STAT 135
5. Maximum Likelihood Estimation

Spring 2022

Lecturer: Dr Rebecca Barter (*she/her*)
Office hours: Tu 9:30-10:30 (in person), Th 1:30-2:30 (virtual)
Office: Evans 339

Email: rebeccabarter@berkeley.edu
Twitter: @rlbarter
GitHub: rlbarter

Computing parameter estimates

We previously used the sample mean, $\frac{1}{n}\sum_{i=1}^{n} x_i$, as an estimate of the population mean parameter.

We showed that this estimate was unbiased and had nice variance properties (e.g., the variance decreased as the sample size increased).

Computing parameter estimates

What if instead of caring about the population mean or proportion, our data came from a $Exponential(\lambda)$ distribution.

Can you think of any natural estimators of λ?

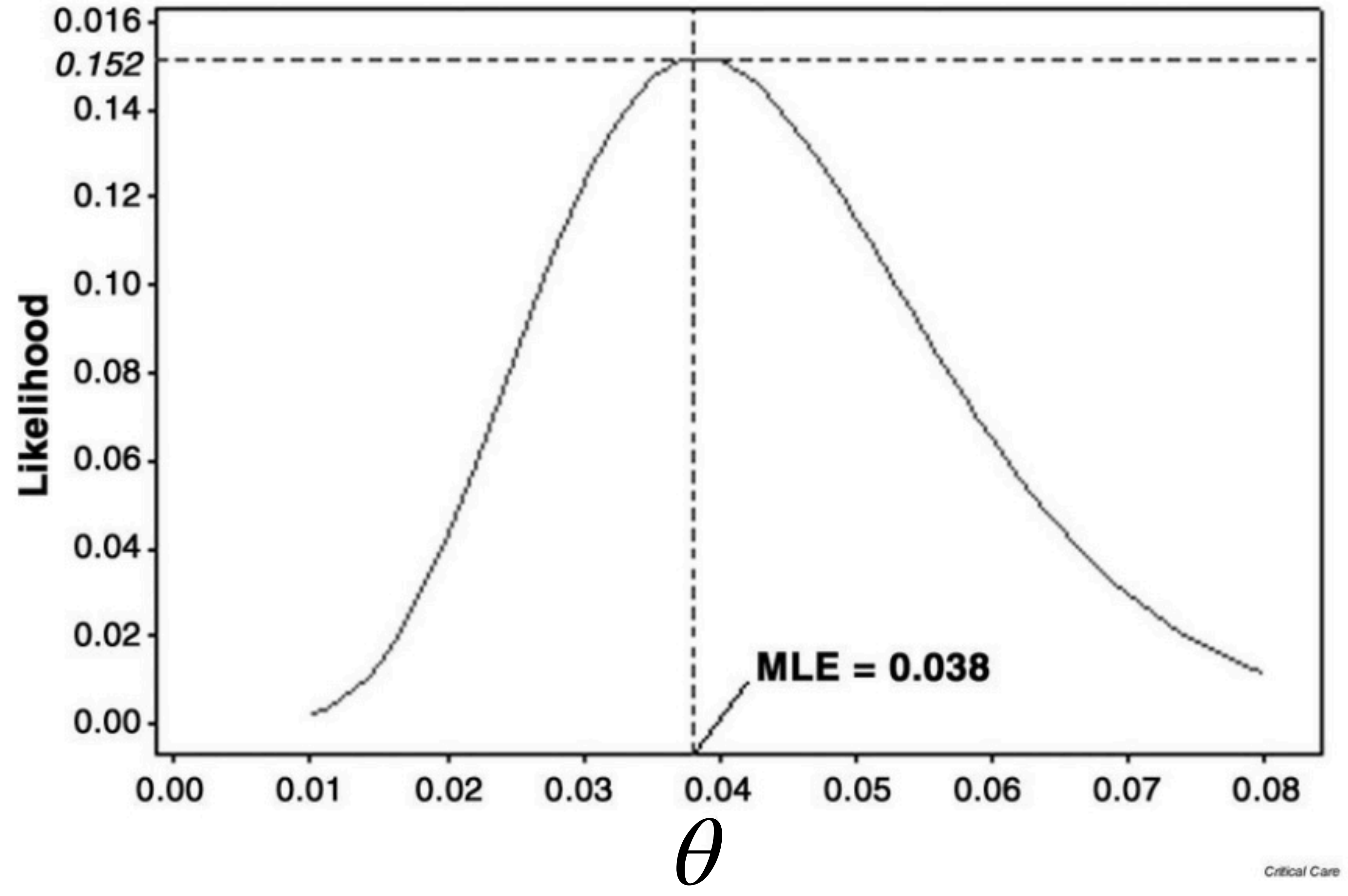
Maximum likelihood estimation (MLE) is a generating technique for identifying reasonable estimates of the parameters from any distribution.

The idea is choose the value parameter based that is “most likely” to have led to our observed data.

The likelihood function

The likelihood function

The **likelihood function**, $lik(\theta)$, corresponds to the probability of observing the particular data in our sample for various values of θ (assuming the data came from some distribution with parameter θ).



Critical Care

Our goal is to find the value of θ that maximizes the likelihood function

The likelihood function

Assume that our sample corresponds to IID random variables $X_1, \ldots, X_n$ from some distribution that has density/frequency function f_θ (a general density function with parameter θ). Denote the actual values that we observe in our data by $x_1, \ldots, x_n$.

