Parametric estimation: some methods and examples

Avanti Athreya

SAMSI/Duke University
Undergraduate Workshop, May 2010

May 18, 2010
Unknown parameters

In many real-life problems, important parameters are unknown.

Note: in this context, unknown does NOT mean random!

Example: what is the average height of an American woman?

Partial answer: It’s some number, say $\mu$, which we can only determine by measuring every American woman and then taking an average.

Not reasonable for most us (aside from US Census Bureau) to gather this data!
Unknown parameters

In many real-life problems, important parameters are unknown. Note: in this context, unknown does NOT mean random!

Example: what is the average height of an American woman?
Partial answer: It’s some number, say $\mu$, which we can only determine by measuring every American woman and then taking an average.
Not reasonable for most us (aside from US Census Bureau) to gather this data!
Unknown parameters

In many real-life problems, important parameters are *unknown*. Note: in this context, *unknown* does NOT mean *random*!

Example: what is the average height of an American woman?

Partial answer: It’s some number, say \( \mu \), which we can only determine by measuring *every American woman* and then taking an average.

Not reasonable for most us (aside from US Census Bureau) to gather this data!
In many real-life problems, important parameters are unknown. Note: in this context, unknown does NOT mean random!

Example: what is the average height of an American woman?

Partial answer: It’s some number, say \( \mu \), which we can only determine by measuring every American woman and then taking an average.

Not reasonable for most us (aside from US Census Bureau) to gather this data!
Unknown parameters

In many real-life problems, important parameters are \textit{unknown}. Note: in this context, \textit{unknown} does NOT mean \textit{random}!

Example: what is the average height of an American woman?

Partial answer: It’s some number, say $\mu$, which we can only determine by measuring every \textit{American} woman and then taking an average.

Not reasonable for most us (aside from US Census Bureau) to gather this data!
Instead of determining the true population mean $\mu$ exactly, we approximate it.

Idea: take a random sample of, say, 100 women, find the average sample height, and use this as a good guess—more precisely, an estimate of the true value $\mu$.

To do this, we assume:

1. We have a random sample, i.e. a collection of i.i.d. random variables $X_i$ whose distributions have a known functional form but with unknown parameters.

2. We construct an estimator, i.e. a function of our random sample, which is “close” to the parameter we don’t know.
Estimating unknown parameters

Instead of determining the *true population* mean $\mu$ exactly, we approximate it.

Idea: take a random sample of, say, 100 women, find the average sample height, and use this as a good guess—more precisely, an *estimate* of the true value of $\mu$.

To do this, we assume:

1. We have a *random sample*, i.e. a collection of i.i.d. random variables $X_i$ whose *distributions have a known functional form* but with *unknown* parameters.

2. We construct an *estimator*, i.e. a function of our random sample, which is “close” to the parameter we don’t know.
Suppose we take a random sample of women and measure their heights.

Some assumptions:

1. Each woman’s height is a normal random variable $X_i$ with mean $\mu$ and variance $\sigma$;
2. The variables $X_i$ are independent.

The strong law of large numbers says that with probability one, as $n \to \infty$,

$$\hat{\mu} = \frac{\sum_{i}^{n} X_i}{n} \to \mu$$

So $\hat{\mu} = \frac{\sum_{i}^{n} X_i}{n}$ is a good estimator for $\mu$. 
Estimating the sample mean

Suppose we take a random sample of women and measure their heights.

Some assumptions:

1. Each woman’s height is a normal random variable $X_i$ with mean $\mu$ and variance $\sigma$;
2. The variables $X_i$ are independent.

The strong law of large numbers says that with probability one, as $n \to \infty$,

$$\hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n} \to \mu$$

So $\hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n}$ is a good estimator for $\mu$. 
Estimating the sample mean

Suppose we take a random sample of women and measure their heights.

Some assumptions:

1. Each woman’s height is a normal random variable $X_i$ with mean $\mu$ and variance $\sigma$;
2. The variables $X_i$ are independent.

The strong law of large numbers says that with probability one, as $n \to \infty$,

$$\hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n} \to \mu$$

So $\hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n}$ is a good estimator for $\mu$. 
Estimators are random variables! They are functions of the sample (i.e. the data).

The sample mean $\hat{\mu}$, for instance, is random.

One can describe:

1. The distribution function of an estimator;
2. Its mean, variance, and other moments;
3. Its asymptotic behavior as the sample size $n$ grows large.
Estimators are random variables! They are functions of the sample (i.e. the data).

