Parametric Families

We have seen several Families of probability distributions whose behavior is determined by a finite set of adjustable parameters θ:

Family	Parameters
$X \sim \text{Bern}(p)$	$\theta = \{p\}$
$X \sim \text{Geometric}(p)$	$\theta = \{p\}$
$X \sim \text{Poisson}(\lambda)$	$\theta = \{\lambda\}$
$X \sim N(\mu, \sigma^2)$	$\theta = \{\mu, \sigma^2\}$
$X \sim \text{Gamma}(\alpha, \beta)$	$\theta = \{\alpha, \beta\}$
$X \sim t_\nu$	$\theta = \{\nu\}$
$\vdots$	$\vdots$

Parametric statistical inference

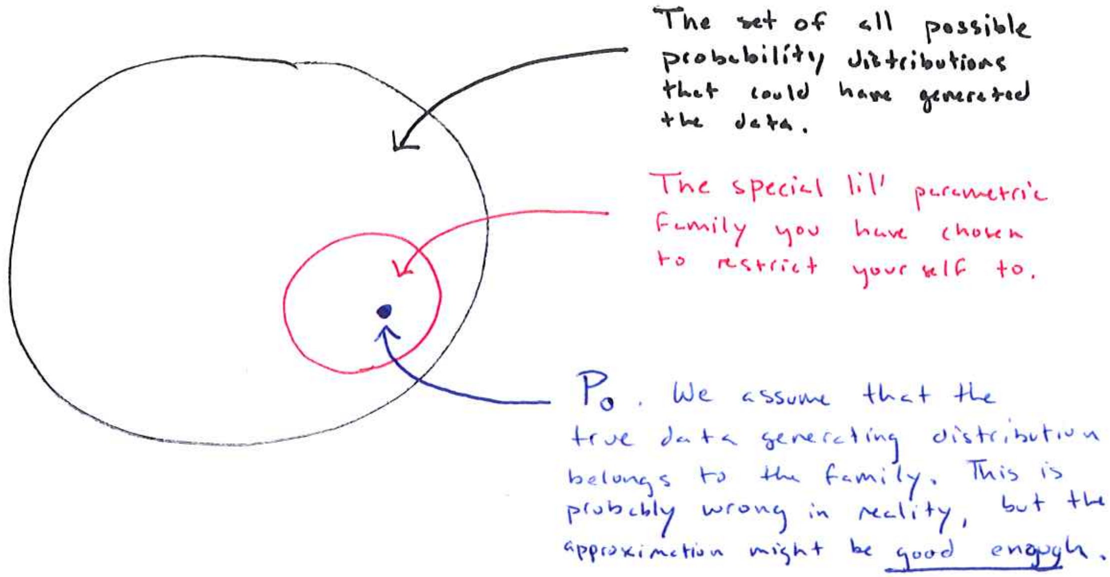
In statistics we observe data from some unknown probability distribution and use data to try to learn the distribution: $X_1, X_2, \ldots, X_n \overset{iid}{\sim} P_0$.

In parametric statistics, we make the massive simplifying assumption that the unknown distribution P_0 belongs to some familiar parametric family. In that case, in order to learn P_0, all you have to do ~~we~~ is learn the parameters. Then you're done.

Data: $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\theta$ ← pdf or pmf

Estimator: $\hat{\theta}_n = \hat{\theta}(X_1, X_2, \ldots, X_n)$

Assumption:



<u>Example</u>: $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Exponential}(\lambda)$.

$\lambda > 0$ is unknown. How do we use the data to estimate it?

What is the joint pdf of the data?

independence

$$
\begin{aligned}
f(x_1, x_2, \ldots, x_n \mid \lambda) &= f_1(x_1) f_2(x_2) f_3(x_3) \cdots f_n(x_n) \\
&= f(x_1) f(x_2) f(x_3) \cdots f(x_n) \\
&= \left(\lambda e^{-\lambda x_1}\right)\left(\lambda e^{-\lambda x_2}\right)\left(\lambda e^{-\lambda x_3}\right)\cdots\left(\lambda e^{-\lambda x_n}\right) \\
&= \lambda^n e^{-\lambda \sum_{i=1}^n x_i}.
\end{aligned}
$$

identically distributed

This is a function of n arguments: $x_1, x_2, \ldots, x_n$. λ is treated as a fixed constant. But stare at this formula and flip a switch in your brain:

$$\lambda^n e^{-\lambda \sum_{i=1}^n x_i}.$$

Instead of treating it as a function of the x_i with λ fixed, think of it as a function of λ with the x_i fixed. This is called the <u>likelihood function</u>:

$$L(\lambda \mid x_1, \ldots, x_n) = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}.$$

It describes how the joint probability mass/density of a fixed dataset varies for different choices of the unknown parameter.