The likelihood function can be written:

$$\begin{aligned} lik(\theta) &= P(X_1 = x_1, \ldots, X_n = x_n) \\ &= P(X_1 = x_1) \ldots P(X_n = x_n) && \text{(since the } X_i \text{ are IID)} \\ &= \prod_{i=1}^{n} f_\theta(X_i) && \text{(since the } X_i \text{ each have density } f_\theta) \end{aligned}$$

(This isn't rigorous notation for continuous RVs since they take specific values with probability 0)

Example: Bernoulli

If $X_1, \ldots, X_n \sim Bernoulli(p)$, then $f_p(x) = p^x(1-p)^{1-x}$, and

$$
\begin{aligned}
lik(p) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\
&= p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}
\end{aligned}
$$

What value of p maximizes this expression?

Maximizing the likelihood function

The **Maximum likelihood estimate**, $\hat{\theta}_{MLE}$, of a parameter θ Is the value that maximizes the likelihood function based on the observed data

The likelihood functions themselves are usually hard to maximize. The **logarithm** of the likelihood is usually much easier to maximize

(The value of θ that maximizes the likelihood function, also maximizes the *log*-likelihood function!)

MLE via maximizing the log-likelihood

The likelihood function

Our goal is to compute the format of the parameter θ that maximizes the likelihood function:

1. Write down the likelihood function: $lik(\theta) = \prod_{i} f_\theta(X_i)$
2. Calculate the *logarithm* of the likelihood function, i.e., the log-likelihood function: $\ell(\theta) = \log\left(\prod_{i} f_\theta(X_i)\right) = \sum_{i} \log(f_\theta(X_i))$
3. Differentiate the log-likelihood function with respect to θ, set to zero, and solve for θ

Example: Bernoulli

If $X_1, \ldots, X_n \sim Bernoulli(p)$, then $f_p(x) = p^x(1-p)^{1-x}$, and

1. Write down the likelihood function:

$$lik(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \quad = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}$$

2. Calculate the log-likelihood:

$$\ell(p) = \log(lik(p)) = \log p \sum_i x_i + \log(1-p)\left(n - \sum_i x_i\right)$$

$$= n\bar{x}\log p + n(1-\bar{x})\log(1-p)$$

Example: Bernoulli

If $X_1, \ldots, X_n \sim Bernoulli(p)$, then $f_p(x) = p^x(1-p)^{1-x}$, and

2. Calculate the log-likelihood

$$\ell(p) = n\bar{x}\log p + n(1-\bar{x})\log(1-p)$$

3. Differentiate with respect to p:

$$\ell'(p) = \frac{n\bar{x}}{p} - \frac{n(1-\bar{x})}{1-p}$$

Set to 0 and solve for p:

Multiply by $p(1-p)$
and divide by n

$$\frac{n\bar{x}}{p} - \frac{n(1-\bar{x})}{1-p} = 0 \implies \bar{x}(1-\hat{p}) - (1-\bar{x})\hat{p} = 0 \implies \hat{p} = \bar{x}$$

The sample mean is the MLE estimator for p!

Example: Normal (μ)

If $X_1, \ldots, X_n \sim N(\mu, \sigma)$, then $f_\mu(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, and

1. Write down the likelihood function:

$$lik(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2}$$

Example: Normal (μ)

If $X_1, \ldots, X_n \sim N(\mu, \sigma)$, then $f_\mu(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, and

1. Write down the likelihood function:

$$lik(\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2}$$

2. Compute the log-likelihood:

$$\ell(\mu, \sigma^2) = \log(lik(\mu)) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2$$

Example: Normal (μ)

If $X_1, \ldots, X_n \sim N(\mu, \sigma)$, then $f_\mu(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, and

2. Compute the log-likelihood

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2$$

3. Differentiate with respect to μ:

$$\frac{\partial \ell}{\partial \mu} = -\frac{1}{\sigma^2}\sum_i (x_i - \mu)$$

So the MLE for μ is the sample mean too!

Set to 0 and solve for μ:

$$\sum_i (x_i - \hat{\mu}) = 0 \qquad \Longrightarrow \sum_i x_i - n\hat{\mu} = 0 \qquad \Longrightarrow \hat{\mu} = \frac{\sum_i x_i}{n} = \bar{x}$$

Extra practice: Poisson

If $X_1, \ldots, X_n \sim Poisson(\lambda)$, then

$$f_\lambda(x) = \frac{\lambda^x}{x!}e^{-\lambda}, \qquad x = 0,1,2,3,...; \lambda > 0$$

Show that the MLE for λ is $\hat{\lambda}_{MLE} = \bar{X}$

Extra practice: Exponential

If $X_1, \ldots, X_n \sim Exponential(\lambda)$, then

$$f_\lambda(x) = \lambda e^{-\lambda x}, \qquad x = 0,1,2,3,...; \quad \lambda \geq 0$$

Show that the MLE for λ is $\hat{\lambda}_{MLE} = \dfrac{1}{\bar{X}_n}$

Properties of MLE estimators

Asymptotic results for the MLE

The MLE has some really nice properties as the sample size increases.

