# Parametric Inference, Chapter 9

Elizabeth Purdom

*This document has last been compiled on Sep 16, 2024.*

## Contents

# 1 Parametric Models

A parametric model means the statistical models we are chosing from have the form:

$$\mathcal{F} = \{F(x;\theta) : \theta \in \Theta\}$$

where $\Theta \subseteq \mathbb{R}^k$ is the parameter space. In this case, we often call $\mathcal{F}$ a parametric model of distributions.

We typically choose a class $\mathcal{F}$ based on knowledge about the particular problem. We might state this as making certain assumptions about the data generating mechanism. It's good practice when using a parametric model to consider how the procedures will behave if these assumptions are violated.

For now we are going to focus on methods to estimate parameters $\theta$ that describe the distribution, i.e. we have a distribution $F_\theta$ and we want to estimate $\theta$. We will return to functions of $\theta$ afterward.

We'll begin with two methods for constructing estimators of $\theta$ for parametric models: the method of moments and maximum likelihood estimation. But before we do that, let's talk about some special cases and terminology around parametric models.

**Multivariate $X$ versus multivariate $\theta$** You need to keep clear the difference between the dimension of the data $X$ and that of the parameter $\theta$.

In particular, we often have $X = X_1, \ldots, X_n$, meaning that $X$ is a random vector. So that $X \in R^n$ has a joint (multivariate) distribution. We often assume i.i.d data, so it's a simple joint distribution. If $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, 1)$, then the joint distribution is

$$\begin{aligned} f(x_1, \ldots, x_n) &= \prod_{i=1}^{n} f(x_i; \mu, 1) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{(x_i - \mu)^2}{2}\} \\ &= \frac{1}{(2\pi)^{n/2}} \exp\{-\sum_{i=1}^{n}(x_i - \mu)^2\} \end{aligned}$$

However, there is only 1 parameter $\mu$, so the parameter space $\Theta = R$.

On the other hand, $X \sim N(\mu, \sigma)$ is a single random variable, but $\theta = (\mu, \sigma)' \in R^2$ is a multivariate parameter.

## 1.1 Nuisance Parameters and Parameters of Interest

When the parametric model has more than one parameter ($\theta$ is a vector), then the actual scientific questions we are trying to address from the data is often formed in terms of one of the parameters.

For example, if we want to estimate the average of the population, we might make the simple assumption that our data is i.i.d $N(\mu, \sigma^2)$. Then $\theta = (\mu, \sigma^2)$ and our parameter space is

$$\Theta = \{(\mu, \sigma) : \mu \in \mathbb{R}, \sigma > 0\}.$$

To address our actual question, $\mu$ is of relevance and called our **parameter of interest**. However, as we have seen, we need an estimate of $\sigma^2$ in order to perform inference on $\mu$. Then other parameters are often called a **nuisance parameter**.

# 2 Types of Parametric Models

Often our parametric model $\mathcal{F}$ is a set of distributions that share a density function, and the set of models comes from the values the parameter takes on. We usually write the models in terms of the distribution of the data. For example

$$\mathcal{F} = \{F : F = Poisson(\lambda), \lambda > 0\} \rightarrow X_i \overset{i.i.d}{\sim} Poisson(\lambda)$$
$$\mathcal{F} = \left\{F : F = Normal(\mu, \sigma^2), \sigma > 0\right\} \rightarrow X_i \overset{i.i.d}{\sim} Normal(\mu, \sigma^2)$$

We can also talk about parametric "families". This usually refers to classes of parametric models that share certain properties. Most of the time a family is too broad to define a parametric model for the estimation procedure we will talk about. But because the family of models shares certain properties in common, we will sometimes have results that are true for the entire family, without needing to specify a particular distribution.

## 2.1 Location and Scale Family

We can also construct a parametric model by starting with a single distribution and creating parameters. A common class of such a family of distributions are location and scale families.

Let $X$ be a random variable with distribution $F$. Without loss of generality, let's assume $E(X) = 0$ and $var(X) = 1$.

Let $F_\mu$ be the distribution function of $X + \mu$. The family $\mathcal{F} = \{F_\mu : -\infty < \mu < \infty\}$ is called a **location family** with $\mu$ is the location parameter. For example, if $F = \mathcal{N}(0, 1)$ then $\mathcal{F}$ consists of all normal distributions with variance 1.

Similarly, let $F_\sigma$ be the distribution of $\sigma X$. The family $\{F_\sigma : \sigma > 0\}$ is called a **scale family** and $\sigma^2$ is the scale parameter. Similarly, if $F = \mathcal{N}(0, 1)$ then $F_\sigma$ consists of all normal distributions with mean 0.

Putting these together we can also have $F_{\mu,\sigma}$ be the distribution of $\sigma X + \mu$. The family $\{F_{\sigma,\mu} : -\infty < \mu < \infty, \sigma > 0\}$ is called a **location scale family**. Clearly the set of all normal distributions is a location scale family.

**Exercise 1.** If $X \sim F$ is continuous with density $f_X$, then for the location-scale family $\{F_{\sigma,\mu}\}$, if $Y \sim \{F_{\sigma,\mu}\}$, then the density of $Y$ is

$$f_Y = \frac{1}{b} f_X(\frac{y - a}{b})$$

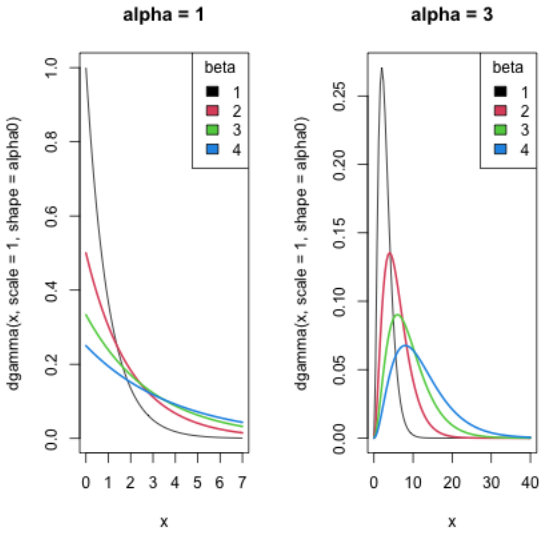
For any distribution $F$ you can construct location-scale families as a class of distributions $\mathcal{F}$. The ubiquity of the normal distribution makes these families natural to consider, but its important to recognize that many common parametric distributions are not parameterized in a way that makes them location-scale families.

**Example: Gamma distribution** For example, consider the Gamma distribution, with

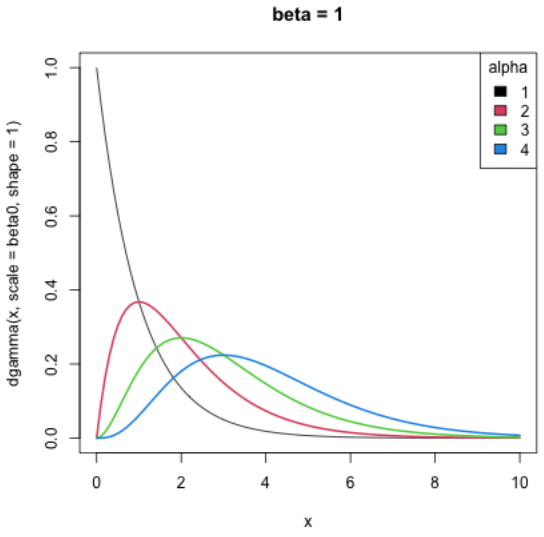
$$f(x; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad x > 0$$

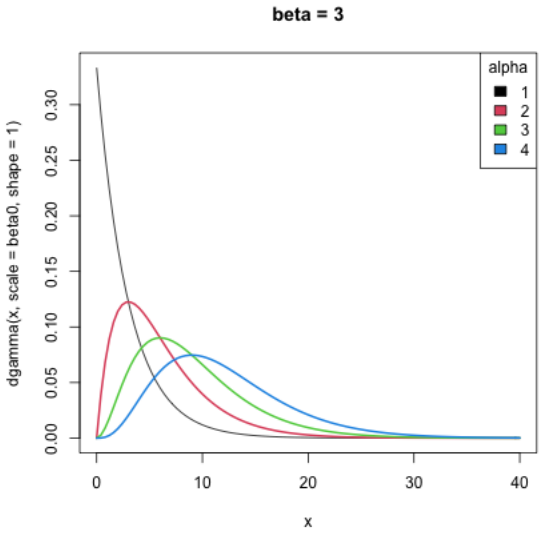
for $\alpha, \beta > 0$

If we *fix* $\alpha = \alpha_0$ (i.e. assume it is known) then the family $\Gamma(\alpha_0, \beta)$ is a scale family.




However, $\alpha$ is not a location (or scale) parameter, if $\beta$ is fixed, nor is $Gamma(\alpha, \beta)$ a location-scale family.






## 2.2 Exponential Family

A very common class of distributions is the **exponential family**. (Be careful to not confuse this with the exponential *distribution*). Many common distributions you have been taught are members of the exponential family of distributions.

**Definition 2.1** (Univariate Exponential Family)**.** A univariate distribution is a member of the exponential family if the density $f(x;\theta)$ can be written as

$$f(x;\theta) = h(x)c(\theta)\exp\left\{\eta(\theta)T(x)\right\}$$

where $h$ and $c$ are non-negative.

Let's make this concrete with an example.

**Example: Normal Distribution** The normal distribution $N(\mu, 1)$ has

$$\begin{aligned} f(x;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}}\exp\{-\frac{(x-\mu)^2}{2}\} \\ &= \frac{1}{\sqrt{2\pi}}\exp\{-\frac{x^2-2x\mu+\mu^2}{2}\} \\ &= \underbrace{\frac{1}{\sqrt{2\pi}}\exp\{-\frac{x^2}{2}\}}_{h(x)}\underbrace{\exp\{-\frac{\mu^2}{2}\}}_{c(\theta)}\exp\{x\mu\} \end{aligned}$$

with $\eta(\theta) = \theta$ and $T(x) = x$

Notice that

- Often times, the standard way of the writing the equation doesn't match the defintion, and you have to manipulate the density.
- Each of the functions in the definition rely either entirely on $x$ ($h(x), T(x)$) or $\theta$ ($c(\theta), \eta(\theta)$). So we can decompose the likelihood into products of function that separate out the $x$ part from the $\theta$ part.

**Exercise 2.** Show that each belongs to the exponential family

- $Binomial(n, p)$ with $n$ known

- $Exponential(\lambda)$

**Exercise 3.** Show that $Unif(0, \theta)$ does not belong to the exponential family.

**Dimensions** Our definition does not assume that $X \in R$. In particular $X$ can be a vector of data, like $X = (X_1, \ldots, X_n)$.

$T$ and $h$ must be real-valued, but their domain can be in higher dimensions. For example, $T(x)$ can be a function from $R^n \to R$, like

$$T(x) = T(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

In this case $T(X)$ takes as input $n$ values $X_i$ and returns $\bar{X}$, a single value.

Our definition (so far) assumes a univariate parameter $\theta \in R$ but we will expand that in just a moment.

**Exercise 4.** Show that if $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, 1)$, then the joint density

$$f_n(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \mu)$$

is an member of the exponential family, where $f(x_i; \mu)$ is the normal density for a $N(\mu, 1)$.

**Alternative description (Canonical Form)** Sometimes, to make it easier to work with, we will describe the distribution with respect to $\eta = \eta(\theta)$. $\eta$ is often called the *canonical* or *natural* parameter of an exponential family. This gives us

$$f(x;\eta) = h(x)d(\eta)\exp\{\eta T(x)\} = h(x)\exp\{\eta T(x) - A(\eta)\}$$

($d(\eta) = c(\theta(\eta))$ and we just let $d(\eta) = exp\{-A(\eta)\}$)

In this representation, we have the following useful property:

$$E(T(X)) = A'(\eta)$$

**Exercise 5.** What is the canonical parameter of a Binomial distribution? What is $A'(\eta)$?

### 2.2.1 Multivariate exponential families

We can extend exponential families to multivariate parameters by replacing the product of $\eta(\theta)$ and $T(x)$ with the inner product. Specifically, let

$$\theta = (\theta_1, \ldots, \theta_k).$$

**Definition 2.2** (Multivariate Exponential Family)**.** A multivariate exponential family with parameter $\theta \in R^k$ has a pdf in the following form:

$$f(x;\theta) = h(x)c(\theta)\exp\left\{\sum_{j=1}^{k}\eta_j(\theta)T_j(x)\right\}$$

and all the functions are real-valued functions (but their domains may be vector valued).

