-likelihood
5. Set derivative equal to zero; set $\pi = \hat{\pi}$.
6. Solve for $\hat{\pi}$.
7. (Check second-order conditions.)

ML
in general

Maximum Likelihood Estimator. Suppose we have iid samples $x_1, x_2, ..., x_N$ from pdf or pmf $f(x;\theta)$. Then the joint density/probability is $f(x;\theta) = \prod_{i=1}^{N} f(x_i;\theta)$ and $\log L(\theta) = \sum_{i=1}^{N} \log\left[f(x_i;\theta)\right]$. The ML estimator $\hat{\theta}$ of θ is $\arg\max \log L(\theta)$.

Consistency. Let $\hat{\theta}_N$ be an estimator of θ based on a sample of size N. Say that $\hat{\theta}_N$ is a **consistent** estimator for θ if $\lim_{N\to\infty} \Pr\left(|\hat{\theta}_N - \theta| \geq \varepsilon\right) = 0$ for every $\varepsilon > 0$.

Consistency of ML Estimators. Suppose an ML estimator $\hat{\theta}$ of θ. Under certain *regularity conditions*, $\hat{\theta}$ is a consistent estimator of θ.


Sampling Distribution ofEstimate
Sample size: 100
Density
40
20
0
0.00
0.25
0.50
Estimate

Poisson Model

Suppose we collect N random samples $y = \{y_1, y_2, ..., y_N\}$ and model each draw as a Poisson random variable. Find the ML estimator of λ.

A.4 Poisson Distribution

Table A.4: Summary of the Poisson distribution

Property	Description
Name	Poisson distribution
Notation	$\text{Poisson}(\lambda)$
Parameters	$\lambda > 0$: rate (mean number of events per unit interval)
Support	$x \in \{0, 1, 2, \ldots\}$
PMF	$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$
CDF	$F(x) = \sum_{k=0}^{\lfloor x \rfloor} \frac{e^{-\lambda}\lambda^k}{k!}$
Mean	λ
Variance	λ
Special cases	- Limit of $\text{Binomial}(n, p)$ as $n \to \infty$, $p \to 0$ with $np = \lambda$ fixed - Distribution of counts in a homogeneous Poisson process - Sum of independent $\text{Poisson}(\lambda_i)$ is $\text{Poisson}(\sum \lambda_i)$

Examples

Number of times legislators travel to their districts

Number of ISIS violent events

Number of out-of-circuit judicial citations

Q: It's "nice" for the support of the model to match the support of the observed data. But does it really matter? When?

invariance property

Invariance Property. Suppose an ML estimator $\hat{\theta}$ of θ and a quantity of interest $\tau = \tau(\theta)$ for any function τ. The ML estimate $\hat{\tau}$ of $\tau = \tau(\theta)$ is $\tau(\hat{\theta})$.

Example 1: $\hat{\pi}$ **to odds**

Example 2: $\hat{\lambda}$ **to SD**

Seems obvious?

OLS; no longer unbiased!

posterior mean; no longer the mean!

Among the most important ideas we see!

optim()

Beta Model

A.2 Beta Distribution

Table A.2: Summary of the beta distribution

Property	Description
Name	beta distribution
Notation	$\text{beta}(\alpha, \beta)$
Parameters	$\alpha > 0$: shape; $\beta > 0$: shape
Support	$x \in (0, 1)$
PDF	$f(x) = \dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$
CDF	$F(x) = I_x(\alpha, \beta)$, the regularized incomplete beta function
Mean	$\dfrac{\alpha}{\alpha+\beta}$
Variance	$\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Special cases	$\text{beta}(1, 1)$ is $\text{uniform}(0, 1)$.

$$L(\alpha, \beta) = \prod_{i=1}^{n} \frac{y_i^{\alpha-1}(1-y_i)^{\beta-1}}{B(\alpha, \beta)}$$

$$\log L(\alpha, \beta) = \sum_{i=1}^{n} \log \frac{y_i^{\alpha-1}(1-y_i)^{\beta-1}}{B(\alpha, \beta)}$$

$$= \sum_{i=1}^{n} \left[\log y_i^{\alpha-1} + \log(1-y_i)^{\beta-1} - \log B(\alpha, \beta)\right]$$

$$= \sum_{i=1}^{n} [(\alpha-1)\log y_i + (\beta-1)\log(1-y_i) - \log B(\alpha, \beta)]$$

$$= \sum_{i=1}^{n} [(\alpha-1)\log y_i + (\beta-1)\log(1-y_i)] - n\log B(\alpha, \beta)$$

$$\log L(\alpha, \beta) = (\alpha-1)\sum_{i=1}^{n}\log y_i + (\beta-1)\sum_{i=1}^{n}\log(1-y_i) - n\log B(\alpha, \beta)$$