Consistency: as the sample size gets larger the MLE approaches the true parameter value

Normality: as the sample size gets larger the distribution of the MLE (as in if you were able to compute various versions of the MLE from many different random samples) becomes *Normal.*

Properties of MLE estimators

1. Consistency

Consistency

An estimate $\hat{\theta}_n$, of θ is **consistent** if

$$\hat{\theta}_n \xrightarrow{P} \theta \quad \text{as} \quad n \to \infty$$

Does this remind you of something? (LLN)

Where $\hat{\theta}_n \xrightarrow{P} \theta$ technically means that, for all $\epsilon > 0$,

$$P(|\hat{\theta}_n - \theta| > \epsilon) \to 0 \quad \text{as} \quad n \to \infty$$

But don't worry too much about this… just think of it as when *n* gets really large, the probability that $\hat{\theta}_n$ differs from θ becomes increasingly small

Consistency of the MLE

Theorem: The MLE $\hat{\theta}_{MLE,n}$, is a **consistent estimator** of the parameter, θ, that it is estimating, which means that

$$\hat{\theta}_{MLE,n} \overset{P}{\to} \theta \quad \text{as} \quad n \to \infty$$

The consistency of the MLE implies that the MLE is **asymptotically unbiased:**

$$E[\hat{\theta}_{MLE,n}] \to \theta \quad \text{as} \quad n \to \infty$$

Consistency of the MLE

Which statement is stronger?

Consistency:

The **actual value** of the parameter estimate approaches the true value of the parameter

$$\hat{\theta}_{MLE,n} \xrightarrow{P} \theta \quad \text{as} \quad n \to \infty$$

Asymptotic unbiasedness:

The **expected value** of the parameter estimate approaches the true value of the parameter

$$E[\hat{\theta}_{MLE,n}] \to \theta \quad \text{as} \quad n \to \infty$$

Continuous mapping theorem

Continuous mapping theorem:

For any continuous function g, if $\hat{\theta} \xrightarrow{P} \theta$ as $n \to \infty$, then

$$g(\hat{\theta}) \xrightarrow{P} g(\theta) \text{ as } n \to \infty$$

e.g., $\bar{X} \xrightarrow{P} \mu$ as $n \to \infty$, implies that $\bar{X}^2 \xrightarrow{P} \mu^2$ as $n \to \infty$

Sketch of proof of the consistency of the MLE

The LLN implies that

$$\frac{1}{n}\sum_{i=1}^{n}\log f_\theta(X_i) \xrightarrow{P} E_{\theta_0}(\log f_\theta(X))$$

Which implies that the value of θ maximizing the LHS (the MLE) converges to the value that maximizes the RHS, which we claim is θ_0 (the true value of θ)

So we just need to prove that the value that maximizes the RHS is θ_0

The proof of this is based on the fact that log is concave. Using Jensen's inequality you can show that

$$E_{\theta_0}\left[\log\frac{f_\theta(X)}{f_{\theta_0}(X)}\right] \leq \log E_{\theta_0}\left[\frac{f_\theta(X)}{f_{\theta_0}(X)}\right] = \log\int\frac{f_\theta(x)}{f_{\theta_0}(x)}f_{\theta_0}(x)dx = \log\int f_\theta(x)dx = 0$$

Which shows that $E_{\theta_0}[\log f_\theta(X)]$ is maximized at $\theta = \theta_0$

Properties of MLE estimators

1. Consistency
2. Asymptotic normality

Asymptotic normality of the MLE

The MLE is asymptotically Normal. If $\hat{\theta}_{ML,n}$ is the MLE estimate of a parameter θ whose true value is θ_0, then as $n \to \infty$, we have the

$$\hat{\theta}_{ML,n} \xrightarrow{D} N(\mu_{\theta_0}, \sigma^2_{\theta_0}) \quad \text{as} \quad n \to \infty$$

μ_{θ_0} and $\sigma^2_{\theta_0}$ are mean and variance parameters for the sampling distribution of $\hat{\theta}_{MLE}$ that might depend on the true value of the parameter, θ_0

The question is: what are μ_{θ_0} and $\sigma^2_{\theta_0}$????

Asymptotic normality of the MLE

The MLE is asymptotically Normal. If $\hat{\theta}_{ML,n}$ is the ML estimate of a parameter θ whose true value is θ_0, then as $n \to \infty$, we have that

$$\hat{\theta}_{ML,n} \xrightarrow{D} N\left(\theta_0, \frac{1}{nI(\theta_0)}\right) \quad \text{as } n \to \infty$$

This is from the consistency of the MLE

This is called the “Fisher Information”

Fisher information

$$\boxed{\hat{\theta}_{ML,n} \xrightarrow{D} N\left(\theta_0, \frac{1}{nI(\theta_0)}\right) \quad \text{as } n \to \infty}$$

The **Fisher information** is defined by

$$I(\theta_0) = E\left[\left(\frac{d}{d\theta}\log(f_\theta(x))\Big|_{\theta_0}\right)^2\right]$$

Or alternatively,

$$I(\theta_0) = -E\left[\frac{d^2}{d^2\theta}\log(f_\theta(x))\Big|_{\theta_0}\right]$$

Fisher information

$$\boxed{\hat{\theta}_{ML,n} \xrightarrow{D} N\left(\theta_0, \frac{1}{nI(\theta_0)}\right) \quad \text{as } n \to \infty}$$

$$I(\theta_0) = -E\left[\frac{d^2}{d^2\theta}\log(f_\theta(x))\Big|_{\theta_0}\right]$$

