lecture 03

se

sampling distribution; parametric bootstrap; information matrix; delta method; coverage

review

an example exam question

an easy one

```
1  ll_fn <- function(theta, y) {
2    alpha <- theta[1]
3    beta <- theta[2]
4    ll <- sum(dbeta(y, shape1 = alpha, shape2 = beta, log = TRUE))
5    return(ll)
6  }
7
8  est_beta <- function(y) {
9    est <- optim(par = c(2, 2), fn = ll_fn, y = y,
10                 control = list(fnscale = -1),
11                 method = "BFGS") # for >1d problems
12   if (est$convergence != 0) print("Model did not converge!")
13   res <- list(est = est$par)
14   return(res)
15 }
16
17 fit <- est_beta(y)
```

1. On line 4, what does `log = TRUE` do?
2. On line 10, what does `control = list(fnscale = -1)` do?
3. Describe the function `est_beta()`, both (i) the arguments it takes and (ii) the results it returns.
4. Write R code to compute $\hat{\mu} = \frac{\hat{\alpha}}{\hat{\alpha}+\hat{\beta}}$ after running the code above.

Suppose we collect N random samples $y = \{y_1, y_2, ..., y_N\}$ and model these data with a Poisson distribution with pmf $f(y_i; \lambda) = \dfrac{e^{-\lambda}\lambda_i^y}{y_i!}$ for $y_i \in \{0, 1, 2, ...\}$ and $\lambda > 0$. Find the ML estimator of λ.

Can you sketch the process?

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.)

Do you remember the solution?

$$\hat{\lambda} = \text{avg}(y)$$

sampling distribution

We have an estimator $\hat{\theta}$.

This estimator is a **random variable**

because $\hat{\theta}$ is a function of y,
which is random *in some way*.

random sampling randomization probability model

If $\hat{\theta}$ is a random variable,
then it has a distribution.

This distribution is called the
sampling distribution.

The **sampling distribution** is the **most important** statistical concept (by far).

When you run a study, you get an estimate $\hat{\theta}$.

The estimate is random—it comes from a noisy procedure.

We must figure out that noise (SE, p-value, CI).