$$\log L(\alpha, \beta) = (\alpha - 1) \sum_{i=1}^{n} \log y_i + (\beta - 1) \sum_{i=1}^{n} \log(1 - y_i) - n \log B(\alpha, \beta)$$

```r
# manually typing log-likelihood
ll_fn <- function(theta, y) {
  alpha <- theta[1]
  beta <- theta[2]
  ll <- alpha*sum(log(y)) + beta*sum(log(1 - y)) -
    length(y)*log(beta(alpha, beta))
  return(ll)
}
```

```r
# using dbeta shortcut (good!)
ll_fn <- function(theta, y) {
  alpha <- theta[1]
  beta <- theta[2]
  ll <- sum(dbeta(y, shape1 = alpha, shape2 = beta, log = TRUE))
  return(ll)
}
```

```r
# using dbeta shortcut (good!)
ll_fn <- function(theta, y) {
  alpha <- theta[1]
  beta <- theta[2]
  ll <- sum(dbeta(y, shape1 = alpha, shape2 = beta, log = TRUE))
  return(ll)
}


# use optim()
est <- optim(par = c(2, 2), [ starting values of parameters ]
             fn = ll_fn, [ function to optimize ]
             y = y, [ data ]
             control = list(fnscale = -1), [ maximize! ]
             method = "Nelder-Mead")
```

```
# using dbeta shortcut (good!)
ll_fn <- function(theta, y) {
  alpha <- theta[1]
  beta <- theta[2]
  ll <- sum(dbeta(y, shape1 = alpha, shape2 = beta, log = TRUE))
  return(ll)
}

# make function that fits beta model
est_beta <- function(y) {
  est <- optim(par = c(2, 2), fn = ll_fn, y = y,
               control = list(fnscale = -1),          same as before!
               method = "BFGS") # for >1d problems
  if (est$convergence != 0) print("Model did not converge!") [ automatically check! ]
  res <- list(est = est$par) [ return a *list* of results... only ests for now ]
  return(res)
}

# fit beta model
ml_est <- est_beta(y) [ very compact ]
print(ml_est, digits = 3)
```

(Continue in R—there's a Gist.)

https://gist.github.com/carlislerainey/17f281357ffd91972dce3dc670f35a37

predictive distribution

In my view, the **predictive distribution** is the best way to (1) understand, (2) evaluate, and then (3) improve models. (It's not something you should see reported in papers often, but you should use it internally for sure.)

You can use the predictive distribution as follows:

1. Fit your model with maximum likelihood.

2. Simulate a new outcome variable using the estimated model parameters (i.e., $f(y; \hat{\theta})$). Perhaps simulate several for comparison.

3. Compare the simulated outcome variable(s) to the observed outcome variable. You can use histograms, plots of the ECDF, or other summaries.

Example: Holland (2015)