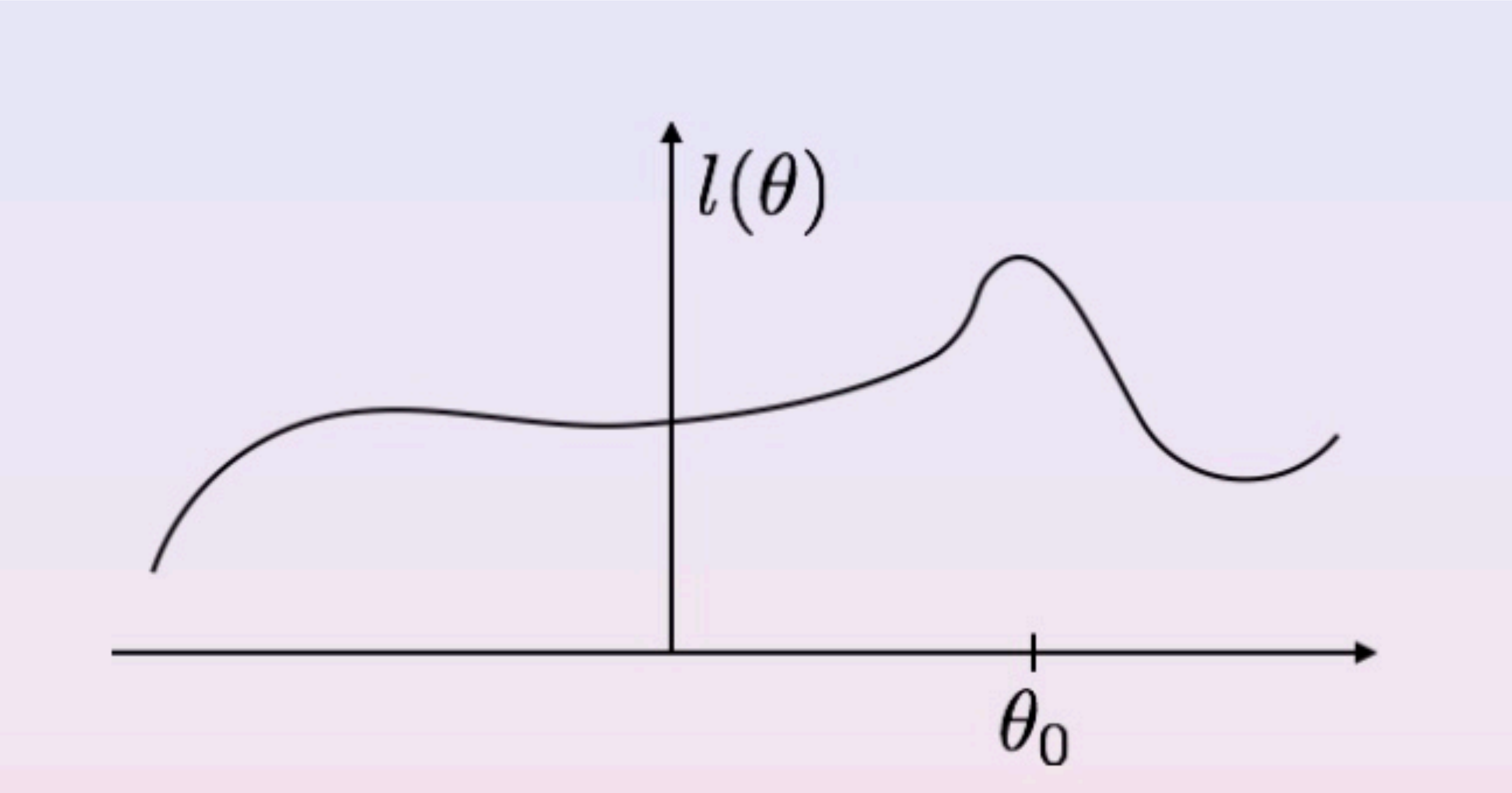
Wikipedia says: "*The Fisher information is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ upon which the probability of X depends*"

Thanks for that wikipedia.

Fisher information

$$\boxed{\hat{\theta}_{ML,n} \xrightarrow{D} N\left(\theta_0, \frac{1}{nI(\theta_0)}\right) \quad \text{as } n \to \infty}$$

$$I(\theta_0) = -E\left[\frac{d^2}{d^2\theta}\log(f_\theta(x))\bigg|_{\theta_0}\right]$$



$I(\theta_0)$ measures how “peaked” $\ell(\theta)$ is around θ_0

If $I(\theta_0)$ is large, then it is easier to detect θ_0 $\Longrightarrow$lower variance

Sketch of proof of the asymptotic normality of the MLE

$$\boxed{\hat{\theta}_{ML,n} \xrightarrow{D} N\left(\theta_0, \frac{1}{nI(\theta_0)}\right) \text{ as } n \to \infty}$$

A first order Taylor expansion to the equation $0 = \ell'(\hat{\theta})$ around $\hat{\theta} = \theta_0$:

$$0 \approx \ell'(\theta_0) + (\hat{\theta} - \theta_0)\ell''(\theta_0)$$

Which implies that

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx -\sqrt{n}\frac{\ell'(\theta_0)}{\ell''(\theta_0)} = -\frac{\frac{1}{\sqrt{n}}\ell'(\theta_0)}{\frac{1}{n}\ell''(\theta_0)} \longrightarrow \frac{1}{I(\theta_0)}N(0,\ I(-\theta_0)) = N(0,\ I(\theta_0)^{-1})$$

Where this limit comes from the LLN, the CLT, the continuous mapping theorem and Slutsky’s lemma

Asymptotic dist of MLE: Normal

If the $X_1, \ldots, X_n$ are IID from a N(μ_0,σ_0^2) distribution. What is the limiting distribution of the MLE of μ?