Example

The Poisson ML Estimate

https://gist.github.com/carlislerainey/ae1e37fe392f69017a4ce5909c4bbd29

three important features

of the sampling distribution

location

approx. unbiased

spread

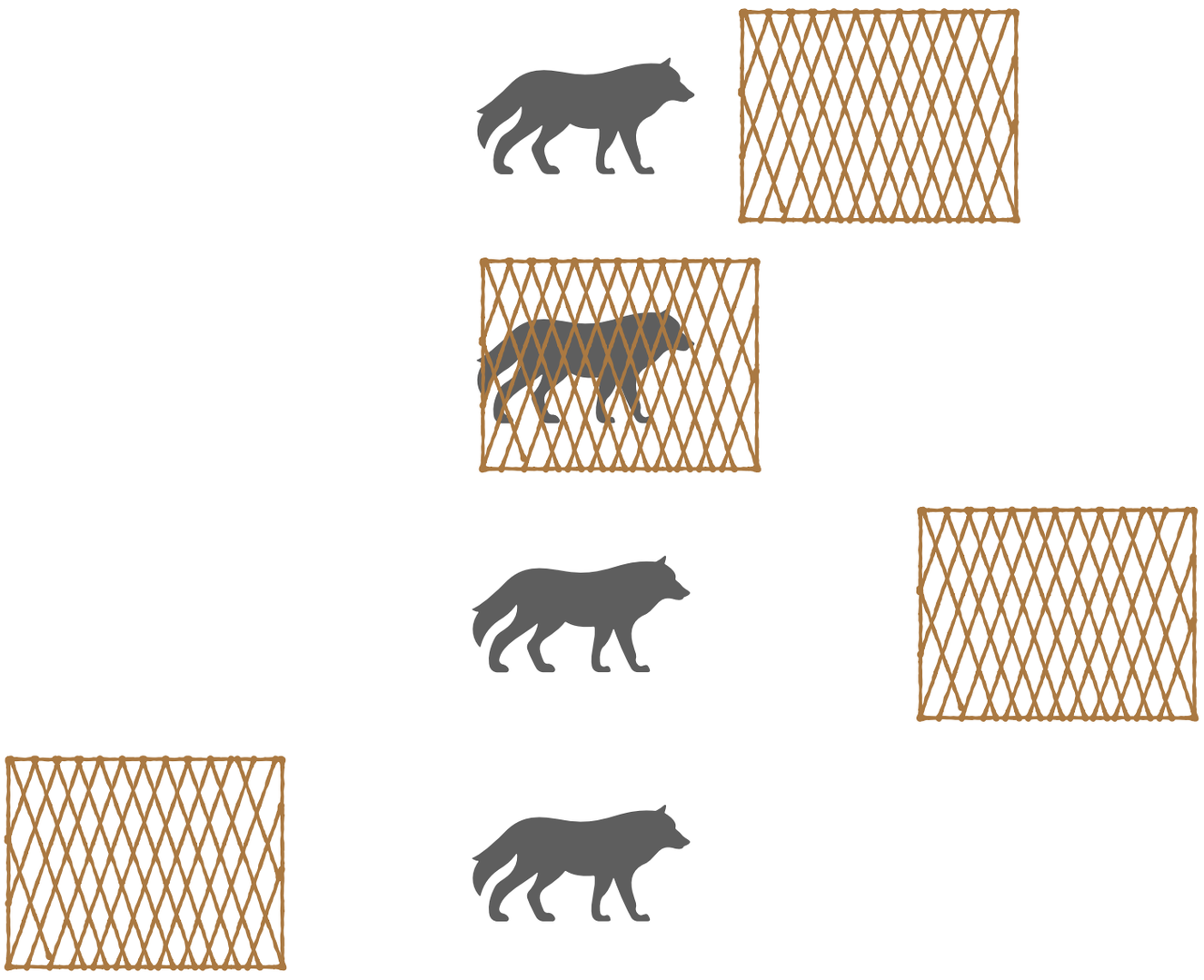
this is the question

shape

approx. normal

How do we estimate the spread of the sampling distribution?