For computational convenience, it will help to also define the <u>log likelihood function</u>:

$$
\begin{aligned}
\ell(\lambda \mid x_1, \ldots, x_n) &= \ln L(\lambda \mid x_1, \ldots, x_n) \\
&= \ln\left(\lambda^n e^{-\lambda \sum_{i=1}^n x_i}\right) \\
&= \ln(\lambda^n) + \ln\left(e^{-\lambda \sum_{i=1}^n x_i}\right) \\
&= n \ln \lambda + (-\lambda) \sum_{i=1}^n x_i.
\end{aligned}
$$

Maximum likelihood estimation:

To estimate the unknown parameter $\lambda > 0$, we pick the value that makes the likelihood of our observed data as large as possible. This is the maximum likelihood estimator (MLE):

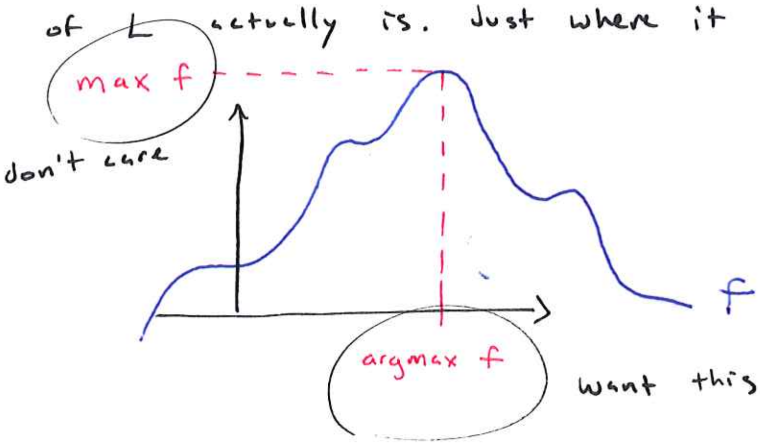
$$\hat{\lambda}_n^{(MLE)} = \underset{\lambda > 0}{\operatorname{argmax}}\, L(\lambda \mid X_{1:n})$$

$$= \underset{\lambda > 0}{\operatorname{argmax}}\, \ln L(\lambda \mid X_{1:n})$$

$$= \underset{\lambda > 0}{\operatorname{argmax}} \left[n \ln \lambda - \lambda \sum_{i=1}^{n} X_i \right].$$

natural log is an order-preserving function.

"argmax" means "the argument that does the maximizing." We want the location of the maximum value. We don't care what the maximum value of L actually is. Just where it happens.



To find the MLE, we're back in calc I:

$$\frac{d \ln L}{d\lambda} = \frac{d}{d\lambda}\left(n \ln \lambda - \lambda \sum_{i=1}^{n} X_i\right) = \frac{d}{d\lambda}(n \ln \lambda) - \frac{d}{d\lambda}\left(\lambda \sum_{i=1}^{n} X_i\right)$$

$$= \frac{n}{\lambda} - \sum_{i=1}^{n} X_i.$$

$$\frac{d \ln L}{d\lambda} = 0 \implies \frac{n}{\lambda} - \sum_{i=1}^{n} X_i = 0 \implies \frac{n}{\lambda} = \sum_{i=1}^{n} X_i$$

$$\frac{n}{\sum_{i=1}^{n} X_i} = \lambda$$

(yada yada check the second derivative)

$$\hat{\lambda}_n^{(MLE)} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}_n}.$$

Given data, this is our best guess for the unknown parameter $\lambda > 0$. Like all estimators, $\hat{\lambda}_n^{(MLE)}$ is a function of the data, and the data are random. As such, the MLE is a random variable with a sampling distribution that describes how it varies across alternative random samples. We itemize the statistical properties of the estimator by itemizing the properties of its sampling distribution: where is it centered, how spread out is it, what happens to it as $n \to \infty$, and so on. For example...