The sample mean $\hat{\mu}$, for instance, is random.

One can describe:

1. The distribution function of an estimator;
2. Its mean, variance, and other moments;
3. Its asymptotic behavior as the sample size $n$ grows large.
Desirable properties of estimators

1. **Unbiasedness**: \( \hat{\mu} \) is unbiased for \( \mu \) if \( E[\hat{\mu}] = \mu \).

2. **Minimal variance**: \( V(\hat{\mu}) \) should be small.

3. **Consistency**: \( \hat{\mu}_n \) is consistent if for any \( \epsilon > 0 \),

\[
\lim_{n \to \infty} P\{|\hat{\mu}_n - \mu| > \epsilon\} = 0
\]

4. **Asymptotic normality**: \( \hat{\mu}_n \) is asymptotically normal if

\[
\sqrt{n}(\hat{\mu}_n - \mu) \to_d N(0, \sigma^2)
\]
Properties of estimators: unbiasedness

Suppose \( \hat{\mu} \) is an estimator for \( \mu \).

We say \( \hat{\mu} \) is unbiased for \( \mu \) if \( E[\hat{\mu}] = \mu \).

That is, on average, the estimator will equal the true value of the parameter.
Suppose $\hat{\mu}$ is an estimator for $\mu$.

We say $\hat{\mu}$ is *unbiased for* $\mu$ if $E[\hat{\mu}] = \mu$.

That is, on average, the estimator will equal the true value of the parameter.
Suppose $X_1, \cdots, X_n$ is a random sample from a distribution with mean $\mu$ and variance $\sigma^2$.

The sample mean $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ is an unbiased estimator for $\mu$:

$$E(\bar{X}) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{\sum_{i=1}^{n} E(X_i)}{n} = \frac{n\mu}{n} = \mu$$

The sample variance

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

is an unbiased estimator for the true variance $\sigma^2$. 

Examples of unbiased estimators: sample mean and sample variance
Examples of unbiased estimators: sample mean and sample variance

Suppose $X_1, \cdots, X_n$ is a random sample from a distribution with mean $\mu$ and variance $\sigma^2$.

The sample mean $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ is an unbiased estimator for $\mu$:

$$E(\bar{X}) = E \left( \frac{\sum_{i=1}^{n} X_i}{n} \right) = \frac{\sum_{i=1}^{n} E(X_i)}{n} = \frac{n \mu}{n} = \mu$$

The sample variance

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1}$$

is an unbiased estimator for the true variance $\sigma^2$. 
More on the sample variance

Why is the sample variance unbiased?

\[
S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} = \frac{1}{n-1} \left[ \sum_{i=1}^{n} X_i^2 - \left( \frac{\sum_{i=1}^{n} X_i}{n} \right)^2 \right]
\]

Taking expectations, we get that \(E(S^2)\) is given by

\[
\frac{1}{n-1} \left\{ \sum_{i=1}^{n} E(X_i^2) - \frac{1}{n} E \left[ \left( \sum_{i=1}^{n} X_i \right)^2 \right] \right\}
= \frac{1}{n-1} \left\{ \sum_{i} (\sigma^2 + \mu^2) - \frac{1}{n} \left[ V(\sum_{i} X_i) + \left( E \left( \sum_{i} X_i \right) \right)^2 \right] \right\}
= \frac{1}{n-1} \left\{ n\sigma^2 + n\mu^2 - \frac{1}{n} n\sigma^2 - \frac{1}{n} n^2 \mu^2 \right\} = \sigma^2
\]
More on the sample variance

Why is the sample variance unbiased?

\[ S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1} = \frac{1}{n - 1} \left[ \sum_{i=1}^{n} X_i^2 - \left( \frac{\sum_{i=1}^{n} X_i}{n} \right)^2 \right] \]

Taking expectations, we get that \( E(S^2) \) is given by

\[
\frac{1}{n - 1} \left\{ \sum_{i=1}^{n} E(X_i^2) - \frac{1}{n} E \left[ \left( \sum_{i=1}^{n} X_i \right)^2 \right] \right\} \\
= \frac{1}{n - 1} \left\{ \sum_{i} (\sigma^2 + \mu^2) - \frac{1}{n} \left[ V(\sum_{i} X_i) + \left( E \left( \sum_{i} X_i \right) \right)^2 \right] \right\} \\
= \frac{1}{n - 1} \left\{ n\sigma^2 + n\mu^2 - \frac{1}{n} n\sigma^2 - \frac{1}{n} n^2 \mu^2 \right\} = \sigma^2
Why is the sample variance unbiased?