Be careful about the domain of the functions here. $X$ will generally be of different dimensions than $\theta$. So for example, if $X = (X_1, \ldots, X_n)$, then we would have

$$\begin{aligned} h &: R^n \to R \\ T_j &: R^n \to R \end{aligned}$$

while,

$$\begin{aligned} c &: R^k \to R \\ \eta_j &: R^k \to R \end{aligned}$$

**Canonical Form** We can write the multivariate version with respect to our canonical parameter $\eta_j = \eta_j(\theta)$, a function from $R^k \to R$,

$$f(x;\eta) = h(x)\exp\left\{\sum_{j=1}^{k}\eta_j T_j(x) - A(\eta)\right\}.$$

Then we have

$$E_\eta(T_j) = \frac{\partial}{\partial\eta_j}A(\eta)$$
$$cov_\eta(T_j, T_k) = \frac{\partial}{\partial\eta_j\partial\eta_k}A(\eta)$$

**Vector Form** We often write the density more succinctly in vector form:

$$\begin{aligned} f(x;\theta) &= h(x)c(\theta)\exp\left\{\sum_{j=1}^{k}\eta_j(\theta)T_j(x)\right\} \\ &= h(x)c(\theta)\exp\left\{\eta(\theta)'T(x)\right\} \end{aligned}$$

where

$$\eta(\theta) = (\eta_1(\theta), \ldots, \eta_k(\theta))$$
$$T(x) = (T_1(x), \ldots, T_k(x)).$$

Or in canonical form,

$$f(x;\eta) = h(x)\exp\left\{\eta'T(x) - A(\eta)\right\}.$$

**Exercise 6.** Show that the following are multivariate exponential families and identify $T_j(x)$ and $\eta_j(\theta)$.

1. $N(\mu, \sigma^2)$
2. $Beta(\alpha, \beta)$

# 3 Method of Moments

The first method of finding an estimate of a parameter is the **Method of Moments** which is historically one of the earliest methods for finding estimates and are very easy to compute, though their performance is not always as good as other options.

Suppose $\theta = (\theta_1, \ldots, \theta_k)$. For $j = 1, \ldots, k$, define the $j^{th}$ moment

$$\alpha_j \equiv \alpha_j(\theta) = E_\theta[X^j] = \int x^j dF_\theta(x)$$

and the $j^{th}$ sample moment

$$\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j.$$

Notice the $j^{th}$ sample moment is the $j$th moment of $\hat{F}_n$.

The method of moments estimator $\hat{\theta}_n$ is found by setting equal the theoretical moments (functions of $\theta$) and the sample moments (functions of $X_1, \ldots, X_n$). The $\theta$ value that solves this system of equations is designated as $\hat{\theta}$ and is our methods of moments estimator.

**Definition 3.1** (Methods of Moments (MOM) Estimator)**.** We define the MOM estimator $\hat{\theta}_n$ to be the value of $\theta$ such that:

$$\begin{array}{rcl} \alpha_1(\hat{\theta}_n) & = & \hat{\alpha}_1 \\ \alpha_2(\hat{\theta}_n) & = & \hat{\alpha}_2 \\ \vdots & \vdots & \vdots \\ \alpha_k(\hat{\theta}_n) & = & \hat{\alpha}_k \end{array}$$

Notice that there is not a guarantee that there is a solution to the system of equations (nor that it is unique).

**Example** For the case of a normal distribution, We have that $\theta = (\mu, \sigma^2)$ and

$$\begin{aligned} \alpha_1(\theta) &= E_\theta[X^1] = \mu \\ \alpha_2(\theta) &= E_\theta[X^2] = \sigma^2 + \mu^2 \end{aligned}$$

While

$$\begin{aligned} \hat{\alpha}_1 &= \frac{1}{n}\sum_i X_i = \bar{X}_n \\ \hat{\alpha}_2 &= \frac{1}{n}\sum_i X_i^2 \end{aligned}$$

This gives the system of equations

$$\hat{\mu} = \bar{X}$$
$$\hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n}\sum_i X_i^2$$

Solving this system gives

$$\hat{\mu} = \bar{X}$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_i X_i^2 - \bar{X}^2$$

**Exercise 7.** Suppose that $Y \sim Binomial(n, p)$, what is the MOM estimate of $p$? How is this related to Example 9.4 in Wasserman?

**Exercise 8.** Find a method of moments estimator for $X_1, \ldots, X_n \sim U(a, b)$.

**MOM generalization:** Instead of using $\alpha_j(\theta) = E_\theta[X^j]$, we can consider transformations $g(X)$ and use their expectations, i.e.

$$\alpha_j(\theta) = E_\theta[g(X)^j]$$

and find $\hat{\theta}_n$ s.t. $\alpha_j(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^n g(X_i)^j$, $j = 1, ..., n$.

**Exponential Families** We noted that for our exponential families (using the $\eta$ parameterization), we have

$$E(T(X)) = A'(\eta).$$

We can use this to get a MOM estimator for $\eta$, i.e. $\hat{\eta}$ is the solution to the following equation:

$$\frac{1}{n}\sum_{i=1}^n T(X_i) = A'(\eta).$$

## 3.1 Properties

If you make assumptions on the parametric model, you can get consistency and asymptotic normality of the estimator.

**Theorem 1.** *Let $\hat{\theta}_n$ be the MOM estimator of $\theta$. Under appropriate conditions*

1. *The estimate $\hat{\theta}_n$ exists with probability tending to 1*

2. *The estimate is consistent:* $\hat{\theta}$

3. *The estimate is asymptotically normal*

$$\sqrt{n}(\hat{\theta}_n - \theta) \overset{D}{\approx} N(0, \Sigma)$$

You can find the equation for $\Sigma$ in Wasserman (Theorem 9.6). Notice that these are asymptotic statements for the vector $\hat{\theta}$. We will discuss later in this module what these multivariate statements mean when we discuss the MLE.

# 4 Maximum Likelihood Estimator (MLE)

**Joint Density and Likelihood** We assume that the joint density of our data $X_1, \ldots, X_n$ is given by

$$f(x_1, \ldots, x_n; \theta)$$

If our data $X_i$ are i.i.d with density $f(x; \theta)$ (the density of each $X_i$), then the joint density of our data $X_1, \ldots, X_n$ is given by

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta) \quad \text{if the data are independent}$$

**Definition** (Likelihood Function)**.** The **likelihood** function $\mathcal{L}_n(\theta)$ is the joint density of the data *when evaluated at the observed data*

$$\begin{aligned} \mathcal{L}_n(\theta) &= f(X_1, \ldots, X_n; \theta) \\ &= \prod_{i=1}^{n} f(X_i; \theta) \quad \text{if the data are independent} \end{aligned}$$

Keep track the difference between the likelihood and the density:

- The likelihood $\mathcal{L}_n(\theta)$ is viewed as a function of the unknown $\theta \in R^k$ while $f(x_1, \ldots, x_n; \theta)$ is a function on $R^n$
- $f(x_1, \ldots, x_n; \theta)$ is a deterministic function while $\mathcal{L}_n(\theta)$ is a random function
- $\mathcal{L}_n(\theta)$ is not a density function (it doesn't integrate to 1 with respect to $\theta$).

**Definition** (Maximum Likelihood Estimator (MLE))**.** Then the MLE $\hat{\theta}_n$ is the value that maximizes the likelihood,

$$\hat{\theta}_n = \text{argmax}_\theta \, L_n(\theta)$$

**Solving**

1. It's often easier to work with the log-likelihood function
$$\ell_n(\theta) = \log \mathcal{L}_n(\theta) = \sum_{i=1}^{n} \log f(X_i; \theta)$$
2. If the log-likelihood is differentiable with respect to $\theta$, possible candidates for the MLE are those in the interior of $\Theta$ that solve
$$\frac{\partial}{\partial \theta_j} \ell_n(\theta) = 0, \quad j = 1, \ldots, k$$
We still need to check that we've found the *global* maximum.
3. If the maximum occurs on the boundary of $\Theta$, the first derivative may not be zero at the maximum *over* $\Theta$.
4. It's not always possible to maximize the likelihood analytically, and in these cases we turn to numerical maximization methods.
5. We are not guaranteed in general that there is a unique solution $\hat{\theta}$, or that there even is a solution.

**Exercise 9.** What does the log-likelihood look like for data that i.i.d. data from a distribution in the exponential family? What about the solution when you set the gradient to 0? Assume we write the density in the canoncial form given above using $\eta$.

**Example (Wasserman, p. 123)** Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$. Find the MLE for $\theta$.
$$\mathcal{L}(\mu, \sigma^2) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{1}{2\sigma^2}(X_i - \mu)^2$$
$$\ell(\mu, \sigma^2) = -\frac{n}{2} log(2\pi\sigma^2) + \sum_i \{-\frac{1}{2\sigma^2}(X_i - \mu)^2\}$$

We will solve the MLE by setting the gradient to equal zero
$$0 = \frac{\partial \ell(\mu, \sigma^2)}{\partial \mu}$$
$$0 = \frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2}$$

Solving this will give $\hat{\mu} = \bar{X}$ and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_i (X_i - \bar{X}_n)^2$$

To see how this works, first let's manipulate $\ell(\mu, \sigma^2)$ a bit more:

$$\begin{aligned}
\ell(\mu, \sigma^2) &= -\frac{n}{2} log(\sigma^2) - \frac{1}{2\sigma^2}\sum_i (X_i - \bar{X}_n + \bar{X}_n - \mu)^2 + constant \\
&= -\frac{n}{2} log(\sigma^2) - \frac{1}{2\sigma^2}\sum_i (X_i - \bar{X}_n)^2 - \frac{1}{2\sigma^2}\sum_i (\bar{X}_n - \mu)^2 + constant
\end{aligned}$$

We solve the system of equations (by setting the gradient =0),

$$\begin{aligned}
0 &= \frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} \\
&= \frac{1}{2\sigma^2}\sum_i \frac{\partial}{\partial \mu}(\bar{X}_n - \mu)^2 \\
&= -\frac{1}{2\sigma^2}\sum_i 2(\bar{X}_n - \mu) \Rightarrow \mu = \bar{X}_n \\
0 &= \frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} \\
&= -\frac{n}{2}\frac{\partial}{\partial \sigma^2} log(\sigma^2) - \frac{\partial}{\partial \sigma^2}\frac{1}{2\sigma^2}\left(\sum_i (X_i - \bar{X}_n)^2 + \sum_i (\bar{X}_n - \mu)^2\right) \\
&= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\left(\sum_i (X_i - \bar{X}_n)^2 + \sum_i (\bar{X}_n - \mu)^2\right)
\end{aligned}$$

$$\begin{aligned}
\mu = \bar{X}_n \Rightarrow & \\
\frac{n}{2}\frac{1}{\sigma^2} &= \frac{1}{2(\sigma^2)^2}\sum_i (X_i - \bar{X}_n)^2 \\
\hat{\sigma}^2 &= \frac{1}{n}\sum_i (X_i - \bar{X}_n)^2
\end{aligned}$$

Furthermore, we can verify that it is a maximum by looking at the Hessian.

**Exercise 10.** Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$, but with the restriction $\Theta = [0, \infty)$. Find the MLE for $\theta$.

**Exercise 11.** Let $X_1, \ldots, X_n \overset{iid}{\sim} Unif(0, \theta)$. Find the MLE for $\theta$.

**Exercise 12.** Let $X_1, \ldots, X_n \overset{iid}{\sim} Unif(\theta, \theta + 1)$. Find the MLE for $\theta$.

# 5 Asymptotic Properties of MLE (Univariate)

To begin with, we will focus only on settings where $\theta \in R$, and will generalize to talk about multivariate $\theta$ afterward.

The following three properties are asymptotic properties of the MLE *under certain conditions*:

1. Existance and Consistency: With probability approaching 1 the MLE $\hat{\theta}$ exists and is consistent,
$$\hat{\theta}_n \xrightarrow{P} \theta^*,$$
where $\theta^*$ is the true value of the parameter.[1]
2. Asymptotic normality: $(\hat{\theta}_n - \theta^*)/se(\hat{\theta}_n) \xrightarrow{D} N(0, 1)$.
3. Asymptotic efficiency: The MLE has the smallest asymptotic variance among asymptotically normal estimators.

These properties will hold, if we assume certain **regularity conditions** on the density $f$ of the observed data. We will discuss (briefly) those regularity conditions later.