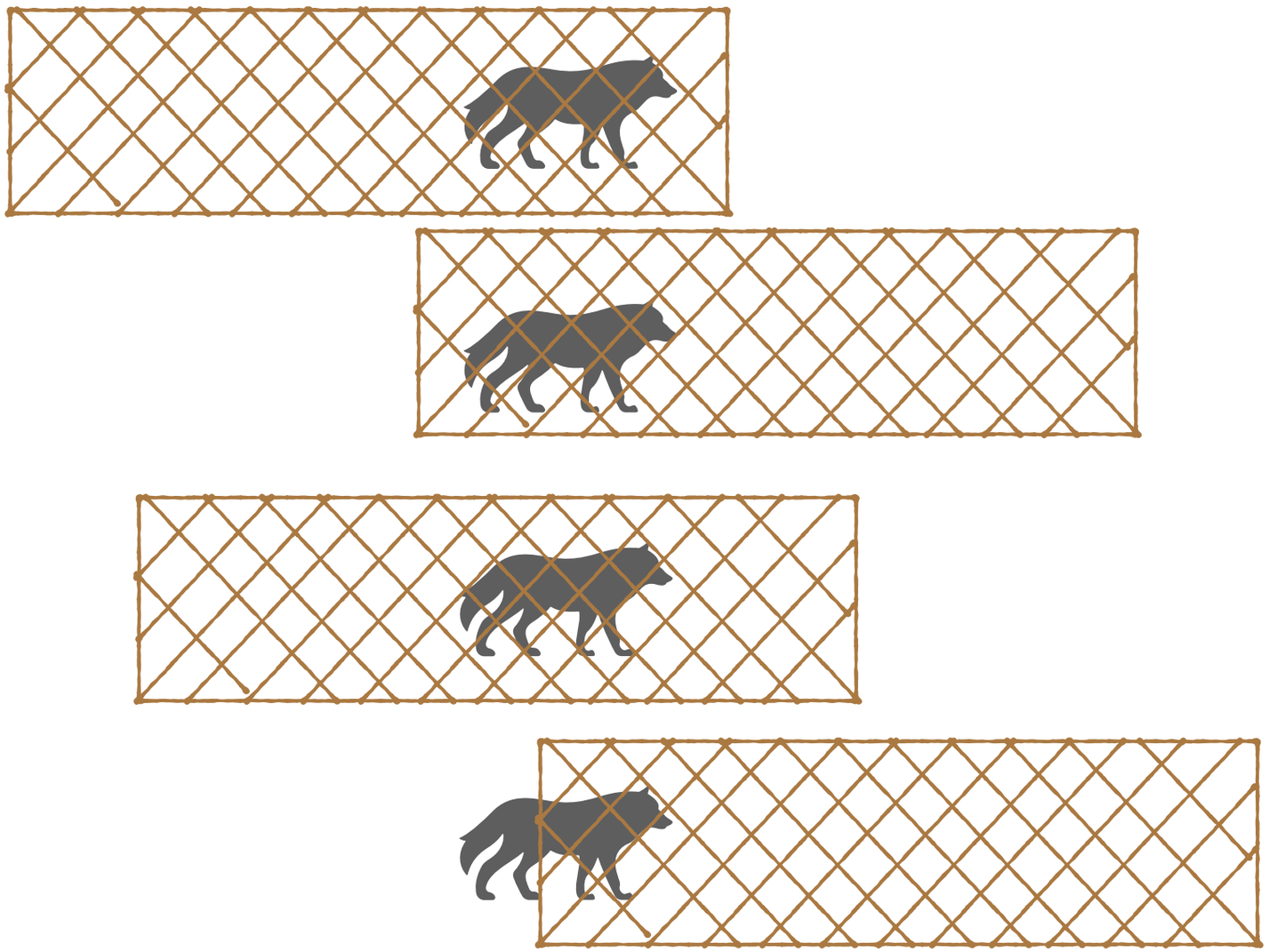
Too Narrow



Just Right

miss left about 2.5% of the time, miss right about 2.5% of the time

“almost always” catch the wolf



So how wide does our net need to be?

90% CI = estimate ± 1.64 SE

95% CI = estimate ± 1.96 SE

How do we estimate the spread of the sampling distribution?

parametric bootstrap

(building our intuition)

Inputs: observed data, a parametric model $f(\theta)$, an estimator $\hat{\theta}$, and perhaps a quantity of interest $\hat{\tau} = \tau(\hat{\theta})$.

Compute $\hat{\theta}$ from the observed data.

Do the following 2,000 times:

1. Simulate a new **b.s. data set** *from the fitted parametric model*.
2. Use the b.s. data set to obtain a **b.s. estimate** your quantity of interest. Store the estimate.

Summarize the empirical distribution of the b.s. estimates. (e.g., SD for SE, quantiles for CI)