\[ S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} = \frac{1}{n-1} \left[ \sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n} \right] \]

Taking expectations, we get that \( E(S^2) \) is given by

\[
\frac{1}{n-1} \left\{ \sum_{i=1}^{n} E(X_i^2) - \frac{1}{n} E \left[ \left( \sum_{i=1}^{n} X_i \right)^2 \right] \right\}
\]

\[
= \frac{1}{n-1} \left\{ \sum_{i} (\sigma^2 + \mu^2) - \frac{1}{n} \left[ V(\sum_{i} X_i) + \left( E \left( \sum_{i} X_i \right) \right)^2 \right] \right\}
\]

\[
= \frac{1}{n-1} \left\{ n\sigma^2 + n\mu^2 - \frac{1}{n} \frac{n \sigma^2}{n} - \frac{1}{n} \frac{n \mu^2}{n^2} \right\} = \sigma^2
\]
More on the sample variance

Why is the sample variance unbiased?

\[ S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1} = \frac{1}{n - 1} \left[ \sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n} \right] \]

Taking expectations, we get that \( E(S^2) \) is given by

\[
\frac{1}{n - 1} \left\{ \sum_{i=1}^{n} E(X_i^2) - \frac{1}{n} E \left[ \left( \sum_{i=1}^{n} X_i \right)^2 \right] \right\} \\
= \frac{1}{n - 1} \left\{ \sum_{i} (\sigma^2 + \mu^2) - \frac{1}{n} \left[ V(\sum_i X_i) + \left( E \left( \sum_i X_i \right) \right)^2 \right] \right\} \\
= \frac{1}{n - 1} \left\{ n\sigma^2 + n\mu^2 - \frac{1}{n} n\sigma^2 - \frac{1}{n} n^2 \mu^2 \right\} = \sigma^2
\]
Minimum variance unbiased estimators

General guideline: if there are multiple unbiased estimators for a parameter $\mu$, choose the one with minimum variance.
Recall that $\hat{\mu}_n$ is consistent if for any $\epsilon > 0$,

$$
\lim_{n \to \infty} P\{|\hat{\mu}_n - \mu| > \epsilon\} = 0
$$

Idea: for large samples, the probability is small that $\hat{\mu}$ differs from $\mu$ by very much.

Exercise: Suppose $X_1, \cdots, X_n$ is a random sample from a distribution with mean $\mu$ and finite variance $\sigma^2$. How can we show that $\bar{X}$ is consistent for $\mu$?
Recall that $\hat{\mu}_n$ is consistent if for any $\epsilon > 0$,

$$\lim_{n \to \infty} P\{|\hat{\mu}_n - \mu| > \epsilon\} = 0$$

Idea: for large samples, the probability is small that $\hat{\mu}$ differs from $\mu$ by very much.

Exercise: Suppose $X_1, \cdots, X_n$ is a random sample from a distribution with mean $\mu$ and finite variance $\sigma^2$. How can we show that $\bar{X}$ is consistent for $\mu$?
Recall that $\hat{\mu}_n$ is asymptotically normal if

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

This is useful because it gives us a limiting distribution for the estimator—we can approximate the probability that $\hat{\mu}$ is far from $\mu$.

What’s an example of an asymptotically normal estimator for the mean $\mu$ of a random sample from a distribution with finite variance?
Asymptotic normality

Recall that $\hat{\mu}_n$ is asymptotically normal if

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

This is useful because it gives us a limiting distribution for the estimator—we can approximate the probability that $\hat{\mu}$ is far from $\mu$.

What’s an example of an asymptotically normal estimator for the mean $\mu$ of a random sample from a distribution with finite variance?
Asymptotic normality

Recall that \( \hat{\mu}_n \) is asymptotically normal if

\[
\sqrt{n}(\hat{\mu}_n - \mu) \to_d N(0, \sigma^2)
\]

This is useful because it gives us a limiting distribution for the estimator—we can approximate the probability that \( \hat{\mu} \) is far from \( \mu \).