Asymptotic normality of MLE says:

$$\hat{\mu}_{ML,n} = \bar{X}_n \xrightarrow{D} N\left(\mu_0, \boxed{\frac{1}{nI(\mu_0)}}\right)$$

$$I(\theta_0) = -E\left[\frac{d^2}{d^2\theta}\log(f_\theta(x))\Big|_{\theta_0}\right]$$

CLT says:

$$\bar{X}_n \xrightarrow{D} N\left(\mu_0, \boxed{\sigma_0^2/n}\right)$$

What do you think $I(\mu_0)$ will be equal to?

Asymptotic dist of MLE: Normal

If the $X_1, \ldots, X_n$ are IID from a N(μ_0, σ^2) distribution.

$$I(\mu_0) = -E\left[\frac{\partial^2}{\partial^2\mu}\log(f_\mu(x))\Big|_{\mu_0}\right]$$

$$f_\mu(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$\log f_\mu(x) = -\log(\sqrt{2\pi}) - \log\sigma - \frac{1}{2\sigma^2}(x-\mu)^2$$

$$\frac{\partial^2}{\partial\mu^2}\log f_\mu(x) = -\frac{1}{\sigma^2}$$

$$I(\mu_0) = -E\left[\frac{\partial^2}{\partial^2\mu}\log(f_\mu(x))\Big|_{\mu_0}\right] = \frac{1}{\sigma^2}$$

Which means that: $\hat{\mu}_{ML,n} = \bar{X}_n \xrightarrow{D} N\left(\mu_0, \frac{1}{nI(\mu_0)}\right) = N\left(\mu_0, \frac{\sigma^2}{n}\right)$ Which agrees with the CLT!

Asymptotic dist of MLE: Binomial

If $X_1, \dots, X_n \overset{IID}{\sim} Bernoulli(p)$, then $\hat{p}_{MLE} = \bar{X}_n$

Then $P(X = k) = p^k(1-p)^{n-k}$

$$\boxed{I(\theta_0) = -E\left[\frac{d^2}{d^2\theta}\log(f_\theta(x))\bigg|_{\theta_0}\right]}$$

$$f_p(x) = p^x(1-p)^{1-x}$$

$$\log\Big(f_p(X)\Big) = X\log p + (1-X)\log(1-p)$$

$$\frac{d}{dp}\log\Big(f_\theta(X)\Big) = \frac{X}{p} - \frac{(1-X)}{1-p}$$

$$\frac{d^2}{dp^2}\log\Big(f_p(X)\Big) = -\frac{X}{p^2} - \frac{(1-X)}{(1-p)^2}$$

$$I(p_0) = -E\left[\frac{d^2}{d^2p}\log(f_p(x))\bigg|_{p_0}\right] = \frac{p_0}{p_0^2} + \frac{1-p_0}{(1-p_0)^2} = \frac{1}{p_0(1-p_0)}$$

Asymptotic dist of MLE: Bernoulli

If $X_1, \ldots, X_n \overset{IID}{\sim} Bernoulli(p)$

$$I(p_0) = -E\left[\frac{d^2}{d^2p}\log(f_p(x))\bigg|_{p_0}\right] = \frac{1}{p_0(1-p_0)}$$

Asymptotic normality of MLE says:

$$\hat{p}_{ML} = \bar{X}_n \xrightarrow{D} N\left(p_0, \frac{1}{nI(p_0)}\right)$$, which

$$\hat{p}_{ML} \xrightarrow{D} N\left(p_0, \frac{p_0(1-p_0)}{n}\right)$$

Extra practice: Poisson

If $X_1, \ldots, X_n \sim Poisson(\lambda)$, then

$$f_\lambda(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \qquad x = 0,1,2,3,...; \lambda > 0$$

The MLE for λ is $\hat{\lambda}_{MLE} = \bar{X}$

Show that the distribution of $\hat{\lambda}_{MLE} \to N(\lambda_0,\ \lambda_0)$

Extra practice: Exponential

If $X_1, \ldots, X_n \sim Exponential(\lambda)$, then

$$f_\lambda(x) = \lambda e^{-\lambda x}, \qquad x = 0,1,2,3,...; \quad \lambda \geq 0$$

The MLE for λ is $\hat{\lambda}_{MLE} = \dfrac{1}{\bar{X}_n}$

Show that the distribution of $\hat{\lambda}_{MLE} \rightarrow N(\lambda_0,\, \lambda_0^2)$

The delta method (for the mean)

The delta method:

By the CLT, we know that $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2)$ as $n \to \infty$.

For any function g such that $g'(\mu)$ exists and is non-zero, then:

$$\sqrt{n}(g(\bar{X}_n) - g(\mu)) \xrightarrow{D} N(0, \sigma^2[g'(\mu)]^2)$$

For a proof of the general case,
see http://en.wikipedia.org/wiki/Delta_method

The delta method (for the mean)

What is the limiting distribution of $\frac{1}{\bar{X}_n}$?