**Optimization algorithms** Notice that we don't have an analytical expression for $\hat{\theta}_n$. We just say that it is that value that satisfies an optimization criteria:

$$\hat{\theta}_n = \text{argmax}_\theta \, \ell_n(\theta).$$

Despite not having an equation, we can say something about it's asymptopic properties!

[1] Actually, the exact statement is that with probability tending to 1 as $n \to \infty$, the likelihood equation $\frac{\partial}{\partial \theta} \ell_n(\theta) = 0$ has a root $\hat{\theta}_n$ such that $\hat{\theta}_n \xrightarrow{P} \theta^*$. (Lehman Theorem 3.7)

## 5.1 Asymptotic Normality and Fisher Information (Univariate)

Asymptotic normality of $\hat{\theta}_n$ means:

$$\frac{\hat{\theta}_n - \theta}{se(\hat{\theta}_n)} \Rightarrow N(0,1)$$

In order to use asymptotic normality in any practical sense, however, we need to know what is $se(\hat{\theta}_n)$.

Recall from our first module, however, that operationally we don't to actually know $se(\hat{\theta}_n)$ for any specific $n$. We just need a quantity $\hat{se}(\hat{\theta}_n)$ so that

$$\frac{\hat{\theta}_n - \theta}{\hat{se}(\hat{\theta}_n)} \Rightarrow N(0,1)$$

The *Fisher Information* will give us that quantity.

**Definition** (Fisher Information)**.** The **Fisher information** (based on $n$ observations) is

$$I_n(\theta) \;\; = \;\; Var_\theta\left(\frac{\partial}{\partial\theta}\ell_n(\theta)\right)$$

**Definition** (Score Function)**.** Notice the critical quantity

$$s_n(X;\theta) = \frac{\partial}{\partial\theta}\ell_n(\theta) = \frac{\partial}{\partial\theta}f(X_1,\ldots,X_n;\theta)$$

which we call the **Score Function**

When we maximize the likelihood, we set the score function to 0:

$$\frac{\partial}{\partial\theta}\ell(\theta) = s(X;\theta) = 0$$

Furthermore, assuming our density allows us to switch the integral and derivative, we have that at the true value of $\theta$, on average $s(X;\theta)$ is actually 0. This means that our MLE has the property that $s(X;\hat{\theta}_n) = 0$ and our true $\theta$ has the property on average $s(X;\theta)$ is 0. So we can think of the MLE as picking the $\hat{\theta}_n$ so that the estimated value of $s(X;\theta)$ (i.e. by plugging in $\hat{\theta}_n$) matches its expected value.

**Theorem.** *Assuming that our density $f$ is smooth enough to switch the order of integration and differentiation*[2]*,*

$$E_\theta s_n(X;\theta) = 0.$$

[2]Our regularity conditions for consistency and asymptotic normality will imply this is satisfied.

*Proof.* We rely on the simple fact from the chain rule that

$$\frac{\partial}{\partial\theta}\log f(X_i;\theta) = \frac{1}{f(X_i;\theta)}\frac{\partial}{\partial\theta}f(X_i;\theta).$$

Then we have

$$\begin{aligned}
E_\theta s_n(X;\theta) &= E_\theta \frac{\partial}{\partial\theta}\log f(X_i;\theta) \\
&= \int \frac{\partial}{\partial\theta}\log f(x_i;\theta) f(x_i;\theta)dx \\
&= \int \frac{1}{f(X_i;\theta)}\frac{\partial}{\partial\theta}f(X_i;\theta) f(x_i;\theta)dx \\
&= \int \frac{\partial}{\partial\theta}f(X_i;\theta)dx \\
&= \frac{\partial}{\partial\theta}\int f(X_i;\theta)dx \\
&= \frac{\partial}{\partial\theta}1 \\
&= 0
\end{aligned}$$

□

**Alternative Formula for $I_n(\theta)$** When $E_\theta s_n(X;\theta) = 0$, we can usually rewrite the Fisher Information in an easier form to calculate formula:

$$I_n(\theta) = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\ell_n(\theta)\right] = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X_1,\ldots,X_n;\theta)\right]$$

*Proof.* The proof is similar manipulations as above.

$$\begin{aligned}
-E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right] &= -E_\theta\left[\frac{\partial}{\partial\theta}\left(\frac{1}{f(X;\theta)}\frac{\partial}{\partial\theta}f(X;\theta)\right)\right] \\
&= -E_\theta\left[-\frac{1}{f(X;\theta)^2}\left(\frac{\partial}{\partial\theta}f(X;\theta)\right)^2 + \frac{1}{f(X;\theta)}\frac{\partial^2}{\partial\theta^2}f(X;\theta)\right] \\
&= E_\theta\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2 - E_\theta\left[\frac{1}{f(X;\theta)}\frac{\partial^2}{\partial\theta^2}f(X;\theta)\right] \\
&= I(\theta) - \int\left[\frac{1}{f(x;\theta)}\frac{\partial^2}{\partial\theta^2}f(x;\theta)\right]f(x;\theta)dx \\
&= I(\theta) - \frac{\partial}{\partial\theta}\underbrace{\int\left[\frac{1}{f(x;\theta)}\frac{\partial}{\partial\theta}f(x;\theta)\right]f(x;\theta)dx}_{=E_\theta\frac{\partial}{\partial\theta}\log f(X;\theta)=0}
\end{aligned}$$

The last step uses the fact that

$$E_\theta \frac{\partial}{\partial \theta} \log f(X;\theta) = E_\theta \frac{\frac{\partial}{\partial \theta} f(X;\theta)}{f(X;\theta)}$$

and $E_\theta s(X;\theta) = E_\theta \frac{\partial}{\partial \theta} \log f(X;\theta) = 0$) □

**The case of $i.i.d$ data** We can usually write the Fisher Information more simply if we assume that we are dealing with a likelihood based on $i.i.d$ data. In this case, the score function simplifies to the sum of a score function evaluated on each data point:

$$s_n(X;\theta) = \frac{\partial}{\partial \theta} \log f(X_1, \ldots, X_n; \theta) = \sum \frac{\partial}{\partial \theta} \log f(X_i;\theta) = \sum s(X_i;\theta)$$

(I drop the $n$ to differentiate between $s_n$ evaluated on all data to $s$ evaluated on each data point.)

Then our equation for the Fisher's information is

$$\begin{aligned}
I_n(\theta) &= Var_\theta \left( \frac{\partial}{\partial \theta} \ell_n(\theta) \right) \\
&= Var_\theta \left( \sum_{i=1}^{n} s(X_i;\theta) \right) \\
&= \sum_{i=1}^{n} V_\theta(s(X_i;\theta)) \quad \text{(if } X_1, \ldots, X_n \text{ are independent)} \\
&= nV_\theta(s(X_1;\theta)) \quad \text{(if } X_1, \ldots, X_n \text{ are identically distributed)} \\
&= nI_1(\theta) \equiv nI(\theta)
\end{aligned}$$

Again, notice the distinction between $I_n(\theta)$ and $I(\theta)$. $I_n(\theta)$ is the Fisher information on the entire likelihood across all $n$ observations – the variance of the sum of the score function across all observations. $I(\theta)$ is the variance of the score function on a single observation. With i.i.d. data, we can worry about finding $I(\theta)$, and then that gives us $I_n(\theta)$ by multiplying by $n$.

We can similarly simplify the expression of $I(\theta)$:

$$I(\theta) = -E_\theta \left[ \frac{\partial^2}{\partial \theta^2} \log f(X;\theta) \right]$$

**Example:** Suppose $X_1, \ldots, X_n \overset{iid}{\sim} Pois(\lambda)$. Calculate $I_n(\lambda)$.

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}\log f(X;\theta) &= \frac{\partial^2}{\partial\theta^2}\log\frac{\theta^x e^{-\theta}}{x!}\\ &= \frac{\partial^2}{\partial\theta^2}x\log\theta-\theta-\log x!\\ &= \frac{\partial}{\partial\theta}\frac{x}{\theta}-1\\ &= -\frac{x}{\theta^2}\end{aligned}$$

This gives

$$I(\theta) = -E_\theta\left(-\frac{x}{\theta^2}\right) = \frac{1}{\theta}$$

So that $I_n(\theta) = \frac{n}{\theta}$.

## 5.2 Estimating *se* of MLE

Now we can relate can relate the asymptotic variance of $\hat{\theta}$ to $I(\theta)$,

$$se(\hat{\theta}_n) \approx \sqrt{1/I_n(\theta)}$$

Specifically,

**Theorem 2** (Asymptotic normality of MLE)**.**

$$\frac{\hat{\theta}_n-\theta}{\sqrt{1/I_n(\theta)}} \xrightarrow{D} N(0,1)$$

*For i.i.d. data, this can be written as*

$$\frac{\sqrt{n}(\hat{\theta}_n-\theta)}{\sqrt{1/I(\theta)}} \xrightarrow{D} N(0,1)$$

This second formulation is only for i.i.d data and in that case it is prefered because $I(\theta)$ doesn't depend on $n$, so it's clear how this quantity changes with $n$.

**Interpretation** What does "information" in Fisher's information really tell you? Let

$$K_n(\theta) = -\frac{\partial^2}{\partial\theta^2}\ell_n$$

Notice that $K_n(\theta)$ is a random variable, why?

Then $K_n(\theta)$ measures the curvature of the log-likelihood function for any $\theta$. At the MLE, we know that we are finding a maximum, and $K_n(\hat{\theta})$ tells us how steeply curved away the likelihood is at our maximum. .

Fisher's Information $I_n(\theta)$ is it's expectation,

$$I_n(\theta) = E(K_n(\theta)).$$

So $I_n(\theta)$ measures the average value of the curvature of the log-likelihood, over draws of $X$, at $\theta$ (the truth). Large values of average curvature of the likelihood result in low asymptopic variance of your estimator.

### 5.2.1 Estimating $se(\hat{\theta}_n)$

Estimating $se(\hat{\theta}_n)$ becomes estimating $I(\theta)$ (or $I_n(\theta)$).

We can estimate $I_n(\theta)$ from the data in two ways:

1. **Expected Fisher information** is given by $I_n(\hat{\theta}_n) = E_{\theta=\hat{\theta}}(K_n(\theta))$
2. **Observed Fisher information**

$$K_n(\hat{\theta}_n) = -\frac{\partial^2}{\partial\theta^2}\ell_n(\theta)|_{\theta=\hat{\theta}_n}$$

In practice the Observed Fisher information is more commonly recommended.

We can use these estimates to construct approximate $1-\alpha$ confidence intervals for $\theta$,

$$\hat{\theta}_n \pm z_{1-\alpha/2} I_n(\hat{\theta}_n)^{-1/2}$$

or

$$\hat{\theta}_n \pm z_{1-\alpha/2} K_n(\hat{\theta}_n)^{-1/2}$$

The reason we can do this, is that both of these estimators are usually consistent estimators for estimating $I(\theta)$:

1. If $I_n(\theta)$ is continuous function of $\theta$ and the MLE is consistent, then:

$$\frac{1}{n} I_n(\theta) = I(\hat{\theta}_n) \xrightarrow{P} I(\theta)$$

2. Consistency for $K_n(\hat{\theta}_n)$ has more restrictive requirements, but holds in most common settings:

$$\frac{1}{n}K_n(\hat{\theta}_n) \xrightarrow{P} I(\theta)$$

Consistency of these estimators means that asymptotic normality still holds replacing $I(\theta)$ with either of these estimates.

**Example** For the example of the $Poisson(\lambda)$, the MLE is $\hat{\lambda}_n = \bar{X}_n$, and $I(\lambda) = 1/\lambda$. So $I_n(\hat{\theta}_n) = \frac{n}{\bar{X}_n}$.

For the Observed Fisher information, we have

$$K_n(\hat{\lambda}_n) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial\lambda^2} \log f(X_i;\lambda)|_{\lambda=\bar{X}_n} = \sum \frac{X_i}{\bar{X}_n^2} = \frac{n}{\bar{X}_n}$$

So we get the same result.

So this gives us a 95% confidence interval for $\lambda$,

$$\bar{X}_n \pm 1.96\sqrt{\frac{\bar{X}}{n}}$$

Of course, this is the same result we get from using the CLT on $\bar{X}_n$.

**Exercise 13.** Under each of the following models, find the MLE for $\theta$ and calculate an approximate 95% confidence interval using the limiting normal distribution.