$\boxed{\text{What is the bias of } \hat{\lambda}_n^{(MLE)}}$

$$\text{bias}\left(\hat{\lambda}_n^{(MLE)}\right) = E\left(\hat{\lambda}_n^{(MLE)}\right) - \lambda$$

true but unknown value

So, what is the mean of the sampling distribution? To compute, note that $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Gamma}(1, \lambda)$, so $\bar{X}_n \sim \text{Gamma}(n, n\lambda)$. To compute the mean of $\hat{\lambda}_n^{(MLE)} = 1/\bar{X}_n$, we apply <u>LOTUS</u>:

$$E\left(\hat{\lambda}_n^{(MLE)}\right) = E\left(1/\bar{X}_n\right) = \int_0^\infty \frac{1}{x} f_{\bar{X}_n}(x)\,dx$$

$$= \int_0^\infty \frac{1}{x} \frac{(n\lambda)^n}{\Gamma(n)} x^{n-1} e^{-n\lambda x}\,dx$$

$$= \frac{n^n \lambda^n}{\Gamma(n)} \int_0^\infty x^{n-1-1} e^{-n\lambda x}\,dx$$

$$= \frac{n^n \lambda^n}{(n-1)!} \frac{\Gamma(n-1)}{(n\lambda)^{n-1}}$$

$$= \frac{\cancel{n^n}\,\cancel{\lambda^n}}{(n-1)\cancel{!}} \frac{\cancel{(n-2)!}}{\cancel{n^{n-1}}\,\cancel{\lambda^{n-1}}} = \frac{n\lambda}{n-1}.$$

So $$\text{bias}\left(\hat{\lambda}_n^{(MLE)}\right) = \frac{n\lambda}{n-1} - \lambda$$

$$= \frac{n\lambda}{n-1} - \frac{(n-1)\lambda}{n-1}$$

$$= \frac{n\lambda - n\lambda + \lambda}{n-1}$$

$$= \frac{\lambda}{n-1}.$$

Note two things.

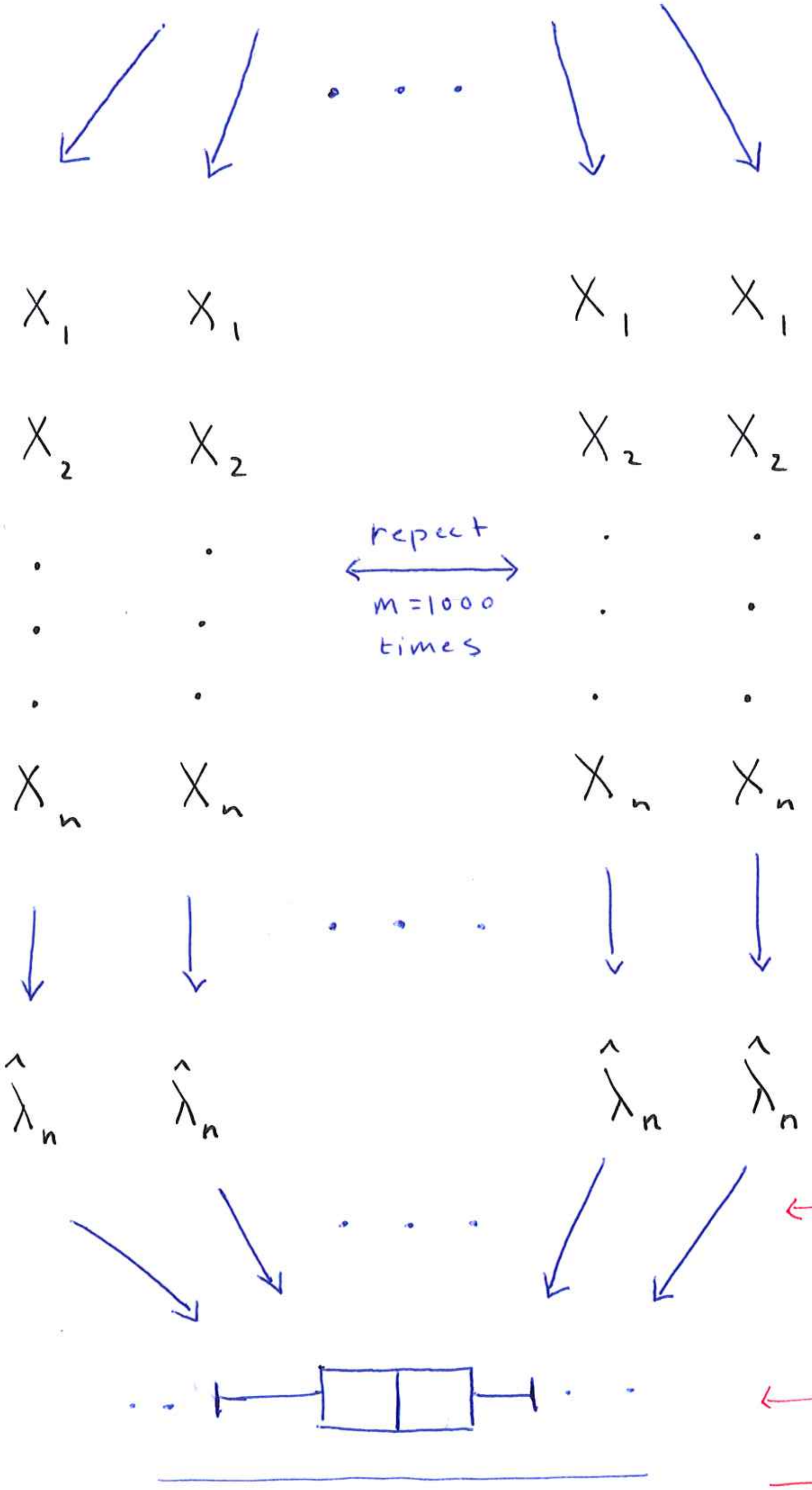
- $\text{bias}\left(\hat{\lambda}_n^{(MLE)}\right) = \frac{\lambda}{n-1} > 0$, so we are always <u>overestimating</u>.
- $\text{bias}\left(\hat{\lambda}_n^{(MLE)}\right) = \frac{\lambda}{n-1} \longrightarrow 0$ as $n \to \infty$.