What’s an example of an asymptotically normal estimator for the mean \( \mu \) of a random sample from a distribution with finite variance?
We will focus on two important methods for constructing estimators:

1. the *method of moments*, and
2. the method of *maximum likelihood*.
Key assumptions

Throughout, we will assume:
Our data consists of i.i.d random variables $X_i$ with distribution $F$ where

1. the *functional form* of $F$ is known (e.g. exponential, Gaussian, binomial)
2. certain *parameter values* that determine $F$ are not known (e.g. $\lambda, \mu, p$)

We take the functional form of $F$ to be known and want to estimate a real number or numbers.
Key assumptions

Throughout, we will assume:
Our data consists of i.i.d random variables $X_i$ with distribution $F$ where

1. the *functional form* of $F$ is known (e.g. exponential, Gaussian, binomial)
2. certain *parameter values* that determine $F$ are not known (e.g. $\lambda$, $\mu$, $p$)

We take the functional form of $F$ to be known and want to estimate a real number or numbers.
Method of moments

Suppose the distribution function $F$ depends on $k$ parameters $\mu_1, \cdots, \mu_k$.

Compute the *true moments* and equate them to the *sample moments*, and use the system to solve for the unknown parameters.

\[
E[X] = \int xf(x; \mu_1, \cdots, \mu_k) \, dx = \frac{\sum_1^n X_i}{n}
\]

\[
E[X^2] = \int x^2 f(x; \mu_1, \cdots, \mu_k) \, dx = \frac{\sum_1^n X_i^2}{n}
\]

\[
E[X^k] = \int x^k f(x; \mu_1, \cdots, \mu_k) \, dx = \frac{\sum_1^n X_i^k}{n}
\]
Specific examples: Exponential and Gamma

1. Use the method of moments to estimate $\lambda$ when $X_i \sim \exp(\lambda)$. Does this produce an unbiased estimator?

2. Use the method of moments to estimate $\alpha, \beta$ when $X_i$ follow a Gamma distribution with parameters $\alpha, \beta$.

Recall that the density function for a Gamma distribution with parameters $\alpha, \beta$ is given by

$$f(t) = \frac{1}{\beta^\alpha \Gamma(\alpha)} t^{\alpha-1} e^{-\frac{t}{\beta}}$$

Also, $E(X) = \alpha \beta$ and $V(X) = \alpha \beta^2$. 
Specific examples: Exponential and Gamma

1. Use the method of moments to estimate $\lambda$ when $X_i \sim \exp(\lambda)$. Does this produce an unbiased estimator?

2. Use the method of moments to estimate $\alpha, \beta$ when $X_i$ follow a Gamma distribution with parameters $\alpha, \beta$. Recall that the density function for a Gamma distribution with parameters $\alpha, \beta$ is given by

$$f(t) = \frac{1}{\beta^\alpha \Gamma(\alpha)} t^{\alpha-1} e^{-\frac{t}{\beta}}$$

Also, $E(X) = \alpha \beta$ and $V(X) = \alpha \beta^2$. 
Using MOM to estimate negative binomial parameters

Recall that $X$ is a *negative binomial random variable* with parameters $r$ and $p$ if $X$ is a count of the number of failures before the $r$th success in a sequence of independent, identical trials, each of which has success probability $p$.

What is the pmf of $X$?

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

We can compute the mean and variance of $X$:

$$E(X) = \frac{r(1 - p)}{p}, \quad V(X) = \frac{r(1 - p)}{p^2}$$
Using MOM to estimate negative binomial parameters

Recall that $X$ is a negative binomial random variable with parameters $r$ and $p$ if $X$ is a count of the number of failures before the $r$th success in a sequence of independent, identical trials, each of which has success probability $p$.

What is the pmf of $X$?

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

We can compute the mean and variance of $X$:

$$E(X) = \frac{r(1 - p)}{p}, \quad V(X) = \frac{r(1 - p)}{p^2}$$
A computational example in R

Find the method of moments estimators for $r$ and $p$.

Do you notice any potential problems with these estimators?

Can the number of goals per game in soccer be modeled effectively by a negative binomial distribution?

Consider the following frequency data for the number of goals per game, which you can import into R and then use to calculate the estimators $\hat{r}$ and $\hat{p}$ from above:

<table>
<thead>
<tr>
<th>Goals</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>29</td>
<td>71</td>
<td>82</td>
<td>89</td>
<td>65</td>
<td>45</td>
<td>24</td>
<td>7</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>
A computational example in R

Find the method of moments estimators for \( r \) and \( p \).

Do you notice any potential problems with these estimators?

Can the number of goals per game in soccer be modeled effectively by a negative binomial distribution?