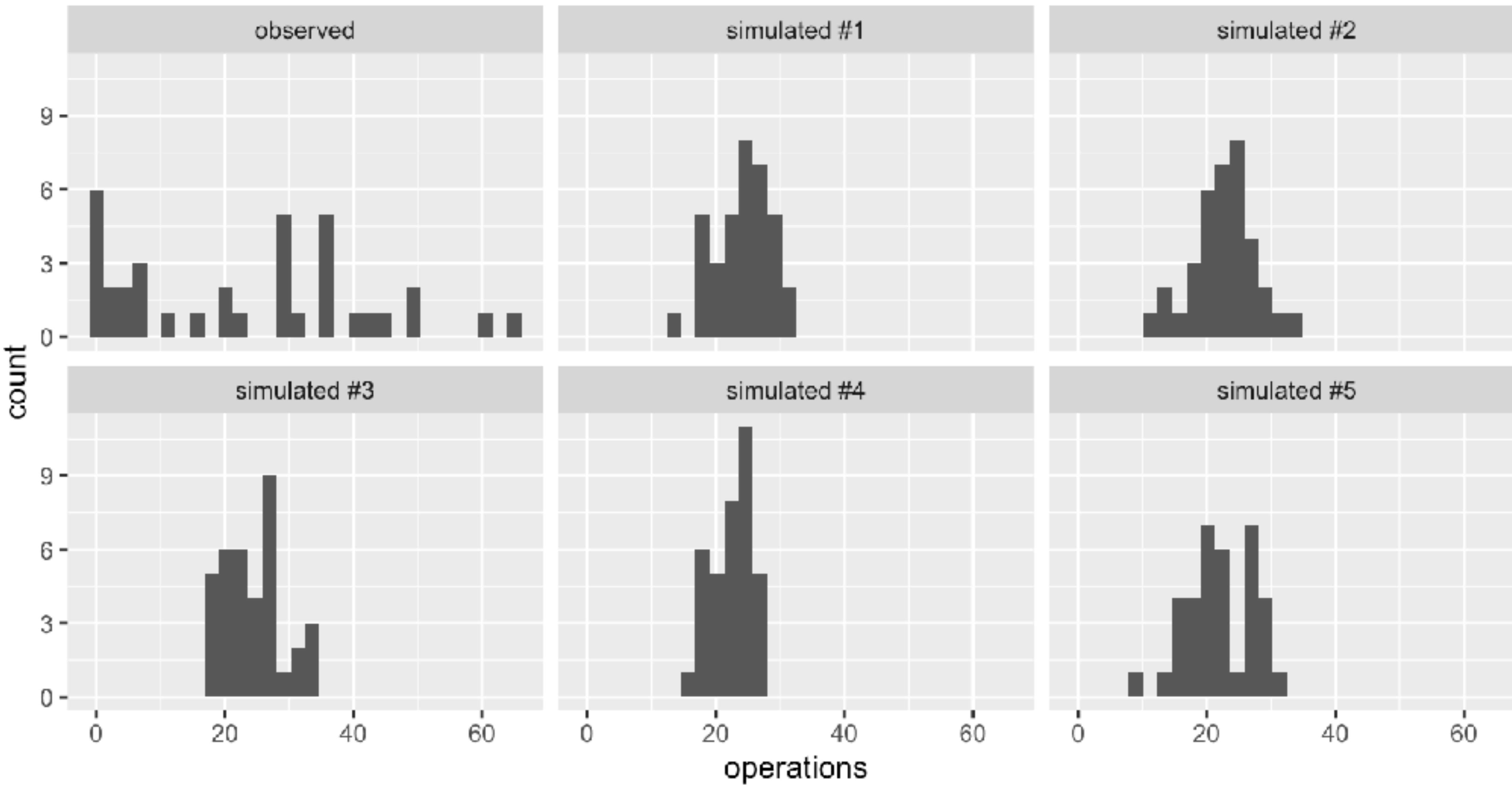
number of enforcement operations in a city

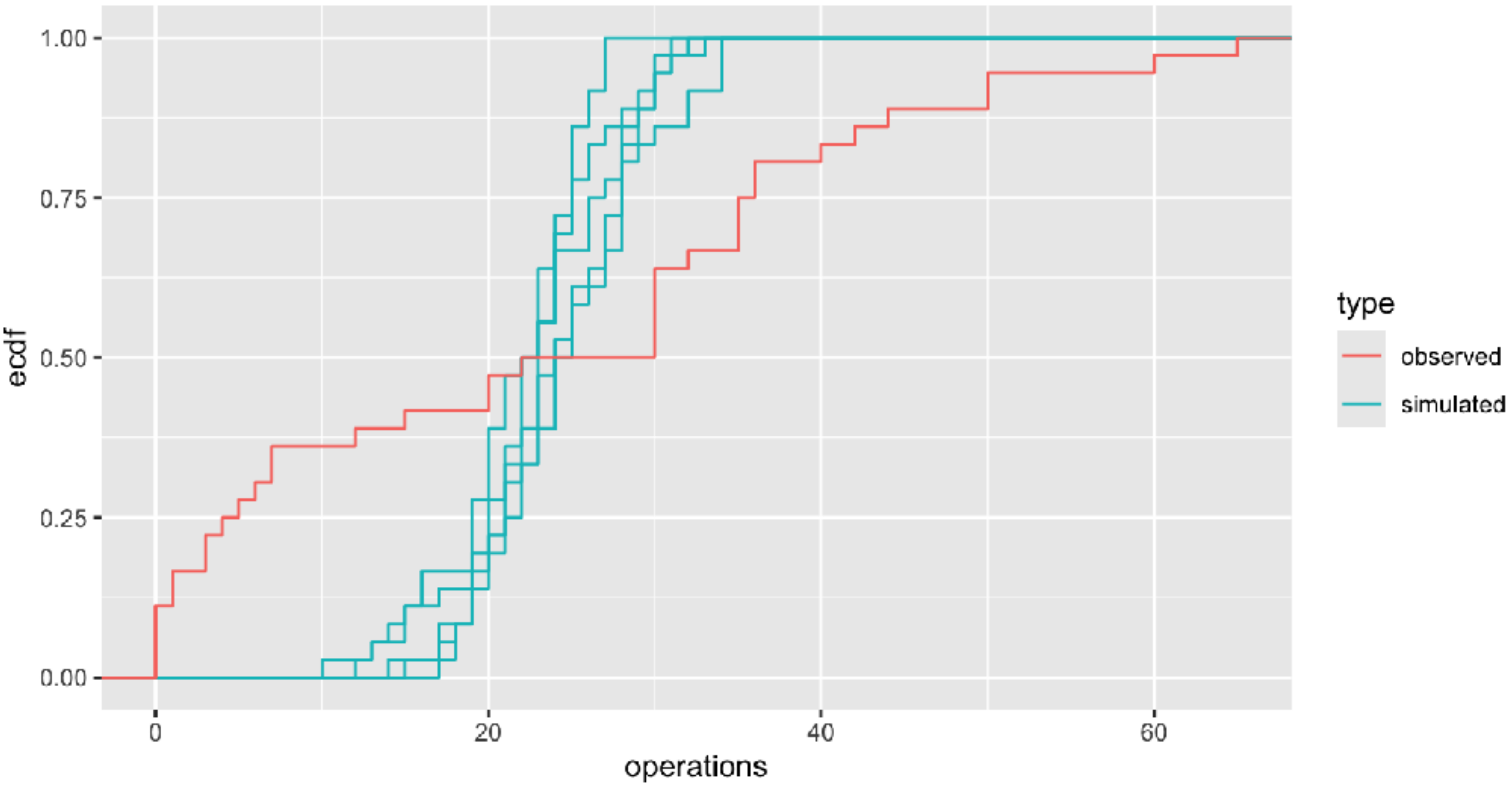
Observed data:

```
7   0  20   1   1  50   3   0  36  32…
```