So for finite n, the MLE is <u>biased</u>, but in the limit, it is <u>consistent</u>:

$$\hat{\lambda}_n^{(MLE)} \xrightarrow{\text{prob}} \lambda.$$

Let's investigate properties of the MLE via simulation:

$$P_0 = \text{Exponential}(\overbrace{\lambda = 2}^{\text{ground truth parameter value}})$$



simulate many different datasets of size n from the true distribution

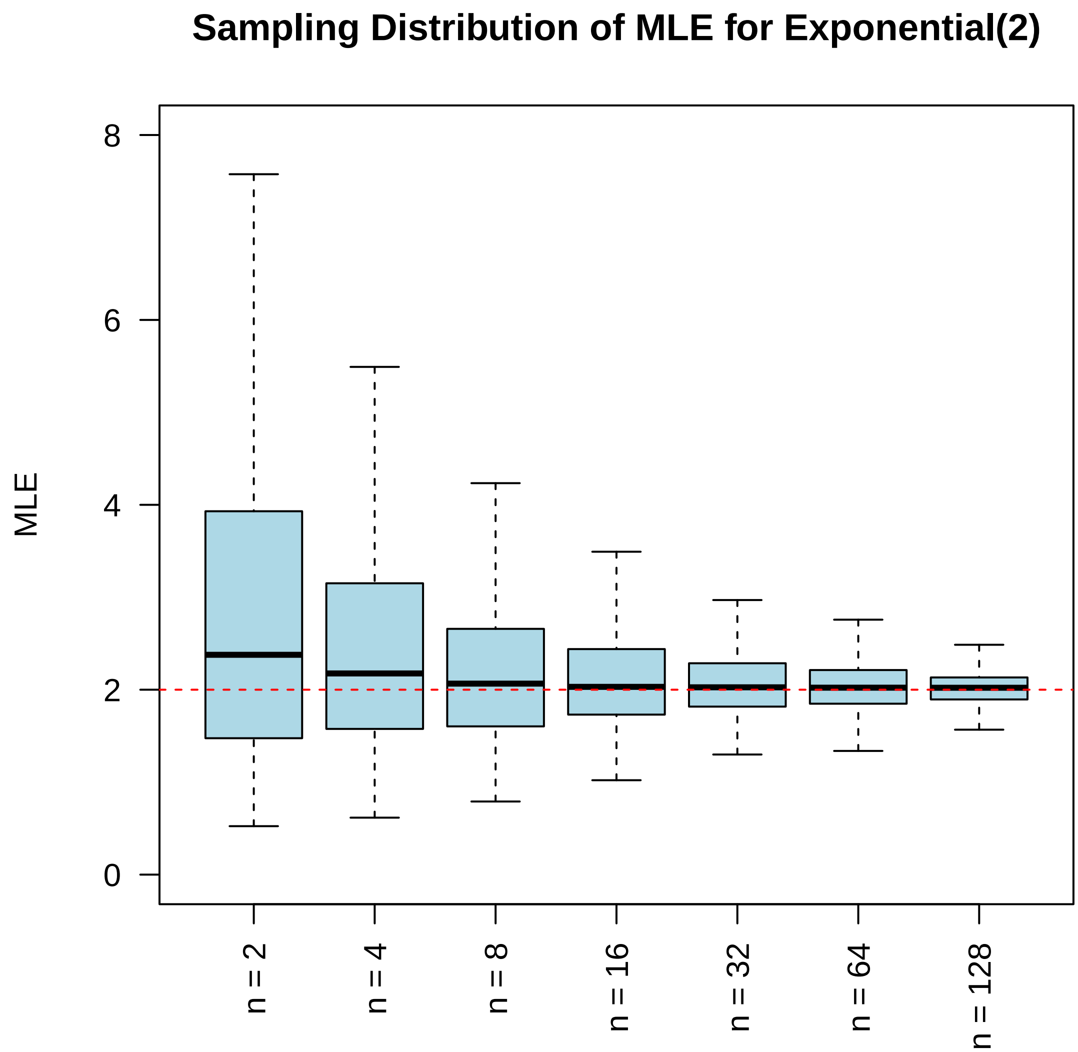
each dataset gives a different point estimate