1. $X_1, \ldots, X_n \overset{iid}{\sim} Exp(\theta)$
2. $X_1, \ldots, X_n \overset{iid}{\sim} Binomial(m, \theta)$ for known $m$
3. $X_1, \ldots, X_n \overset{iid}{\sim} Normal(\theta, \sigma^2)$ for known $\sigma^2$

## 5.3 Efficiency

The MLE is generally not the only possible estimator of a parameter, and is also often not the only possible estimator that is consistent or asymptotically normal.

For example, if $X_1, \ldots, X_n \overset{i.i.d}{\sim} N(\theta, \sigma^2)$, then the MLE $\hat{\theta}$ satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta) \Rightarrow N(0, \sigma^2)$$

while the median $\tilde{\theta}_n$ satisfies

$$\sqrt{n}(\tilde{\theta}_n - \theta) \Rightarrow N(0, \sigma^2 \frac{\pi}{2}).$$

It seems pretty reasonable in this case that for two consistent estimators, the MLE would be prefered since the variance of the estimator will be smaller.[3]

In general, if we require the same regularity conditions, the MLE is **efficient**, meaning it has the smallest asymptotic variance amongst other asymptotically normal estimators:

**Theorem 3.** *If $\tilde{\theta}_n$ is some other estimator s.t. $\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{D} N(0, v(\theta))$, then $v(\theta) \geq 1/I(\theta)$ for all $\theta$.*

For example, this would be a reason we prefer the MLE estimators to MOM estimators.

## 5.4 Regularity conditions

What are the conditions required for the asymptotic properties of the MLE? The conditions are called **regularity conditions** and generally involve the smoothness of the density $f(x;\theta)$.

There are different choices of conditions that will guarantee the asymptotic properties we want. The ones I'm going to give here are common and easily checkable, but they are not *necessary* conditions, meaning that you *could* have these asymptotic properties be true even if these conditions do not hold. For example, the conditions I give here assume i.i.d data. That doesn't mean these properties of the MLE don't ever hold for dependent data; just that these are not the conditions you would need to check.

To show consistency we do not need as many assumptions as the other two properties, which need additional conditions *added to* the conditions for consistency.

**Shared Regularity Conditions** The following are the regularity conditions on $f(x;\theta)$ shared by all three properties. These conditions are sufficient to give consistency, but additional conditions are needed for the other two properties:

1. **(i.i.d)** $X_1, \ldots, X_n$ are *iid* with density $f(x;\theta)$.

[3] Of course there are other considerations, for example the median is more robust to outliers. So in finite samples, we might care to use the median.

2. **(Identifiability)**If $\theta \neq \theta'$, then $f(x;\theta) \neq f(x;\theta')$. This is called **Identifiability** of the model. Meaning that different values of the parameter $\theta$ result in different distributions of the data.

   I would note that identifiability of a parameter is an important concept beyond MLEs. If multiple $\theta$ can give you the same probability distribution of your data, then no matter how much data you collect, you will not be able to distinguish between the multiple $\theta$.

3. **(common support)** The densities $f(x;\theta)$ have common support, in other words
$$\{x : f(x;\theta) > 0\}$$
   is the same for all $\theta$.

4. **(truth is interior point)** The parameter space $\Theta$ contains an open set $\omega$ of which the true parameter value $\theta^*$ is an interior point.

5. **(f differentiable)**The function $f(x;\theta)$ is differentiable with respect to $\theta$ in $\Omega$.

Exponential families all satisfy these conditions.

**Exercise 14.** Consider $f(x;\theta) = U(0,\theta)$. What property does this not satisfy?

**Exercise 15.** Consider $f(x;\theta) = Binomial(n,\theta)$, with $\theta \in [0,1]$. Since $x$ only takes on discrete values, does this mean that it violates the differentiabity requirement?

**Conditions for Asymptotic normality and efficiency** We will add additional requirements *in addition to these* for asymptotic normality and efficiency to hold. We need a bunch of further smoothness conditions on the density $f$ beyond differentiability and conditions on $I(\theta)$, which we won't go into.[4]

**What do you need to know?** First of all, regularity conditions are only relevant for asymptotic results. That means they are mainly important for constructing confidence intervals; you can calculate MLEs without verifying these assumptions (assuming the MLE exists). Furthermore, these assumptions are finicky and I haven't given very precise definitions here, so unless I ask explicitly, you do not need to show regularity conditions in homeworks or exams to construct a confidence interval or determine the asymptotic standard error.

[4]Lehman Theorem 2.6 p441 + Theorem 3.1 p449 gives an exact set of conditions.

### 5.4.1 A side note on Wasserman's Theorem on Consistency (Optional)

In fact the above conditions are more than are needed (sufficient but not necessary).

Wasserman gives a theorem (Theorem 9.13) for consistency that has a completely different set of requirements and look nothing the same. The requirements I give above are far superior for functionality (i.e. being able to actually check them), but Wasserman's cover wider class of models.

Wasserman's requirements are more like the most minimal structure of what you need to be true for consistency – once you know his two requirements are true, then the conclusion of consistency follows pretty immediately. It is pretty useful, however, to look at the conditions because

- The section on consistency (9.5) introduces Kullback Leibler divergence, which is a useful measure for quantifying how well one distribution approximates another.
- The conditions shows the general structure of what needs to be true for optimization algorithms to give consistent results, not just the MLE. In otherwords, if you had an estimate that came from an optimization problem (not just the MLE), if you could show these conditions hold, you would probably be able to use this same proof to get consistency.

**Definition** (Kullback-Leiber Divergance)**.** The Kullback-Leiber Divergence[5] between two distributions $F$ and $G$ with pdfs $f$ and $g$, is given by

$$D(f,g) = \int f(x) \log(\frac{f(x)}{g(x)}) = E_f\{\log f(x) - \log g(x)\} \geq 0.$$

You can show that $D(f,g) = 0$ if and only if $f = g$, i.e. the two distributions are the same distribution. Notice that the definition implicitly makes the assumption that $g$ is strictly positive on the support of $F$.

If $f_{\theta_1}$ and $f_{\theta_2}$ are pdfs from the same parametric family and differ only in the value of $\theta$, we write

$$D(f_{\theta_1}, f_{\theta_2}) := D(\theta_1, \theta_2)$$

[5]Wasserman calls it a "distance", even though he notes it isn't a distance because doesn't satisfy the most basic properties of a distance (not symmetric). I prefer to use the term "Divergance" to make that clear

Wasserman's Consistency conditions consider the following quantity:

$$\begin{aligned} M_n(\theta) &= \frac{1}{n}\sum_i log\frac{f(X_i;\theta)}{f(X_i;\theta_*)} \\ &= \frac{1}{n}\ell_n(\theta) - \frac{1}{n}\ell_n(\theta_*) \end{aligned}$$

where $\theta_*$ is the true (unknown) value of $\theta$.

Then notice that for the MLE, $\hat{\theta}$ is the value of $\theta$ that maximizes $M_n(\theta)$ (because $\frac{1}{n}\ell_n(\theta_*)$ is just a unknown constant value it doesn't affect the maximization).

Further, notice that

$$E_{\theta_*} log\frac{f(X_i;\theta)}{f(X_i;\theta_*)} = -D(\theta_*,\theta) := M(\theta)$$

So $M_n(\theta)$ is an empirical estimate of $-D(\theta_*,\theta) = M(\theta)$ (i.e. replacing average over i.i.d data for expectation), so we expect that

$$M_n(\theta) \approx M(\theta)$$

We know that $M(\theta)$ is maximized at zero when $\theta = \theta_*$ and otherwise less than zero . So maximizing $M_n(\theta)$ should give us a $\hat{\theta}$ near the truth $\theta_*$.

**Wasserman's Consistency Requirements** Wasserman then gives the folowing requirements for consistency:

1. The empirical estimate $M_n(\theta)$ converges to the truth $M(\theta)$,

$$sup_\theta |M_n(\theta) - M(\theta)| \xrightarrow{P} 0$$

2. $M(\theta)$ is well-behaved, so that for $\theta$ near $\theta_*$, you won't find the maximum at a place far away from the truth (namely $M(\theta_*)$):

$$sup_{\theta:|\theta-\theta_*|\geq\epsilon} M(\theta) < M(\theta_*)$$

These two conditions are the general quantities you need to be true when you try to optimize a function of $\theta$ to get an estimate of the true $\theta_*$.

But the second one, in particular, is trivial in the case of the MLE if we have an *identifiable model*, because in the case of the MLE

$$M(\theta_*) = -D(\theta_*,\theta_*) = 0$$

and

$$M(\theta) = -D(\theta_*,\theta) < 0, \theta \neq \theta_* \text{ (Assuming identifiability)}$$

But for other optimization algorithms or settings it might not be the case.

## 5.5 Summary for Exponential families

For an exponential family, notice that all of these things come together, including the regularity conditions:

1. MOM and MLE are equivalent (check by yourself).
2. MLE is consistent.
3. MLE is asymptotically normal after an appropriate linear transformation.
4. MLE is asymptotically efficient (smallest asymptotic variance defined through fisher information).

# 6 Estimating Functions of $\theta$ (Univariate)

Sometimes the parameter we are interested in are not the parameters that naturally describe the distribution. In particular, in our above notation, $\theta$ is the parameter that parameterizes the distribution/density. We might be interested in estimating other quantities than just the elements of $\theta$.

For example, Wasserman describes a situation where $X_i$ are $N(\mu, \sigma^2)$ and are the result of a diagnostic test where $X_i > 1$ indicates a positive result. Then we might want to estimate the probability of a positive in the population,

$$\tau = P(X > 1) = 1 - \Phi(\frac{1-\mu}{\sigma}),$$

where $\Phi$ is the cdf of a standard normal. $\tau$ would be our parameter of interest and is not naturally a parameter that defines the normal distribution, but rather a function of both $\mu$ and $\sigma$, $\tau = g(\mu, \sigma^2)$.

This above example is where $\theta$ is a multivariate parameter ($\theta \in R^2$), but we are going to focus first with a univariate parameter, and then expand to multivariate parameters.

## 6.1 Reparameterization with 1-1 functions

When $\tau = g(\theta)$ is a 1-1 function of $\theta$ we can simply **reparameterize** the distribution to be with respect to $\tau$ instead of $\theta$. Meaning that we started with the family of

distributions being written in terms of $\theta$, $F(x;\theta)$, but we can rewrite the distribution to be $F(x;\tau)$ by substituting in $g^{-1}(\tau)$ for $\theta$ in the density.

So for example, if $E_\theta(X) = f(\theta)$, then

$$E_\tau(X) = E_{\theta=g^{-1}(\tau)}(X) = f(g^{-1}(\tau)).$$

**Example: Poisson** For example, suppose we have $X \sim Poisson(\theta)$. We could be interested in considering $\tau = g(\theta) = 1/\theta$. Then our standard density is

$$f(x;\theta) = \frac{\lambda^x e^{-\theta}}{x!}, \lambda > 0$$

Since $g$ is invertible, we can write $\theta = g^{-1}(\tau) = 1/\tau$, and our density of $X$ can equivalently be written as

$$h(x;\tau) = f(x; g^{-1}(\tau)) = \frac{(1/\tau)^x e^{-1/\tau}}{x!} = \frac{e^{-1/\tau}}{x!\tau^x}, \tau > 0$$

I use $h$ to emphasize that this is a different density function, even though it describes the same probability distribution for $X$. Also note that constraints on $\theta$ will carry over to induce constraints on $\tau$.

**Example: Multivariate $\theta$** Suppose, for example, in the problem where $X_i$ are i.i.d $N(\mu, \sigma^2)$, and we had $\tau = (\tau_1, \tau_2) = (\mu/\sigma, 1/\sigma)$. Then this is a 1-1 function between $\theta$ and $\tau$ and we could rewrite the density of our $X_i$ as

$$\frac{\tau_2}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(\tau_2 x - \tau_1)^2\}, \tau_2 > 0$$

**Functions of Parameters vs. Functions of Random Variables** Notice the difference between taking functions of the parameters (reparameterization) compared to taking functions of random variables. For reparameterization, I can just substitute in $g^{-1}(\tau)$ in place of $\theta$ to get the new density expressed in terms of $\tau$. But if I define a new random variable $Y = g(X)$ based on an invertible function $g$ then in order to get the density of $Y$, I have[6]

$$f_Y(y) = f_x(g^{-1}(y))\frac{d}{dy}g^{-1}(y).$$

[6] I give the equation for a univariate random variable, but it generalizes to multivariate $Y$.

**Non invertible functions** It is only possible to rewrite the density $f(x;\theta)$ with respect to $\tau$ if the function is invertible. Notice that I *cannot* do this with all functions $g$, only invertible functions (bijections).