Simulated data from Poisson model:

```
23  25  20  22  19  19  22   8  14  25…
```


observed
simulated #1
simulated #2
simulated #3
simulated #4
simulated #5
count
operations
0
3
6
9
0
20
40
60



1.00
0.75
0.50
0.25
0.00
ecdf
0
20
40
60
operations
type
observed
simulated

(Continue in R—there's a Gist.)

https://gist.github.com/carlislerainey/c3cb055454a076ab49ebca1eca6b0c0c a bit of fun

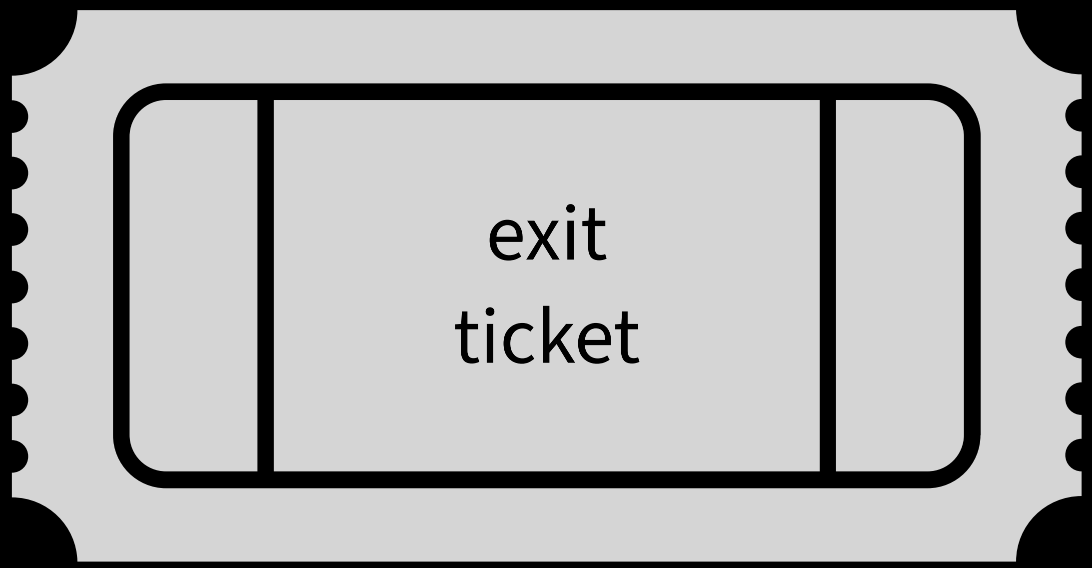
German Tank Problem

Suppose a discrete uniform distribution from 0 to K. The pmf is $f(x;K) = \frac{1}{K+1}, \quad x \in \{0, 1, \ldots, K\}$. Suppose I have a sample of size 3 from the distribution: 276, 159, and 912.

1. Find the ML estimate of K. *Hint: The log-likelihood is discontinuous, so the usual optimization routine might mislead you. But the maximum is immediately apparent once you write out the likelihood. A verbal argument is okay, too!*

2. Find the method of moments estimate of K. *Hint: The mean of this uniform distribution is $\frac{K}{2}$. Set the sample mean equal to the model mean (i.e., $\frac{K}{2} = \text{avg}(x)$) and solve.*

3. Discuss any problems you notice with each estimator.



List three important ideas from today's class. For each, briefly connect it to one or more ideas from last week.