The δ-method tells us that:

$$\sqrt{n}(g(\bar{X}_n) - g(\mu)) \xrightarrow{D} N(0,\ \sigma^2[g'(\mu)]^2)$$

Here $g(x) = \frac{1}{x}$, and $g'(x) = -\frac{1}{x^2}$

So $\sqrt{n}\left(\frac{1}{\bar{X}_n} - \frac{1}{\mu}\right) \to N\left(0, \frac{\sigma^2}{\mu^4}\right)$

Method of Moments

Method of Moments

The MLE is just one way to formulate a parameter estimate

Another way is to relate the **sample moments** to the **theoretical moments**

The k-th moment of X is

$$\mu_k = E(X^k)$$

Theoretical moment	Sample moment
$E(X)$	$\bar{x} = \frac{1}{n}\sum_i x_i$
$E(X^2)$	$\frac{1}{n}\sum_i x_i^2$
$Var(X) = E(X^2) - E(X)^2$	$s^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$

Not technically a moment, but a function of moments →

Method of Moments: Normal example

If $X_1, \ldots, X_n \sim Normal(\mu, \sigma^2)$

Then $\mu = E(X)$ $\sigma^2 = Var(X)$

So the MOM estimators are:

$$\hat{\mu}_{MOM} = \frac{1}{n} \sum_i x_i, \quad \hat{\sigma}^2_{MOM} = \frac{1}{n} \sum_i (x_i - \bar{x})^2$$

Which is the same as the MLE

Method of Moments: Poisson

If $X_1, \ldots, X_n \sim Poisson(\lambda)$

Then $E(X) = \lambda$

$$\implies \hat{\lambda}_{MOM} = \frac{1}{n} \sum_i x_i$$

Which is the same as the MLE

Method of Moments: Gamma example

If $X_1, \ldots, X_n \sim Gamma(\alpha, \beta)$

Then $$E(X) = \frac{\alpha}{\beta} \qquad Var(X) = \frac{\alpha}{\beta^2}$$

So the MOM estimators will satisfy:

$$\begin{tikzcd}
\bar{x} = \frac{\hat{\alpha}}{\hat{\beta}}, \arrow[d] & \hat{s}^2 = \frac{\hat{\alpha}}{\hat{\beta}^2} \arrow[d] \\
\hat{\alpha} = \bar{x}\hat{\beta} \arrow[r] & \hat{s}^2 = \frac{\bar{x}\hat{\beta}}{\hat{\beta}^2}
\end{tikzcd}$$

$$\implies \hat{\beta}_{MOM} = \frac{\bar{x}}{s^2}$$

$$\implies \hat{\alpha}_{MOM} = \frac{\bar{x}^2}{s^2}$$

Method of Moments: Gamma example

If $X_1, \ldots, X_n \sim Gamma(\alpha, \beta)$

The MOM estimators are:

$$\hat{\beta}_{MOM} = \frac{\bar{x}}{s^2}$$

$$\hat{\alpha}_{MOM} = \frac{\bar{x}^2}{s^2}$$

The MLE estimators are:

$$\hat{\beta}_{MLE} =$$

$$\hat{\alpha}_{MLE} =$$

???

MOM vs MLE

MOM estimators are consistent (i.e., $\hat{\theta}_{MOM} \to \theta_0$)

This follows from the LLN for moments: if $X_1, X_2, \ldots, X_n$ is an IID sample, then:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} E(X_1^k), \quad \text{as } n \to \infty$$

But MOM estimators don't have limiting distribution results like the MLE

Cramer-Rao lower bound

Multiple estimators

While the MLE and MOM often yield the same estimators, they will sometimes differ.

How should we compare two possible estimators for the same parameter?

Compare their bias

Compare their **variance**

Cramer-Rao lower bound

Theorem (Cramer-Rao lower bound): If X_i are IID from a distribution with density f_θ, under smoothness conditions on f_θ, we have that:

If $\hat{\theta}$ is an unbiased estimator for θ, then

$$var(\hat{\theta}) \geq \frac{1}{nI(\theta)} \longleftarrow$$

This is the variance of the MLE

This result essentially states that the price to pay for having an unbiased estimator is a certain amount of variance

This means that the MLE has the lowest possible variance among unbiased estimators!

Efficiency

Efficiency

Efficiency: Given two unbiased estimators $\hat{\theta}$ and $\tilde{\theta}$ of a parameter θ, the efficiency of $\hat{\theta}$ relative to $\tilde{\theta}$ is

$$eff(\hat{\theta}, \tilde{\theta}) = \frac{var(\tilde{\theta})}{var(\hat{\theta})}$$

If $eff(\hat{\theta}, \tilde{\theta}) \leq 1$, then $var(\hat{\theta}) \geq var(\tilde{\theta})$, which implies that $\hat{\theta}$ is *less* efficient than $\tilde{\theta}$.

Efficient estimators

Efficient estimator: An **unbiased** estimator that achieves the Cramer-Rao lower bound is called **efficient**. The Cramer-Rao lower bound is:

$$Var(\hat{\theta}) = \frac{1}{nI(\theta)}$$

Note: unbiased estimators cannot do better in terms of variance than the Cramer-Rao lower bound.

The MLE is *asymptotically efficient.*

(But MLE is not necessarily efficient for finite samples.)