Consider for example, $X \sim N(\mu, 1)$. Suppose $\tau = \mu^2$. It is not possible to write down the density of $X$ knowing only $\tau$. We can see this because the density of $X \sim N(3, 1)$ is quite different than $X \sim N(-3, 1)$, but $\tau = 9$ for both cases. So I cannot make probability calculations for $X$ knowing only $\tau = 9$ – the probability $P(2 < X < 3)$ is quite different for those two different values of $\mu$, and knowing only $\tau = 9$ I can't tell you which probability statement is correct. This means I cannot define the distribution of $X$ based on $\tau$.

**Exercise 16.** Can I reparameterize the $Binomial(n, p)$ with respect to $\tau = p^2$? If so, give $h(x;\tau)$

**Exercise 17.** Consider the negative binomial distribution, a discrete distribution defined on integer values,

$$P(X = x) = f(x;\theta) = \binom{x + r - 1}{x}(1 - p)^x p^r$$

where $\theta = (p, r)$, and $p \in (0, 1)$ and $r > 0$.

Consider the following function from $R^2 \to R^2$,

$$\tau = g(p, r) = (\frac{1 - p}{p}r, \frac{1 - p}{p^2}r)$$

Can you reparameterize the density with respect to $\tau$? If so, give $h(x;\tau)$

**Exercise 18.** Can we reparameterize the density with respect to $\tau$ in the above problem where

$$\tau = P(X > 1) = 1 - \Phi(\frac{1 - \mu}{\sigma})?$$

If so, give $h(x;\tau)$

## 6.2 MLE of $\tau$

Suppose you have the MLE for the parameter $\theta$, but you are interested in $\tau = g(\theta)$ What is the MLE of $\tau$?

The answer is simple: $\hat{\tau}_n = g(\hat{\theta})$. This is the **equivalence or invariance property of the MLE** and is another reason that MLE estimators are attractive (MOM estimators, for example, do not have this propoerty).

But let's consider what this means more carefully by considering two settings: $g$ is invertible versus not.[7]

**$g$ invertible** Let's first ignore that we have a MLE $\hat{\theta}$. If $g$ is invertible, I can reparameterize the problem with respect to $\tau$. So I can define the joint density $h_n(x;\tau)$ and write down the likelihood of $\tau$,[8]

$$\ell_n^h(\tau) = \log h_n(x;\tau)$$

So the MLE $\hat{\tau}$ is

$$\hat{\tau}_n = argmax_\tau \ell_n^h(\tau)$$

We also have a MLE $\hat{\theta}_n$ of $\theta$ based on the likelihood $\ell_n^f(\theta) = \log f_n(x;\theta)$

So the question is whether $\hat{\tau}_n$ that maximizes $\ell_n^h(\tau)$ equivalent to $g(\hat{\theta}_n)$?

We know that

$$h_n(x;\tau) = f_n(x; g^{-1}(\tau))$$

so we have

$$\ell_n^h(\tau) = \log h_n(x;\tau) = \log f_n(x; g^{-1}(\tau))$$

If $\hat{\theta}$ maximizes $\log f_n(x;\theta)$ then if we choose $\hat{\tau}$ so that $g^{-1}(\hat{\tau}) = \hat{\theta}$, then $\hat{\tau}$ must also maximize $\log f_n(x; g^{-1}(\tau))$, which means it also maximizes $\log h_n(x;\tau)$. This means that setting $\hat{\tau} = g(\hat{\theta})$ will give a MLE for $\tau$.[9]

**Non-invertible** What about $\tau = g(\theta)$ where $g$ is not invertible? This is trickier, because as we saw above, we cannot rewrite our density $f(x;\theta)$ with respect to $\tau$. This means we don't even have a definition of what is the likelihood with respect to $\tau$.

So we need a definition of what the MLE of $\tau$ actually means. We **define** the MLE of $\tau$ as the value $\tau$ maximizes what is known as the *induced likelihood function*

[7] We would note that Wasserman's "proof" only covers the case of $g$ invertible.

[8] In this section only, I will write $\ell^f$ and $\mathcal{L}^f$ to be clear which density the likelihood is using.

[9] Notice that we aren't guaranteed a unique MLE.

**Definition 6.1** (MLE of $\tau = g(\theta)$)**.** We write the *induced likelihood function* of $\tau$ as

$$\mathcal{L}^*(\tau) = \sup_{\theta:g(\theta)=\tau} \mathcal{L}^f(\theta)$$

The MLE $\hat{\tau}$ is **defined** as the $\tau$ that maximizes this quantity

$$\hat{\tau} = \text{argmax}_\tau \, \mathcal{L}^*(\tau)$$

In other words, $\mathcal{L}^*(\tau)$ is the largest value of $\mathcal{L}^f(\theta)$, when maximized over all $\theta$ where $g(\theta) = \tau$.

Notice that this definition works also if $g$ is invertible. Then we have

$$\mathcal{L}^*(\tau) = \sup_{\theta:g(\theta)=\tau} \mathcal{L}^f(\theta) = \mathcal{L}^f(g^{-1}(\tau)) = \mathcal{L}^h(\tau),$$

i.e. you can just equivalently rewrite the density with respect to $\tau$. This is just reparameterizing the density with respect to $\tau$ instead of $\theta$.

**Exercise 19.** Consider $\tau = \theta^2$ What is the induced likelihood function?

**Equivalence/Invariance of MLE** With this definition of the MLE of $\tau$, you can similarly go from the MLE of $\theta$ to the MLE of $\tau$ for any $g$.

**Theorem 4.** *Let $\tau = g(\theta)$ be a function of $\theta$. Let $\hat{\theta}_n$ be the MLE of $\theta$. Then $\hat{\tau}_n = g(\hat{\theta}_n)$ is the MLE of $\tau$.*

This is true even with $g$ is not 1-1, but Wasserman's proof only covers the case of one-to-one functions that we outlined above.

*Proof.* Suppose first that $g$ is one-to-one. For each $g^{-1}(\tau)$ there is a single $\theta = g^{-1}(\tau)$, so for any particular $\tau_0$ and corresponding $\theta_0$, we have for all $\tau_0$

$$\mathcal{L}^*(\tau) = \mathcal{L}(\underbrace{g^{-1}(\tau)}_{\theta}) \leq \mathcal{L}(\hat{\theta}_n)$$

But because $g$ is one-to-one, $\hat{\theta}_n$ corresponds to a unique $\hat{\tau}_n = g(\hat{\theta}_n)$ and

$$\mathcal{L}(\hat{\theta}_n) = \mathcal{L}^*(\hat{\tau}_n)$$

so $\hat{\tau} = g(\hat{\theta})$ maximizes $\mathcal{L}^*$.

What if $g$ is not 1-1? The general case is only slightly more complicated. We show that the maxima of both $\mathcal{L}^*(\tau)$ and $L(\theta)$ are the same, and that they are both attained at $\tau(\hat{\theta})$ and $\hat{\theta}$ (Casella & Berger, p.320 for detailed proof).

□

**What about if $g(\hat{\theta})$ is not defined?** One example I gave when I was talking about reparameterization was $\tau = \sqrt{\theta}$. What about estimating $\tau = \sqrt{\theta}$? What if $\hat{\theta}_n < 0$?

Going back to principles, our induced likelihood is

$$\mathcal{L}^*(\tau) = \mathcal{L}^f(\tau^2)$$

for $\tau > 0$. This means the MLE of $\hat{\tau}$ requires maximizing $\mathcal{L}^f(\theta)$ over $\theta > 0$, so that $\hat{\tau} = g(\hat{\theta})$ only if $\hat{\theta}$ was maximizing the likelihood assuming $\theta > 0$.

Does this mean that the equivalence theorem is violated? In fact, the entire problem is ill-posed if you do not assume $\theta > 0$. You are implicitly assuming that $\theta > 0$ if you want to even estimate $\tau = \sqrt{\theta}$.

If you assume that $\theta > 0$, you have to similarly adjust your MLE of $\theta$ to be maximizing the likelihood constrained to the set $\theta > 0$. The MLE for one set of models $\mathcal{F}$ is not the same as for another set of models $\mathcal{G}$. We cannot estimate $\theta$ for $\mathcal{F} = \theta : \theta \in R$ and assume that we should use the same $\hat{\theta}$ for the set $\mathcal{G} = \theta : \theta > 0$. For this same reason, it would make no sense to think to use the MLE $\hat{\theta}$ found for the larger set $\mathcal{F}$ for estimating $\tau$, which is implicitly assuming the models $\mathcal{G}$.

However, if both estimators $\hat{\theta}$ and $\hat{\tau}$ are assuming the same set of distributions $\mathcal{F}$ could have generated the data, then the equivalence theorem above works (and for this example, $g$ now becomes invertible on the region $\theta > 0$, so its a trivial example of reparameterization).

## 6.3 Fisher's Information under Reparameterization

We can look at how the Fisher's information changes under reparameterization. We have an equation $I_{f,n}(\theta)$ based on our parameterization $\theta$ with density $f$.

Suppose that $g$ is invertible, so we can reparameterize the distribution with respect to $\tau$. Let's call that density $h(x;\tau)$. We can then write $I_{h,n}(\tau)$

$$I_{h,n}(\tau) = Var_\tau(\frac{\partial}{\partial\tau}\ell_h(\tau))$$

where $\ell_h(\tau)$ is the likelihood written with respect to the density $h(x;\tau)$.

But $h(x;\tau) = f(x; g^{-1}(\tau))$. So

$$\frac{\partial}{\partial\tau}\log h(x;\tau) = \frac{\partial}{\partial\theta}\log f(x;\theta)|_{\theta=g^{-1}(\tau)}\frac{\partial}{\partial\tau}g^{-1}(\tau)$$

Putting this back into our equation for $I_{h,n}(\tau)$, we can write this equation with respect to our original $I_{f,n}(\theta)$ – i.e. we don't have to recalculate our Fisher Information under different parameterizations.

$$\begin{aligned}
I_h(\tau) &= \frac{\partial}{\partial\tau}\ell_h(\tau) \\
&= Var_\tau(\frac{\partial}{\partial\tau}\log h(x;\tau) \\
&= Var_\tau\big(\frac{\partial}{\partial\theta}\ell(\theta)|_{\theta=g^{-1}(\tau)}\frac{\partial}{\partial\tau}g^{-1}(\tau)\big) \\
&= \big(\frac{\partial}{\partial\tau}g^{-1}(\tau)\big)^2 Var_\tau\big(\frac{\partial}{\partial\theta}\ell(\theta)|_{\theta=g^{-1}(\tau)}\big) \\
&= \big(\frac{\partial}{\partial\tau}g^{-1}(\tau)\big)^2 Var_{\theta=g^{-1}(\tau)}\big(\frac{\partial}{\partial\theta}\ell(\theta)\big) \\
&= \big(\frac{\partial}{\partial\tau}g^{-1}(\tau)\big)^2 I_f(g^{-1}(\tau))
\end{aligned}$$

**Example:** Suppose I want to reparameterize my $Poisson(\lambda)$ with respect to $\tau = g(\lambda) = 1/\lambda$. I know that $I(\lambda) = 1/\lambda$.

Then $g^{-1}(\tau) = 1/\tau$ so that

$$\big(\frac{\partial}{\partial\tau}g^{-1}(\tau)\big)^2 = \big(-\frac{1}{\tau^2}\big)^2 = \frac{1}{\tau^4}.$$

This gives us that

$$I(\tau) = \frac{1}{\tau^4}\tau = \frac{1}{\tau^3}$$

You should check that deriving $I(\tau)$ from first principles gets the same result.

## 6.4 Distribution of $\hat{\tau}$: the delta method

We have that if $\hat{\theta}_n$ is the MLE of $\theta$ then $g(\hat{\theta}_n)$ is the MLE of $g(\theta)$.

We've already seen that if $g$ is continuous, we have a Continuous Mapping theorem. This will imply that if $\hat{\theta}_n$ is consistent, then so is $g(\hat{\theta}_n)$, for continous $g$.

But what about the distribution of $g(\hat{\theta}_n)$? In general, it is not as easy mathematically to transfer the distribution of $\hat{\theta}_n$ to the distribution of $g(\hat{\theta}_n)$, but we can do so

assuming that $g$ is a smooth function (differentiable).[10]

**Theorem 5.** *Let $\hat{\theta}_n$ be the MLE which is asymptotically normal, described above. If $\tau = g(\theta)$ and $\hat{\tau}_n = g(\hat{\theta}_n)$, where $g$ is differentiable and $g'(\theta) \neq 0$, then*

$$\frac{\hat{\tau}_n - \tau}{\hat{se}(\hat{\tau}_n)} \Rightarrow N(0,1)$$

*where*

$$\hat{se}(\hat{\tau}_n) = |g'(\hat{\theta}_n)|\hat{se}(\hat{\theta}_n)$$

The proof of this is basically Taylor's expansion applied to $g$ and then the continuous mapping theorem[11]. This use of Taylor's expansion is often called the *Delta Method* when applied to random variables.