Example

Toothpaste Cap Problem

https://gist.github.com/carlislerainey/edb70a6f0da2d27f1116cc2b7606330d

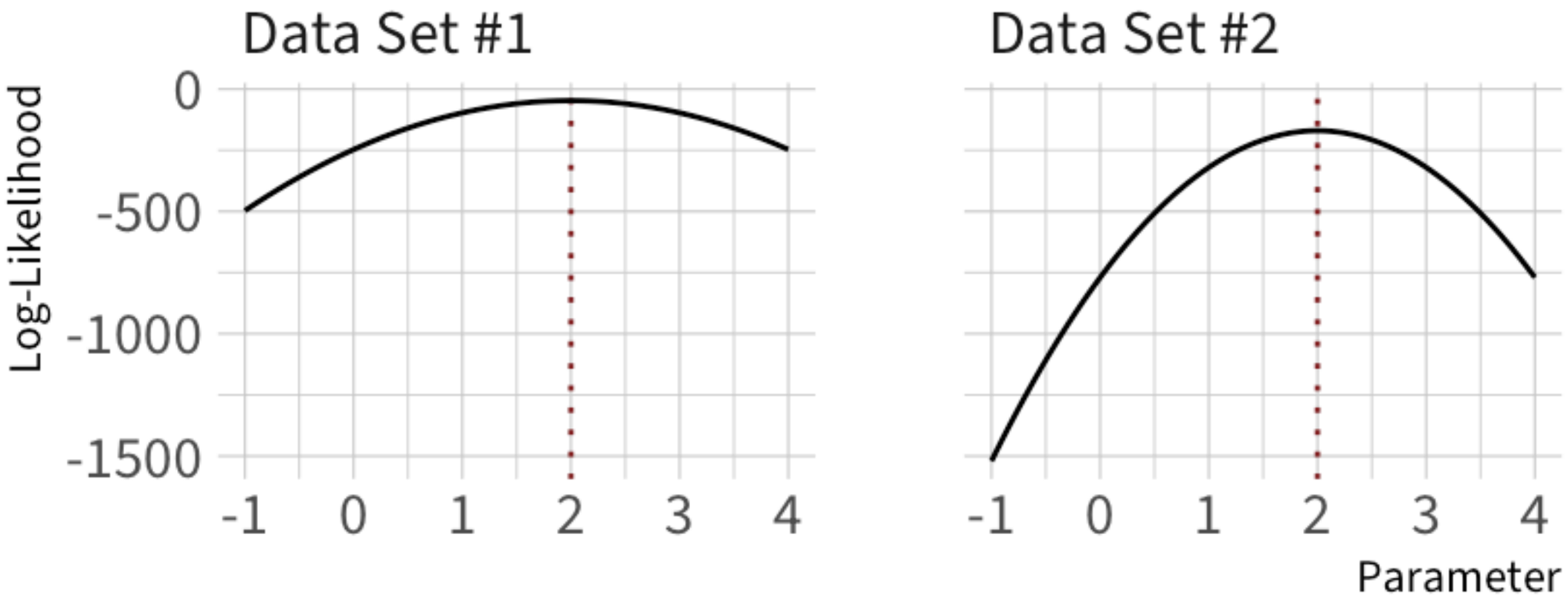
But is there a faster way?

(like a closed-form expression, maybe?)

standard errors

(a shot-in-the-dark guess)

Here are two log-likelihoods. Which produces a more precise estimate?



These are log-likelihoods from a **stylized normal model**.

(my term)

The stylized normal has unknown mean μ and known SD $\sigma = 1$.

usual normal distribution

$$f(y_i; \mu, \sigma^2) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} \exp\!\big(-\tfrac{1}{2\sigma^2}(y_i - \mu)^2\big)$$

stylized normal distribution

$$f(y_i; \mu) \;=\; \frac{1}{\sqrt{2\pi}} \exp\!\big(-\tfrac{1}{2}(y_i - \mu)^2\big)$$

(keep going on board; see handout)

standard errors

(for single parameter models)

just a **magical** *result*

the inverse of the negative curvature
is a great estimate of the variance

(this is for the ***log***-likelihood)

$$\ell''(\hat{\theta})^{-1}$$

asymptotic results

Definition (Fisher Information). For a model $f(y \mid \theta)$ with log-likelihood $\ell(\theta) = \sum_{i=1}^{N} \log f(y_i \mid \theta)$,

- the *expected information* is defined as $\mathcal{I}(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right]$ and
- the *observed information* is defined as $\mathcal{I}_{\text{obs}}(\hat{\theta}) = -\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\Big|_{\theta=\hat{\theta}}$.