Consider the following frequency data for the number of goals per game, which you can import into R and then use to calculate the estimators \( \hat{r} \) and \( \hat{p} \) from above:

<table>
<thead>
<tr>
<th>Goals</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>29</td>
<td>71</td>
<td>82</td>
<td>89</td>
<td>65</td>
<td>45</td>
<td>24</td>
<td>7</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>
Find the method of moments estimators for $r$ and $p$.

Do you notice any potential problems with these estimators?

Can the number of goals per game in soccer be modeled effectively by a negative binomial distribution?

Consider the following frequency data for the number of goals per game, which you can import into R and then use to calculate the estimators $\hat{r}$ and $\hat{p}$ from above:

<table>
<thead>
<tr>
<th>Goals</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>29</td>
<td>71</td>
<td>82</td>
<td>89</td>
<td>65</td>
<td>45</td>
<td>24</td>
<td>7</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>
The maximum likelihood estimator is constructed as follows.

Let \( \mu = (\mu_1, \cdots, \mu_n) \) be the vector of unknown parameters, and \( \vec{x} = (x_1, \cdots, x_n) \) be the vector of data points.

Define the joint likelihood function

\[
L_{\vec{x}}(\mu) = f(x_1, \cdots, x_n; \mu_1, \cdots, \mu_k)
\]

as

\[
L_{\vec{x}}(\mu) = f(x_1, \cdots, x_n; \mu_1, \cdots, \mu_k) = f(x_1; \mu_1, \cdots, \mu_k) \ast \cdots \ast f(x_n; \mu_1, \cdots, \mu_k)
\]
Maximum likelihood estimation

The *maximum likelihood estimator* is constructed as follows.

Let \( \mu = (\mu_1, \cdots, \mu_n) \) be the vector of unknown parameters, and \( \vec{x} = (x_1, \cdots, x_n) \) be the vector of data points.

Define the joint likelihood function
\[
L_{\vec{x}}(\mu) = f(x_1, \cdots, x_n; \mu_1, \cdots, \mu_k)
\]
as
\[
L_{\vec{x}}(\mu) = f(x_1, \cdots, x_n; \mu_1, \cdots, \mu_k)
= f(x_1; \mu_1, \cdots, \mu_k) \ast \cdots \ast f(x_n; \mu_1, \cdots, \mu_k)
\]
The joint likelihood function is exactly the probability density function for the collection $X_1, \cdots, X_n$ if the parameters are $\mu_1, \cdots, \mu_k$.

The joint density function and the likelihood function are the same function of the pair $(\bar{x}, \mu)$.

However, the joint density function $f_{\mu}(x_1, \cdots, x_n)$ is considered a function of $\bar{x}$ for fixed $\mu$,

and the likelihood function $L_{\bar{x}}(\mu)$ is function of $\mu$ for fixed $\bar{x}$. 
Likelihood function vs. joint density function

The joint likelihood function is exactly the probability density function for the collection $X_1, \cdots, X_n$ if the parameters are $\mu_1, \cdots, \mu_k$.

The joint density function and the likelihood function are the same function of the pair $(\bar{x}, \mu)$.

However, the joint density function $f_{\mu}(x_1, \cdots, x_n)$ is considered a function of $\bar{x}$ for fixed $\mu$,

and the likelihood function $L_{\bar{x}}(\mu)$ is function of $\mu$ for fixed $\bar{x}$. 
Likelihood function vs. joint density function

The joint likelihood function is exactly the probability density function for the collection $X_1, \cdots, X_n$ if the parameters are $\mu_1, \cdots, \mu_k$.

The joint density function and the likelihood function are the same function of the pair $(\bar{x}, \mu)$.

However, the joint density function $f_\mu(x_1, \cdots, x_n)$ is considered a function of $\bar{x}$ for fixed $\mu$,

and the likelihood function $L_{\bar{x}}(\mu)$ is function of $\mu$ for fixed $\bar{x}$. 
Likelihood function vs. joint density function

The joint likelihood function is exactly the probability density function for the collection $X_1, \cdots, X_n$ if the parameters are $\mu_1, \cdots, \mu_k$.

The joint density function and the likelihood function are the same function of the pair $(\bar{x}, \mu)$.

However, the joint density function $f_{\mu}(x_1, \cdots, x_n)$ is considered a function of $\bar{x}$ for fixed $\mu$,

and the likelihood function $L_{\bar{x}}(\mu)$ is function of $\mu$ for fixed $\bar{x}$. 
What do we mean by “maximum likelihood”?