**Exercise 20.** If $g$ is invertible, so that we can reparameterize the distribution with respect to $\tau$, show that the above equation is equivalent the result you would have gotten from first principles, i.e. reparameterizing the distribution.

# 7 Asymptotic Properties of MLE (Multivariate)

We have so far only discussed the asymptotic properties of real-valued parameters $\theta$. However, we can expand all of the previous results to multivariate parameters. The results regarding asymptotic normality of the MLE follow for multivariate parameters as well.

Notice that our multivariate assumption is of our parameter $\theta$; our data might or might not be multivariate.

## 7.1 Multivariate Normal

First need to have a multivariate definitions of normality.

[10]The continuous mapping theorem tells us that if $\hat{\theta}_n \Rightarrow Z$ that $g(\hat{\theta}_n) \Rightarrow g(Z)$. But then we still need to figure out what is the distribution of $g(Z)$. So there's a bit more to show that $g(\hat{\theta}_n)$ is approximately normally distributed for any differentiable $g$.

[11]Namely that $X_n \xrightarrow{D} X$ implies $g(X_n) \xrightarrow{D} g(X)$ if $g$ continuous.

**Definition** (Multivariate normal distribution)**.** The multivariate normal distribution for a vector $Z = (Z_1, Z_2, \ldots, Z_k)'$ with mean vector $\mu \in R^k$ and positive definite covariance matrix $\Sigma \in R^{k \times k}$ has pdf

$$f(z; \mu, \Sigma) = (2\pi)^{-k/2} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu) \right\}$$

We can write this $Z \sim N_k(\mu, \Sigma)$.[12]

Let $\mu_i$ denote the $i^{th}$ element of $\mu$, and $\Sigma_{ij}$ the element of $\Sigma$ in the $i^{th}$ row and $j^{th}$ column. Then

- $Z_i \sim N(\mu_i, \Sigma_{ii})$
- $Cov(Z_i, Z_j) = \Sigma_{ij}$

**More on Covariance Matrices** Covariance matrices hold the cross covariances of the random variates $Z_i$ and $Z_j$. This is true for any random vector, not just normally distributed ones. This means all covariance matrices need to be symmetric. Furthermore, all covariance matrices need to be at least positive semidefinite. Recall a symmetric matrix $\mathbf{A}$ is **Positive Definite** if $\forall \boldsymbol{c}$, $\boldsymbol{c}^T \boldsymbol{\Sigma} \boldsymbol{c} > 0$ (and Negative Definite if $< 0$). Non-negative definite or positive semidefinite means you replace $> 0$ with $\geq 0$.

Why? For any random vector $Z$ with covariance matrix $\Sigma$, any deterministic linear combination of a random vector $Z \in R^k$, e.g. $W = u'Z$ for $u \in R^k$, has variance given by $var(W) = u'cov(Z)u \geq 0$. Therefore, if $\Sigma$ is not at least positive semidefinite, we could construct a random variable $W$ with negative variance by choosing any $u$ that makes $u'\Sigma u < 0$.

If $\Sigma$ is not positive definite, but is non-negative definite, that means there is a $\boldsymbol{c}$ so that $var(\boldsymbol{c}^T Y) = 0$ i.e $\boldsymbol{c}^T Y = a$ for some constant $a$. What does that tell us about the elements of the vector $Y$?

For a normal distribution, we require $\Sigma$ positive definite. Otherwise $\Sigma^{-1}$ in the density would not be defined.

[12]Note that if the covariance matrix $\Sigma$ is only positive semidefinite, then the density above is undefined; this is analagous to if you allow $\sigma^2$ to be zero in a univariate normal distribution. Allowing could be considered a degenerate normal distribution, and could occur if some of the $Z_i$ were deterministic linear combinations of the other, e.g. $Z_1, \ldots, Z_{k-1} \sim N(0, \Sigma)$, and $Z_k = \sum_{i=1}^{k-1} Z_i$

**I.i.d Normal values** If we have $n$ i.i.d. observations $X_1, \ldots, X_n$ each distributed $N(\theta, \sigma^2)$, we could write this as

$$X \sim N(\mu, \Sigma)$$
$$\mu = \begin{pmatrix} \theta \\ \theta \\ \vdots \\ \theta \end{pmatrix} = \theta \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \theta 1_n$$
$$\Sigma = \sigma^2 I_n$$

## 7.2 Asy. Normality of the MLE and Fisher Information

Under similar regularity conditions, the MLE $\hat{\theta}$ is asymptotically normal.

**Theorem.** *If $\hat{\theta}_n \in R^k$ is the MLE estimator of $\theta \in R^k$, then if $k$ is a fixed dimension and under sufficient regularity conditions,*

$$\hat{\theta}_n \overset{D}{\approx} N(\theta, V_n(\theta)).$$

(We will discuss later more precisely the definition of multivariate asymptotic normality below). An important additional constraint, in addition to regularity conditions, is that the number of parameters $k$ is *fixed*; we cannot have the number of parameters grow with our sample size.

**Individual elements of $\hat{\theta}$** We are usually interested in the (marginal) distributions of estimators of particular parameters, i.e. $\hat{\theta}_{n,j}$. For example, for a normal distribution $N(\mu, \sigma^2)$, we estimate $\theta = (\mu, \sigma^2)$, but we often only care about creating confidence intervals for $\mu$.

If a multivariate parameter $\hat{\theta}_n$ is asymptotically normal, this gives us the approximate marginal distribution of the individual parameters:

$$\hat{\theta}_{n,j} \overset{D}{\approx} N(\theta_j, V_n(\theta)_{jj}.$$

This means once we define $V_n(\theta)$, we have asymptotic expressions for the standard error of our individual entries,

$$se(\hat{\theta}_{n,j}) \approx \sqrt{V_n(\theta)_{jj}}.$$

Using this, we can build CI for $\theta_j$ based on a estimate of $V_n(\theta)_{jj}$,

$$\hat{se}(\hat{\theta}_j) = \hat{V}_n(\theta)_{jj}.$$

For example, if we have an estimate of $\theta$, we could set $\hat{V}_n(\theta)_{jj} = V_n(\hat{\theta})_{jj}$.[13]

### 7.2.1 Fisher Information

For this to be useful, we need to define $V_n(\theta)$. Like the univariate case, $V_n(\theta)$ can be written in terms of the *multivariate* Fisher Information.

For simplicity we will state the multivariate form of the Fisher Information that corresponds to the univariate version:

$$I_n(\theta) = -E_\theta\left(\frac{\partial^2}{\partial\theta^2}\ell_n(\theta)\right)$$

(the easiest one to calculate)

The corresponding notion of a second derivative is the Hessian matrix of the log-likelihood, which is the matrix of second partial derivatives with elements

$$H_{ij} = \frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell_n(\theta)$$

**Definition 7.1** (Multivariate Fisher information $I(\theta)$)**.** The Fisher information matrix is

$$I_n(\theta) = -E_\theta H(\log f(X;\theta)) = \begin{bmatrix} E_\theta(H_{11}) & \cdots & E_\theta(H_{1k}) \\ E_\theta(H_{21}) & \cdots & E_\theta(H_{2k}) \\ \vdots & \vdots & \vdots \\ E_\theta(H_{k1}) & \cdots & E_\theta(H_{kk}) \end{bmatrix}$$

**i.i.d. Data** Similar to the univariate case, you can write $I_n(\theta) = nI(\theta)$, for i.i.d data, where $I(\theta)$ is in terms of $\log f(X;\theta)$ rather than $\ell_n(\theta)$.

### 7.2.2 Defining $V_n(\theta)$

As in the normal case, we can define $V_n(\theta)$ with respect to the Fisher information,

$$V_n(\theta) = I^{-1}(\theta),$$

so that

$$\hat{\theta} \overset{D}{\approx} N(\theta, I_n(\theta)^{-1})$$

[13]The same issues we learned about in the first module of the class hold true here, in terms of ensuring that you can replace $se$ with an estimate of $se$. You need $\hat{se}$ to satisfy certain properties.

Notice that this is the matrix inverse.

This gives us asymptotic expressions for the standard error of individual parameter estimates,

$$se(\hat{\theta}_j) \approx \sqrt{[I_n(\theta)^{-1}]_{jj}}$$

So if we estimate $I(\theta)$, we invert the matrix, and the diagonal entries of the inverted matrix will be our estimates of $se(\hat{\theta}_j)$

If our data is i.i.d., $I_n(\theta)^{-1} = I(\theta)^{-1}/n$ and this simplifies to

$$se(\hat{\theta}_j) \approx \sqrt{[I(\theta)^{-1}]_{jj}}/\sqrt{n}$$

**Estimating $I_n(\theta)$** Note that we have the same two options for estimating $I_n(\theta)$

1. $I_n(\hat{\theta}_n) = -E_{\theta=\hat{\theta}_n} H(\ell(\theta))$

   Note that the expectation of a matrix is the expectation of the individual elements of the matrix.

2. $K_n(\hat{\theta}_n) = -H(\ell(\theta))|_{\theta=\hat{\theta}_n}$

**Confidence intervals** We will estimate $I(\theta)$ or $I_n(\theta)$ so that for i.i.d. data we have

$$\hat{se}(\hat{\theta}_j) = \sqrt{\frac{1}{n}[\hat{I}^{-1}(\theta)]_{jj}}$$

Note that this is the $jj$ element of the *inverse* of the Information matrix – you have to first invert the matrix:

$$[\hat{I}^{-1}(\theta)]_{jj} \neq \frac{1}{\hat{I}_{jj}(\theta)}$$

(unless you are in a univariate situation). This is a common mistake!

For non i.i.d data we simply use $I_n(\theta)$ instead,

$$\hat{se}(\hat{\theta}_j) = \sqrt{[\hat{I}_n^{-1}(\theta)]_{jj}}$$

**Example** Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. The MLEs for $\mu$ and $\sigma$ are $\hat{\mu}_n = \bar{X}_n$ and $\hat{\sigma}_n = \sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2}$. In addition...

$$I_n(\mu, \sigma) = \begin{bmatrix} n/\sigma^2 & 0 \\ 0 & 2n/\sigma^2 \end{bmatrix}$$

$$V_n(\mu, \sigma) = I_n(\mu, \sigma)^{-1} = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & \sigma^2/(2n) \end{bmatrix}$$

Using the fact that both $\hat{\mu}_n$ and $\hat{\sigma}_n$ are consistent, we can plug in to get

$$\hat{\mu}_n \pm 2\sqrt{\frac{\hat{\sigma}_n^2}{n}} \text{ and } \hat{\sigma}_n \pm 2\sqrt{\frac{\hat{\sigma}_n^2}{2n}}$$

as approximate 95% confidence intervals for $\mu$ and $\sigma$.

## 7.3 Multivariate Asymptotic Normality

We have written

$$\hat{\theta}_n \overset{D}{\approx} N(\theta, V_n(\theta)).$$

Again, what does does that actually mean? (Again, we are going to restrict ourselves to estimators with a well-defined covariance matrix).

To be most similar to the definition we had for univariate parameters, let $S_n$ be the covariance matrix of $\hat{\theta}$.

**Definition** (Multivariate Asymptotic Normality)**.** For $\theta$ a multivariate parameter, we have that $\hat{\theta}_n$ is asymptotically normal if

$$Z = S_n^{-1/2}(\hat{\theta} - \theta) \Rightarrow N_k(0, I_k)$$

where $N_k(0, I_k)$ is a multivariate normal with mean the vector of zeros and identity matrix as covariance.

If you aren't familiar with the square-root of a matrix, do not worry too much about this. In particular, like in the univariate situation, if

$$S_n^{-1/2}(\hat{\theta} - \theta) \sim N(0, I)$$

then we have that

$$\hat{\theta} - \theta \sim N(0, S_n)$$

so this result tells us that

$$\hat{\theta} \overset{D}{\approx} N(0, S_n)$$

Note that the definition of convergence in distribution remains the same in the case of a random vector, since the CDF of a random vector $F(x)$ is a function from $R^k \to R$ and convergence in distribution involves convergence of the CDF.