Theorem (Asymptotic Normality of ML Estimators) Let $y_1, \ldots, y_N$ be iid from $f(y \mid \theta)$ with true parameter θ. Assume the regularity conditions for consistency hold and that the Fisher information $\mathcal{I}(\theta)$ is finite and positive. Then the MLE $\hat{\theta}$ is asymptotically normal so that

$$\hat{\theta} \overset{a}{\sim} \mathcal{N}\big(\theta,\ \mathcal{I}(\theta)^{-1}\big).$$

location

spread

$$\hat{\theta} \overset{a}{\sim} N(\theta, I(\theta)^{-1})$$

shape

Theorem (Asymptotic Variance of ML Estimators). Under the conditions of the asymptotic normality theorem, the variance of the MLE satisfies

$$\mathrm{Var}(\hat{\theta}) \;\approx\; \mathcal{I}(\theta)^{-1}.$$

In practice, replace the unknown θ with $\hat{\theta}$ and use the observed information:

$$\widehat{\mathrm{Var}}(\hat{\theta}) \;\approx\; \left[-\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\bigg|_{\theta=\hat{\theta}} \right]^{-1} = \left[-\ell''(\hat{\theta}) \right]^{-1}.$$

Example

Toothpaste Cap Problem

(on board; see handout)

delta method

In the one-parameter case, $\widehat{\mathrm{Var}}[\tau(\hat{\theta})] \approx \left(\tau'(\hat{\theta})\right)^2 \cdot \widehat{\mathrm{Var}}(\hat{\theta})$.

Example

Toothpaste Cap Problem

(on board; see handout)

standard errors

(for multi-parameter models)

replace the scalars with their matrix equivalents

$\frac{d^2\ell(\theta)}{d\theta^2}$ becomes $\nabla^2\ell(\theta)$

(second derivative becomes the Hessian)

$$\widehat{\mathrm{Var}}\left[\tau(\hat{\theta})\right] \approx \left(\tau'(\hat{\theta})\right)^2 \cdot \widehat{\mathrm{Var}}(\hat{\theta})$$

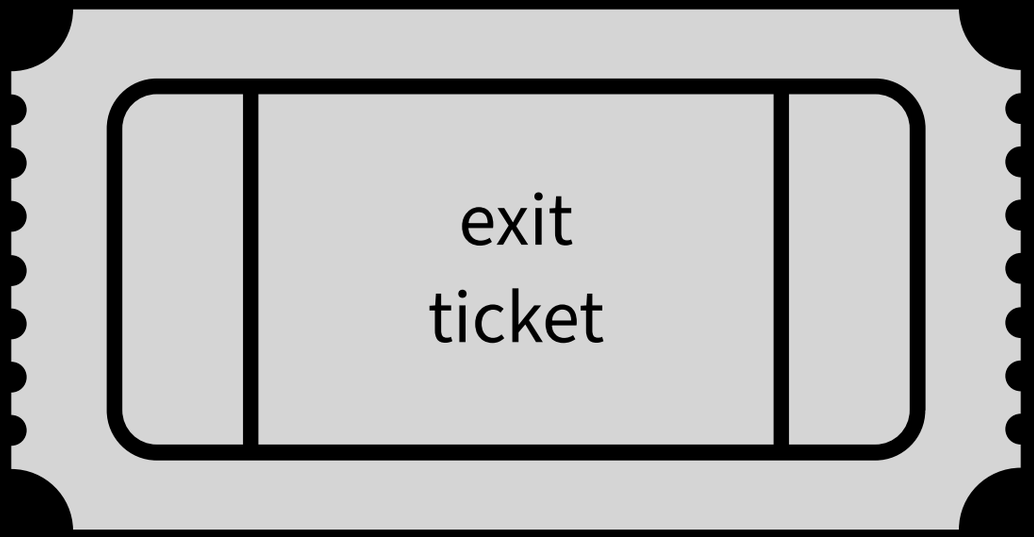
becomes $\widehat{\mathrm{Var}}(\tau(\hat{\theta})) \approx \nabla\tau(\hat{\theta})^\top \cdot \widehat{\mathrm{Var}}(\hat{\theta}) \cdot \nabla\tau(\hat{\theta})$

(derivative becomes gradient; square becomes matrix sandwich analog)

Example

Normal Model of NOMINATE w/ `optim()`

https://gist.github.com/carlislerainey/60d47883a470687f10d5f632e243ddaf



What’s one new idea from today that changed the way you understand something from before?