how do these estimates vary across datasets

visualize the sampling distribution

REPEAT the entire process for different sample sizes n.

```r
set.seed(123)  # for reproducibility

# Parameters
lambda_true <- 2          # true rate parameter
n_sim <- 1000             # number of simulations per sample size
sample_sizes <- 2 * 2^(0:6)

# Container for results
mle_estimates <- vector("list", length(sample_sizes))
names(mle_estimates) <- paste0("n = ", sample_sizes)

# Simulate MLEs
for (i in seq_along(sample_sizes)) {
  n <- sample_sizes[i]
  estimates <- numeric(n_sim)
  for (j in 1:n_sim) {
    sample <- rexp(n, rate = lambda_true)
    estimates[j] <- 1 / mean(sample)  # MLE of lambda
  }
  mle_estimates[[i]] <- estimates
}

# Combine into data frame for plotting
mle_data <- stack(mle_estimates)

# Box plot
boxplot(values ~ ind, data = mle_data,
        main = "Sampling Distribution of MLE for Exponential(2)",
        ylab = "MLE",
        xlab = "",
        col = "lightblue",
        las = 2,
        ylim = c(0, 8),
        outline = FALSE)
abline(h = lambda_true, col = "red", lty = 2)  # true value of lambda
```



Each box plot is displaying the sampling distribution of the MLE for a different sample size (based on $M = 1000$ simulations). For smaller sample sizes, we see that there is some positive bias, but as the sample size increases, the sampling distribution becomes more and more concentrated around the true value of $\lambda = 2$. So via simulation, we demonstrate that this estimator is biased but consistent.

Maximum likelihood in general

Data: $X_1, X_2, \ldots X_n \overset{iid}{\sim} f(x \mid \theta)$

some pmf or pdf

true value of unknown parameter(s)

Likelihood function:

$$L(\theta \mid X_{1:n}) = f(X_1, X_2, \ldots, X_n \mid \theta)$$

$$= \prod_{i=1}^{n} f(X_i \mid \theta)$$

it's a product because of independence, it's the same f in each factor because of identical distribution.

Log-likelihood:

$$\ell(\theta \mid X_{1:n}) = \ln L(\theta \mid X_{1:n})$$

$$= \sum_{i=1}^{n} \ln f(X_i \mid \theta)$$

MLE:

$$\hat{\theta}_n^{(MLE)} = \underset{\theta}{\operatorname{argmax}}\ L(\theta \mid X_{1:n})$$

$$= \underset{\theta}{\operatorname{argmax}}\ \ln L(\theta \mid X_{1:n})$$

some properties:

consistent: $\hat{\theta}_n^{(MLE)} \xrightarrow{prob} \theta$

asymptotic normality: $\dfrac{\hat{\theta}_n^{(MLE)} - \theta}{se(\hat{\theta}_n)} \xrightarrow{dist} N(0,1)$

may or may not be biased: $bias\left(\hat{\theta}_n^{(MLE)}\right) = ?$

may or may not be zero. no guarantees in general.