**Variations** Often, we do not know $S_n$ exactly, but instead have something like the following result,

$$\sqrt{n}(\hat{\theta}_n - \theta) \Rightarrow N_k(0, V(\theta)).$$

This would imply that

$$\hat{\theta} \overset{D}{\approx} N(0, \frac{1}{\sqrt{n}} V(\theta)).$$

This is the case for the MLE from i.i.d. data where $V(\theta) = I(\theta)^{-1}$.

More generally, can have an expression $V_n(\theta)$ so that $V_n(\theta)$ is neither the actual covariance of $\hat{\theta}$ but it depends on $n$ so we can't just put it on the right side of our limit statement. However, $S_n(\theta) \approx V_n(\theta)$ sufficiently so that we can say

$$V_n(\theta)^{-1/2}(\hat{\theta} - \theta) \sim N(0, I)$$

i.e.

$$\hat{\theta} \overset{D}{\approx} N(0, V_n(\theta)).$$

For the MLE, this occurs when we do not have i.i.d data and $V_n(\theta) = I_n(\theta)^{-1}$. $V_n(\theta)$ that is not actually the covariance matrix of the MLE but asymptotically it is close to $S_n(\theta)$.

In the univariate case, we discussed all of these different types of manifestations of how asymptotic normality of a parameter might be expressed. The same ideas are at play here, in terms of when you can make these substitutions and still maintain asymptotic normality. However, we will not get into them in the multivariate setting.

## 7.4 Consistency of the MLE

We also need a multivariate definition of consistency. Previously, we had consistency of a univariate estimator as

$$\hat{\theta}_n \xrightarrow{P} \theta$$

for all choice of $\theta \in \Theta$. This means

$$\lim_{n \to \infty} P(|\theta_n - \theta| \geq \epsilon) = 0$$

for all $\theta$.

Convergence in probability can extend to random vectors by absolute value replaced with the $L_2$ norm,[14]

[14] More generally you can use any distance you choose, so it can be extended beyond random vectors.

**Definition** (Multivariate Convergence In Probability)**.** $W_n$ converges in probability to $W$ if for every $\epsilon$, $\lim_{n\to\infty} P(||W_n - W|| \geq \epsilon) = 0$.

Then $\hat{\theta}_n$ is consistent for $\theta$ if $\hat{\theta}_n$ converges in probability to $\theta$ for all $\theta$. To emphasize that the convergence is based on the $L_2$ norm, we can equivalently write the convergence in terms of the $L_2$ norm converging in probability to 0.

**Definition** ($L_2$ Consistency)**.** An estimator $\hat{\theta}_n$ is consistent for $\theta$ if

$$||\hat{\theta}_n - \theta||_2 \xrightarrow{P} 0$$

for all choice of $\theta \in \Theta$.

Equivalently, $\hat{\theta}_n$ is consistent for $\theta$ if $\hat{\theta}_n$ converges in probability to $\theta$ for all $\theta$.

This can distinguish this concept from *pointwise convergence*, i.e. convergence of the elements of $\hat{\theta}_n$ to the elements of $\theta$

**Definition** (Pointwise Consistency)**.** An estimator $\hat{\theta}_n \in R^k$ is pointwise consistent for $\theta$ if for all $j = 1, \ldots, k$, $\hat{\theta}_{nj}$ is consistent for $\theta_j$.

If the dimension of our vector $\theta$ is fixed and does not grow with $n$, then $L_2$ consistency and pointwise consistency imply each other.

**Theorem.** *Let $\hat{\theta}_n$ be an estimator based on data $X_1, \ldots, X_n$ of $\theta \in R^k$. Assume that the dimension $k$ is fixed and does not change with $n$. Then $\hat{\theta}_n$ is $L_2$ consistent if and only if $\hat{\theta}_n$ is pointwise consistent.*

Note that the assumption of $k$ being fixed is very important. High dimensional parameter estimation, where the number of parameters $k$ is very large relative to $n$ does not follow any of the nice properties given here (including consistency of the MLE!)

**Relationship to Asymptotic Normality** Recall in the univariate setting, we said that asymptotic normality, as Wasserman defined it for univariate parameters, does not itself imply consistency. However, additional constraints on $se(\hat{\theta})$ *did* imply consistence, particularly if $se(\hat{\theta}) \to 0$. So in the univariate case,

$$\sqrt{n}(\hat{\theta}_n - \theta) \Rightarrow N(0, v(\theta))$$

implies consistency of $\hat{\theta}_n$. This also the case for the multivariate case.

**Theorem.** *If an estimator $\hat{\theta}_n \in R^k$, for $k$ a fixed dimension, is asymptotically normal such that*

$$\sqrt{n}(\hat{\theta}_n - \theta) \Rightarrow N_k(0, V(\theta))$$

*then $\hat{\theta}_n$ is $L_2$ consistent estimator of $\theta$.*

This is the case for the MLE, so the asymptotic normality we asserted above also implies consistency.

## 7.5 Functions of Vector Parameters

**Equivariance** In the section on equivariance, we showed that equivariance holds even if $g(\theta)$ is not 1-1. The entire proof (which we didn't do in detail for the case where it is not 1-1), carries over in full to the multivariate case. In particular, it applies to functions that are not 1-1 such as $g : R^k \to R$, such as $\theta = (\theta_1, \theta_2, \theta_3)$ and $\tau = g(\theta) = \theta_1 + \theta_2$.

Again, recall our definition of a MLE when $g$ is not 1-1: the MLE of $\tau$ is the value that maximizes

$$\mathcal{L}^*(\tau) = \sup_{\theta: g(\theta)=\tau} \mathcal{L}(\theta)$$

So in the case of $g(\theta) = \theta_1 + \theta_2$, this would be the $\tau$ that maximizes,

$$\sup_{\theta: \theta_1+\theta_2=\tau} \mathcal{L}(\theta)$$

So the MLE $\hat{\tau}_n$ might correspond to $g(\theta)$ for multiple $\theta$, not just $\hat{\theta}_n$ (e.g. the value of $\theta_3$ doesn't matter, so all $\theta$ with $\theta_1$ and $\theta_2$ equal to the MLE would give the same $\hat{\tau}_n$).

**Multiparameter delta method** Suppose $\tau = g(\theta_1, \ldots, \theta_k)$ is a differentiable function. Let $\nabla g = (\frac{\partial}{\partial \theta_1} g(\theta) \cdots \frac{\partial}{\partial \theta_k} g(\theta))'$ be the gradient of $g$ and suppose that $\nabla g$ evaluated at $\hat{\theta}_n$ is not zero. Then

$$\frac{\hat{\tau}_n - \tau}{se(\hat{\tau}_n)} \xrightarrow{D} N(0, 1)$$

where

$$se(\hat{\tau}_n) = \sqrt{(\nabla g)' I_n(\theta)^{-1} (\nabla g)}$$

To estimate $se(\hat{\tau}_n)$, we can estimate $I_n(\theta)^{-1}$ with $I_n(\hat{\theta}_n)^{-1}$ or $K_n(\hat{\theta}_n)^{-1}$.

Similarly we estimate $\nabla g$ with $\nabla g$ evaluated at $\hat{\theta}_n$, denoted as $\hat{\nabla} g$ in Wasserman.

**Exercise 21.** Let $\tau = g(\mu, \sigma) = \mu/\sigma$. Find the MLE for $\tau$ and its limiting normal distribution.

# 8 More complicated likelihood problems

**Example: Non i.i.d. Bernoulli (logistic regression)** Consider data on whether a person purchased an item based on their age ($x_1$) and salary ($x_2$). For each observation with $(x_{1i}, x_{2i})$ and we observe a 0-1 success as to whether they purchased the item they saw an advertisement for.

Let $Y_1, \ldots, Y_n$ be $0-1$ observed values, and we can modell the successes as Bernouili with

$$P(Y_i = 1) = p_i(\theta) = \frac{\exp\{\theta_3 + \theta_1 x_{1i} + \theta_2 x_{2i}\}}{1 + \exp\{\theta_3 + \theta_1 x_{1i} + \theta_2 x_{2i}\}}$$

Notice that the probability is defined for any $\theta \in R^3$, so there is no constraint.

Then $\theta = (\theta_1, \theta_2, \theta_3)$ and

$$\begin{aligned}
\ell(\theta) =& \sum_i Y_i \log(p_i(\theta)) + (1 - Y_i)\log(1 - p_i(\theta)) \\
=& \sum_i Y_i \log(\frac{\exp\{\theta_3 + \theta_1 x_{1i} + \theta_2 x_{2i}\}}{1 + \exp\{\theta_3 + \theta_1 x_{1i} + \theta_2 x_{2i}\}}) + (1 - Y_i)\log(\frac{1}{1 + \exp\{\theta_3 + \theta_1 x_{1i} + \theta_2 x_{2i}\}}) \\
=& \sum_i Y_i(\theta_3 + \theta_1 x_{1i} + \theta_2 x_{2i}) - Y_i \log(1 + \exp\{\theta_3 + \theta_1 x_{1i} + \theta_2 x_{2i}\}) \\
& - (1 - Y_i)\log(1 + \exp\{\theta_3 + \theta_1 x_{1i} + \theta_2 x_{2i}\}) \\
=& \sum_i Y_i(\theta_3 + \theta_1 x_{1i} + \theta_2 x_{2i}) - \sum_i \log(1 + \exp\{\theta_3 + \theta_1 x_{1i} + \theta_2 x_{2i}\}) \\
=& \theta_3 n\bar{Y} + \theta_1 \sum_i Y_i x_{1i} + \theta_2 \sum_i Y_i x_{2i} - \sum_i \log(1 + \exp\{\theta_3 + \theta_1 x_{1i} + \theta_2 x_{2i}\})
\end{aligned}$$

Unfortunately, we cannot minimize this equation by hand.

**Numerical Approximation** In many cases, it's not possible to find a closed-form expression for the MLE in multiparameter models. This is true even for some common distributions like the Gamma and Beta distributions.

However, numerical optimization is a highly developed field that comes to our rescue in applied problems (that is, when we have actual values for $X_1, \ldots, X_n$ and need to actually have a value of $\hat{\theta}$).

Most of these algorithms are written for minimization, so we need to

- Write a function $g(\theta) = -\ell_n(\theta)$ for the *negative* log-likelihood
- Minimize $g$ it numerically (i.e. with algorithm that finds candidate $\hat{\theta}_n$). Sometimes these algorithms need to know the gradient or Hessian of $g$, but there are versions that approximate these quantities numerically.
- Examine the behavior of the negative log-likelihood at the candidate minimum, $\hat{\theta}_n$
- Optionally, get a numerical approximation of the Hessian and compute the observed Fisher information matrix in order to estimate the standard error.

  $$-H(\ell(\theta))_{\theta=\hat{\theta}_n}$$

  Numerical optimization programs can often return an approximation of the hessian at the minimization point, so we can easily get this quantity. This is an advantage of using observed rather than expected Fisher's information.

**Back to Example** Since there are no constraints on $\theta$, we can solve this using simple numerical optimization, like the `optim` function in R.

```
dataDir <- "../../Data"
ads <- read.csv(file.path(dataDir, "Social_Network_Ads.csv"))
menAds <- ads[ads$Gender == "Male", ]
negLik <- function(theta) {
    z <- menAds$Age * theta[1] + menAds$EstimatedSalary *
        theta[2] + theta[3]
    p <- exp(z)/(1 + exp(z))
    return(-sum(menAds$Purchased * log(p) + (1 - menAds$Purchased) *
        log(1 - p)))
}
phat <- mean(menAds$Purchased)
optim(par = c(0, 0, log(phat) - log(1 - phat)), fn = negLik)$par
```

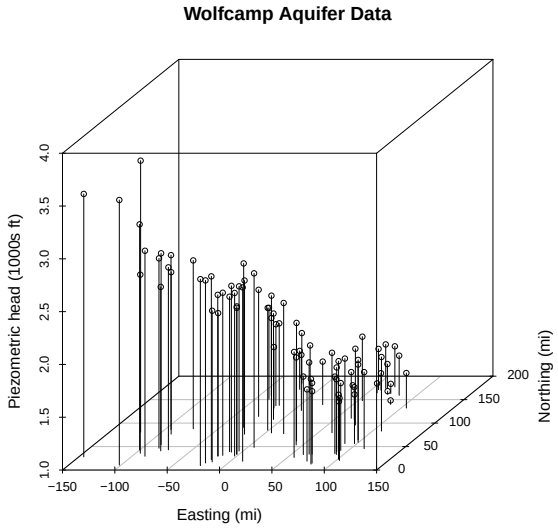
```
## [1]  2.817576e-01  4.416452e-05 -1.477309e+01
```

We can set `Hess=TRUE` in `optim` to get the hessian.