Efficiency example

Suppose we are interested in modeling the distance between mutations on a DNA strand. There are so many mutations on any given DNA strand that it is impossible to look at them all, so we take a sample of distances, $X_1, \ldots, X_n$. It is known that the distances are IID and follow an $Exponential(1/\lambda)$ distribution, but we don't know λ, i.e.:

$$f(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}} \quad , \quad x > 0$$

1. (Exercise) Show that $\hat{\lambda} = nX_{(1)} = n \min(X_1, \ldots, X_n)$ is an unbiased estimator of λ

2. An alternative estimator is $\tilde{\lambda} = \bar{X}_n$. What is the efficiency of $\hat{\lambda}$ relative to $\tilde{\lambda}$

Efficiency example

$X_1, \ldots, X_n$~$Exponential(1/\lambda)$, $\qquad f(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}} \quad , \quad x > 0$

1. (Exercise) Show that $\hat{\lambda} = nX_{(1)} = n \min(X_1, \ldots, X_n)$ is an unbiased estimator of λ

   Hint: Show that $X_{(1)} \sim Exponential(n/\lambda)$

Efficiency example

$$Var(X_{(1)}) = \frac{\lambda^2}{n^2}$$

$X_1, \ldots, X_n \sim Exponential(1/\lambda)$, $\quad f(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}}$, $\quad Var(X_i) = \lambda^2$

2. An alternative estimator is $\tilde{\lambda} = \bar{X}_n$. What is the efficiency of $\hat{\lambda} = nX_{(1)}$ relative to $\tilde{\lambda} = \bar{X}_n$

$$Var(\hat{\lambda}) = Var(nX_{(1)}) = n^2 \left(\frac{\lambda}{n}\right)^2 = \lambda^2$$

$$Var(\tilde{\lambda}) = Var(\bar{X}_n) = \frac{\lambda^2}{n}$$

$$eff(\hat{\lambda}, \tilde{\lambda}) = \frac{Var(\tilde{\lambda})}{Var(\hat{\lambda})} = \frac{1}{n} \leq 1$$

Which implies that $\tilde{\lambda}$ is more efficient

Sufficiency

Sufficiency

$X_1, \ldots, X_n$ can be high-dimensional and might be expensive to store

It'd be neat if there was a function (*T*) of the data that contains all of the information about a parameter of interest θ.

Sufficiency: A statistic T is said to be **sufficient** for θ if

$$X_1, \ldots, X_n \,|\, \Big(T(X_1, \ldots, X_n) = t\Big)$$

does not depend on θ for any t

Examples of statistics, T:

$$T = \bar{X}_n = \frac{1}{n} \sum X_i \qquad T = Var(X_n) = \frac{1}{n} \sum (X_i - \bar{X}_n)^2 \qquad T = max\{X_1, \ldots, X_n\}$$

$$T = 5$$

Sufficiency:

Suppose that $X_1, \ldots, X_n \sim F(\theta)$. Then $T(X_1, \ldots, X_n)$ is a **sufficient statistic** for θ if the statistician who knows the value of T can do just a good a job of estimating the unknown parameter θ as the statistician who knows the entire random sample

Sufficiency: Bernoulli example

Consider $X_1, \dots, X_n$ IID $Bernoulli(p)$

Let $T = \sum_i X_i$

$T \sim Binomial(n, p)$ $\quad P(T = t) = \binom{n}{t} p^t (1-p)^{n-t}$

$$P(X_1 = x_1, \dots, X_n = x_n \,|\, T = t) = \frac{P(X_1 = x_1, \dots, X_n = x_n, T = t)}{P(T = t)}$$

$$= \frac{P(X_1 = x_1, \dots, X_n = x_n)}{P(T = t)}$$

$$= \frac{p^{x_1}(1-p)^{1-x_1} \dots p^{x_n}(1-p)^{1-x_n}}{\binom{n}{t} p^t (1-p)^{n-t}} \quad = \frac{p^t (1-p)^{n-t}}{\binom{n}{t} p^t (1-p)^{n-t}}$$

$$= 1/\binom{n}{t}$$

So T is sufficient for p since its conditional dist does not depend on p

The factorization theorem

Factorization theorem: A necessary and sufficient condition for T to be sufficient for θ is

$$f_\theta(x_1, \ldots, x_n) = g_\theta(T)h(x_1, \ldots, x_n)$$

i.e., the density can be factored into a product such that one factor

- h, which does *not* depend on θ

and another factor,

- g, which *does* depend on θ, and depends on $(x_1, \ldots, x_n)$ only through T

Factorization theorem: Bernoulli example

Factorization theorem: A necessary and sufficient condition for T to be sufficient for θ is

$$f_\theta(x_1, \ldots, x_n) = g_\theta(T)h(x_1, \ldots, x_n)$$

Consider $X_1, \ldots, X_n$ IID $Bernoulli(p)$ Let $T = \sum_i X_i$

$T \sim Binomial(n, p)$ $P(T = t) = \binom{n}{t} p^t (1-p)^{n-t}$

$$P(X_1 = x_1, \ldots, X_n = x_n) = p^{\sum X_i}(1-p)^{n - \sum X_i} \quad = (p/(1-p))^{\sum_i X_i}(1-p)^n$$

So $h = 1$ and $g_p(T) = (p/(1-p))^T(1-p)^n$

Sufficiency:

Two ways to show that a statistic T is sufficient for θ

1. Calculating $P(X_1 = x_1, \ldots, X_n = x_n \,|\, T(X_1, \ldots, X_n))$ and showing it is independent of θ

2. Show that the density can be factorized as
$$f_\theta(x_1, , \ldots, x_n) = g_\theta(T)h(x_1, \ldots, x_n)$$

If you don't already have a sufficient statistic in mind, the factorization approach can be used to *find* sufficient statistics

Another sufficiency example

Consider $X_1, \ldots, X_n$ IID $Poisson(\lambda)$ and that the parameter of interest is $\theta = e^{-\lambda}$. The pmf is given by $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = -\frac{\theta \log(\theta)^x}{x!}$

$$f_\theta(x_1, \ldots, x_n) = \prod_i \left(-\frac{\theta \log(\theta)^{x_i}}{x_i!} \right) = \theta^n (-\log \theta)^{\sum_i x_i} \cdot \frac{1}{\prod_i x_i!}$$

$$\implies g_\theta(T) = \theta^n (-\log \theta)^{\sum_i x_i}, \qquad h(x) = \frac{1}{\prod_i x_i!}$$

So $\sum_i X_i$ is a sufficient statistic for θ

Sufficiency and MLE

Corollary: If T is sufficient for θ, then $\hat{\theta}_{MLE}$ is a function of T

Why?

If $f_\theta(x_1, \dots, x_n) = g_\theta(T)h(x_1, \dots, x_n)$

Then

$$\log(lik(\theta)) = \log(f_\theta(x_1, \dots, x_n)) = \log(g_\theta(T)) + \log(h(x_1, \dots, x_n))$$

So $\log(h(x_1, \dots, x_n))$ plays no role in the maximization of the log-likelihood ($\log(lik(\theta))$ since it does not involve θ

Rao-Blackwell theorem and the bias-variance tradeoff

The Rao-Blackwell theorem

Rao-Blackwell theorem: Suppose that $\hat{\theta}$ is an estimator for θ (with $E(\hat{\theta}^2) < \infty$ and that T is a sufficient statistic for θ. If we define a new estimator to be

$$\tilde{\theta} = E(\hat{\theta}\,|\,T)$$

Then $MSE(\tilde{\theta}) \leq MSE(\hat{\theta})$

i.e., if we know a sufficient statistic T, and we have an estimator $\hat{\theta}$, then we can define an even better estimator, $\tilde{\theta}$, for θ which has smaller MSE

Poisson example

Consider $X_1, \ldots, X_n$ IID $Poisson(\lambda)$ and that the parameter of interest is $\theta = e^{-\lambda}$. The pmf is given by $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = -\frac{\theta \log(\theta)^x}{x!}$

We already showed that $\sum_i X_i$ is a sufficient statistic for θ

1. Show that $\hat{\theta} = 1\{X_1 = 0\}$ is an unbiased estimator for θ
2. Use the Rao-Blackwell theorem to find an estimator for θ with smaller MSE than $\hat{\theta}$

Poisson example

Consider $X_1, \ldots, X_n$ IID $Poisson(\lambda)$ and that the parameter of interest is $\theta = e^{-\lambda}$. The pmf is given by $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = -\frac{\theta \log(\theta)^x}{x!}$

1. Show that $\hat{\theta} = 1\{X_1 = 0\}$ is an unbiased estimator for θ

$$
\begin{aligned}
E(1\{X_1 = 0\}) &= 0 \times P(X_1 \neq 0) + 1 \times P(X_1 = 0) \\
&= P(X_1 = 0) \\
&= e^{-\lambda}\lambda^0/0! \\
&= e^{-\lambda} \\
&= \theta
\end{aligned}
$$

Poisson example

$X_1, \ldots, X_n$ IID $Poisson(\lambda)$, $\theta = e^{-\lambda}$, and $P(X = x) = \dfrac{e^{-\lambda}\lambda^x}{x!} = -\dfrac{\theta \log(\theta)^x}{x!}$

2. Use the Rao-Blackwell theorem to find an estimator for θ with smaller MSE than $\hat{\theta}$

Rao-Blackwell tells us that the following estimator, $\tilde{\theta}$, must have smaller MSE than $\hat{\theta}$:

$$\tilde{\theta} := E\left(\hat{\theta} \,\middle|\, \sum_i X_i = t\right) = E\left(1\{X_1 = 0\} \,\middle|\, \sum_i X_i = t\right) = P\left(X_1 = 0 \,\middle|\, \sum_i X_i = t\right)$$

$$= \frac{P\left(X_1 = 0, \sum_{i=2}^n X_i = t\right)}{P(\sum_{i=1}^n X_i = t)} = \frac{P\left(X_1 = 0\right) P\left(\sum_{i=2}^n X_i = t\right)}{P(\sum_{i=1}^n X_i = t)} = \frac{\exp(-\lambda)\frac{((n-1)\lambda)^t e^{-(n-1)\lambda}}{t!}}{\frac{(n\lambda)^t e^{-n\lambda}}{t!}}$$

$$= \left(\frac{n-1}{n}\right)^t$$ (Since $\sum_i X_i \sim Poisson(n\lambda)$)

So $\tilde{\theta} = \left(\dfrac{n-1}{n}\right)^{\sum_i X_i}$