**Example: Non i.i.d data** MLE is also useful when we do not have i.i.d. data, but have a model for the dependency.

The Deaf Smith County (Texas, bordering New Mexico) was selected as an alternate site for a possible nuclear waste disposal repository in the 1980s. This site was

later dropped on grounds of contamination of the aquifer, the source of much of the water supply for west Texas. In a study conducted by the U.S. Department of Energy, piezometric-head data were obtained at 85 locations (irregularly scattered over the Texas panhandle) by drilling a narrow pipe through the aquifer. The data consists of measurements of hydraulic head from an aquifer giving the height of ground water above the location. The question is how to predict the surface. Creating a smooth surface from the measurements would allow us to predict the path of potential contaminants that might leak into the soil.[15]




In this example, we could imagine that there is a plane that describes this surface, and take

$$E[Z_i] = \beta_0 + \beta_1 x_i + \beta_2 y_i$$

where $(x_i, y_i)$ is the location of observation $i$.

The observations $Z_i$ are clearly not independent, so we need a model that accounts for this correlation as well as the randomness (error) in our measurements.

We will let $Z = (Z_1, \ldots, Z_n)$ be our vector of data. They are not i.i.d, so I want a model for their joint density $f(z_1, \ldots, z_n)$. We could assume that they are jointly normal

$$Z \sim N(\mu, \Sigma)$$

where $\mu \in R^n$ and $\Sigma \in R^{n \times n}$.

We would want these values to have special structure (otherwise we are fitting $n + n(n-1)/2$ parameters with only $n$ data points). We would want the mean to be

[15] Noel Cressie (1989) Geostatistics, The American Statistician, 43:4, 197-202.

related to the $x, y$ coordinates as described above:

$$\mu_i = E[Z_i] = \beta_0 + \beta_1 x_i + \beta_2 y_i$$

And since our data are not independent, $\Sigma$ should not be diagonal. One model would be to have correlation between observations decay with their distance, such as

$$\Sigma_{ij} = Cov(Z_i, Z_j) = \sigma^2 \exp\{-d_{ij}/\rho\}$$

where $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$.

This gives us a parameter vector

$$\theta = (\beta_0, \beta_1, \beta_2, \sigma^2, \rho).$$

Our likelihood would be

$$\ell_n(\theta) = \log\left((2\pi)^{-k/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu)\right\}\right)$$

where $\mu$ and $\Sigma$ are functions of our parameters $\theta$.

Again, there is no closed form expression for the MLE of $\theta$ and we would need to find it via numerical optimization.

# 9 Sufficiency

**Motivation** We hope to separate the information contained in the data into the information relevant for making inference about $\theta$ and the information irrelevant for these inferences. In other words, we would like to compress the data to, e.g. $T(X)$, without loss of information. (Actually, it often turns out that some part of the data carries no information about the unknown distribution that produces the data)

**Benefits:**

1. increasing computational efficiency and decreasing storage requirements
2. involving irrelevant information may increase an estimator's risk (see Rao-Blackwell Theorem)
3. Improving the scientific interpretability of our data

**Definition** (Sufficient Statistic[16])**.** Suppose $X = X_1, \ldots, X_n$ has a distribution from $\mathcal{P} = \{ P_\theta : \theta \in \Omega \}$. A statistic $T$ is **sufficient** for $\theta$ if, for every $t$ in the range $\mathcal{T}$ of $T$, the conditional distribution $P_\theta(X \mid T(X))$ is independent of $\theta$.

In other words, once you know the value of $T(X)$, there is no further information regarding $\theta$ left in the data.

**Example (Wasserman)** : Let $X_i \sim Ber(\theta)$ i.i.d., $i = 1, ..., n$. Show that $T(X) = \sum_{i=1}^n X_i$ is sufficient for $\theta$.

$$P_\theta(X_1 = x_1, \ldots, X_n = x_n | T(X) = t) = \begin{cases} \frac{\theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} & \sum x_i = t \\ 0 & otherwise \end{cases}$$
$$= \begin{cases} \frac{1}{\binom{n}{t}} & \sum x_i = t \\ 0 & otherwise \end{cases}$$

Notice that the joint distribution of $X$ involves our sufficient statistic $T(X)$:

$$P_\theta(X_1 = x_1, \ldots, X_n = x_n) = \theta^{T(X)}(1-\theta)^{n-T(X)}$$

This is a general phenomena,

**Theorem 6** (Neyman Factorization Theorem.)**.** *Suppose a family $\{ P_\theta : \theta \in \Omega \}$ of distributions have joint mass functions or densities $\{ p(x;\theta) : \theta \in \Omega \}$. Then a statistic $T$ is sufficient for $\theta$ if and only if there are functions $h$ and $g$ such that the density/mass function can be written*

$$p(x;\theta) = h(x) \cdot g(T(x), \theta).$$

Note that this clearly means that any 1-1 function applied to a statistic is also sufficient, i.e. $f(T(x))$ is also sufficient.

*Proof.* We will do the case of a discrete distribution

$\Rightarrow$: **Suppose $T(X)$ is a sufficient statistic**
Let $g(t|\theta) = P_\theta(T(X) = t)$ and $h(x) = P(X = x|T(X) = T(x))$. $h(x)$ doesn't depend

[16]This is different from that given in Wasserman, which is not usual

on $\theta$ because $T(X)$ is sufficent. Then basic conditioning rules give the result:

$$\begin{aligned} p(x;\theta) &= P_\theta(X = x) \\ &= P(X = x|T(X) = T(x))P_\theta(T(X) = T(x)) \\ &= h(x) \cdot g(T(x), \theta). \end{aligned}$$

$\Leftarrow$: **Suppose** $p(x;\theta) = h(x)g(T(x),\theta)$
Let $q(t;\theta)$ be the probability mass function for $T(X)$. Note that

$$\begin{aligned} q(t;\theta) &= P_\theta(T(X) = t) \\ &= \sum_{x:T(x)=t} P(T(X) = t|x)P(x) \\ &= \sum_{x:T(x)=t} \underbrace{P(T(X) = t|x)}_{=1} h(x)g(T(x),\theta) \\ &= \sum_{x:T(x)=t} h(x)g(T(x),\theta) \end{aligned}$$

Then

$$\begin{aligned} P_\theta(X = x|T(X) = t) &= \frac{p(x;\theta)}{q(t;\theta)} \\ &= \frac{h(x)g(t,\theta)}{q(t;\theta)} \\ &= \frac{h(x)g(t,\theta)}{\sum_{y:T(y)=t} h(y)g(T(y),\theta)} \\ &= \frac{h(x)g(t,\theta)}{g(t,\theta)} \sum_{y:T(y)=t} h(y) \\ &= \frac{h(x)}{\sum_{y:T(y)=t} h(y)} \end{aligned}$$

This is a constant value that doesn't depend on $\theta$. □

**Exercise 22.** Let $Y_i \sim Uniform(0,\theta)$ i.i.d., $i = 1, ..., n$. Show that $T = Y_{(n)}$ is sufficient for $\theta$.

**Exercise 23.** Let $X_i \sim N(\theta, 1)$ i.i.d., $i = 1, ..., n$. Show that $T = \sum_{i=1}^n X_i$ is sufficient for $\theta$.

## 9.1 The Rao-Blackwell Theorem

Estimators that use information not contained in a sufficient statistic will not do as well. Specifically, an estimator $\hat{\theta}_n$ should only depend on a sufficient statistic, and otherwise you can find a statistic that improves the MSE of $\hat{\theta}_n$.

**Theorem 7** (Rao-Blackwell)**.** *Let $\hat{\theta}$ be an estimator and $T$ a sufficient statistic for $\theta$. Define a new estimator as*

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

*Then*

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

*and the equality is strict unless $\tilde{\theta} = \hat{\theta}$.*[17]

**Proof of Rao-Blackwell:** by Jensen's inequality and iterated expectation.

- Note that if $\hat{\theta}$ is already only a function of the sufficient statistic $T$, then $E(\hat{\theta}|T) = \hat{\theta}$, in other words there was no new information.
- Otherwise, $\tilde{\theta} = E_\theta(\hat{\theta}|T)$ is a new statistic *that depends only on the random variable $T$ (and critically not on $\theta$)* – you've removed whatever randomness in the problem that isn't due to $T$.
- Since $T$ is sufficient, and $\tilde{\theta}$ is a function of only $T$, then the distribution of $X|\tilde{\theta}$, is also independent of $\theta$, so $\tilde{\theta}$ is sufficient.
- This theorem is constructive, in the sense that it also gives a recipe (in principle) for improving upon an estimator that is not sufficient.

**Conditioning on statistics that aren't sufficient** Why limit ourselves to conditioning on sufficient statistics? Consider two observations $X_1, X_2$, i.i.d. $N(\theta, 1)$. Then $\bar{X} = \frac{1}{2}(X_1 + X_2)$ has

$$E_\theta \bar{X} = \theta, var_\theta(\bar{X}) = \frac{1}{2}$$

We could consider the statistic $S(X) = X_1$, which isn't sufficient,

$$\begin{aligned}\tilde{\theta} &= E_\theta(\bar{X}|X_1) \\ &= E_\theta(X_1|X_1) + E_\theta(X_2|X_1) \\ &= \frac{1}{2}X_1 + \frac{1}{2}\theta\end{aligned}$$

Then

$$E_\theta\tilde{\theta} = \theta, var_\theta\tilde{\theta} = \frac{1}{4}$$

so we've reduced the MSE/variance, but $\tilde{\theta}$ is not a valid estimator (it depends on the unknown $\theta$).

[17] In fact we can generalize this to any measure of error (or "loss"), not just MSE if the loss function is convex. We'll see more about loss functions when we talk about decision theory

## 9.2 Minimal Sufficiency

Notice that there is not a unique sufficient statistic. For example, the full data

$$T(X) = (X_1, \ldots, X_n)$$

is a statistic and it is of course sufficient. Similarly, if the data is i.i.d, then the data ordered is also sufficient,

$$T(X) = (X_{(1)}, \ldots, X_{(n)})$$

**Example** If $X_1, \ldots, X_n \overset{i.i.d}{\sim} N(\mu, 1)$ then both $T(X) = \bar{X}_n$ is sufficient and

$$T(X) = (\sum_{i=1}^{\lfloor n/2 \rfloor} X_i, \sum_{i=\lceil n/2 \rceil}^{n} X_i)$$

are sufficient.

We have, however, increasing amounts of reduction for some sufficient statistics, which makes us want to identify a statistic that has the greatest reduction.

**Definition** (Minimal Sufficiency)**.** Suppose $T(X)$ is sufficient for $P = \{P_\theta : \theta \in \Omega\}$. For any other sufficient statistic $S(X)$, if we can always find a function $f$ such that $T = f(S)$, then $T$ is minimally sufficient.

Note that this means that a minimally sufficient statistic if a function of *every other* sufficient statistic.

**Exercise 24.** Let $X_1, ..., X_n$ be iid and follow a normal distribution $N(\mu, \sigma^2)$. Find the minimal sufficient statistic for $\mu$ and $\sigma^2$.

## 9.3 MLE and Sufficiency

Sufficiency is a general property that an estimator can have, and is not directly related to our discussion of MLE. But it raises the natural question are MLEs sufficient?

**Theorem 8.** *If $T(X)$ is sufficient for $\theta$ and a unique MLE $\hat{\theta}_n$ of $\theta$ exists, then $\hat{\theta}_n$ must be a function of $T(X)$. If any MLE exists, an MLE $\hat{\theta}_n$ can be chosen to be a function of $T(X)$.*

*Proof.* We will concentrate on the case of a unique MLE. =From the factorization theorem of a sufficient statistic, the density function can be written as

$$p(x; \theta) = g(T(x), \theta)h(x)$$

so that

$$\ell(y;\theta) = \log g(T(x),\theta) + \log h(x).$$

Hence maximizing $\ell(y;\theta)$ with respect to $\theta$ is equivalent to maximizing $\log g(T(X),\theta)$ with respect to $\theta$. Therefore, the MLE depends on $X$ only through $T(X)$. □

Note that this means that the MLE (if unique) is a function of *any* sufficient statistic. However, a function of a sufficient statistic does not itself have to be sufficient (unless the function is 1-1).

**Theorem 9.** *If the MLE is itself sufficient, it is minimal sufficient.*