Recall that

\[ P\{X_1 = x_1 \pm h, \ldots, X_n = x_n \pm h\} \approx f(x_1; \mu_1, \ldots, \mu_k) \ast \cdots \ast f(x_n; \mu_1, \ldots, \mu_k) \ast 2h^n \]

Given that we see the data \( X_1 = x_1, \ldots, X_n = x_n \), what values of the parameters \( \mu_1, \ldots, \mu_n \) would make observing this data most likely?

We want precisely those values of the parameters that maximize the likelihood function!
What do we mean by “maximum likelihood”?

Recall that

\[ P\{X_1 = x_1 \pm h, \ldots, X_n = x_n \pm h\} \approx f(x_1; \mu_1, \ldots, \mu_k) \ast \cdots \ast f(x_n; \mu_1, \ldots, \mu_k) \ast 2h^n \]

Given that we see the data \( X_1 = x_1, \ldots X_n = x_n \), what values of the parameters \( \mu_1, \ldots, \mu_n \) would make observing this data most likely?

We want precisely those values of the parameters that maximize the likelihood function!
The maximum likelihood estimators

Given the joint likelihood function $f(x_1, \cdots, x_n; \mu_1, \cdots \mu_k)$ the maximum likelihood estimators $\hat{\mu}_1, \cdots \hat{\mu}_k$ satisfy

$$L_\mathcal{X}(\mu) = L_\mathcal{X}(\hat{\mu}_1, \cdots, \hat{\mu}_n) = f(x_1, x_2, \cdots, x_n; \hat{\mu}_1, \cdots, \hat{\mu}_k) \geq f(x_1, \cdots, x_n; \mu_1, \cdots, \mu_k)$$

for all $\mu_1, \cdots, \mu_k$. 
Computing the maximum likelihood estimators

To compute the MLE, we need:

1. to determine the likelihood function;
2. determine how to maximize it as a function of the parameters;
3. solve for the maximizing values of the parameters.
The log-likelihood

Since the log function is monotone increasing, maximizing \( \log(L_{\vec{x}}(\mu)) \) is equivalent to maximizing \( L_{\vec{x}}(\mu) \).

Note that

\[
L_{\vec{x}}(\mu) = f(x_1; \mu)f(x_2; \mu) \cdots f(x_n; \mu)
\]

Thus

\[
\log(L_{\vec{x}}(\mu)) = \sum_{1}^{n} \log(f(x_i; \mu))
\]

and it’s often much easier to deal with sums than products!
Since the log function is monotone increasing, maximizing \( \log(L_{\vec{x}}(\mu)) \) is equivalent to maximizing \( L_{\vec{x}}(\mu) \).

Note that

\[
L_{\vec{x}}(\mu) = f(x_1; \mu) f(x_2; \mu) \cdots f(x_n; \mu)
\]

Thus

\[
\log(L_{\vec{x}}(\mu)) = \sum_{1}^{n} \log(f(x_i; \mu))
\]

and it’s often much easier to deal with sums than products!
Remember: the likelihood and log-likelihood functions depend on all the data \((\vec{x} = (x_1, \cdots, x_n))\) and all the parameters \(\mu = (\mu_1, \cdots, \mu_n)\), but because we don’t know the parameters and have only the data before us, we consider it to be a function of the parameters.

Remember: We want to find the parameter values that will make the data we see the “most likely.”

\textit{THEREFORE}, we want to maximize \(L_{\vec{x}}(\mu)\) as a function of \(\mu\), and the resulting solutions will depend on the data \(\vec{x}\).

\textit{The estimators} \(\hat{\mu}_1, \cdots, \hat{\mu}_n\) \textit{had better depend on the data} \(\vec{x} = (x_1, \cdots, x_n)!\)
Finding maxima

Remember: the likelihood and log-likelihood functions depend on all the data ($\vec{x} = (x_1, \ldots, x_n)$) and all the parameters $\mu = (\mu_1, \ldots, \mu_n)$,

but because we don’t know the parameters and have only the data before us, we consider it to be a function of the parameters.

Remember: We want to find the parameter values that will make the data we see the “most likely.”

THEREFORE, we want to maximize $L_{\vec{x}}(\mu)$ as a function of $\mu$, and the resulting solutions will depend on the data $\vec{x}$.

The estimators $\hat{\mu}_1, \ldots, \hat{\mu}_n$ had better depend on the data $\vec{x} = (x_1, \ldots, x_n)$!
Finding maxima

Remember: the likelihood and log-likelihood functions depend on all the data ($\vec{x} = (x_1, \cdots, x_n)$) and all the parameters $\mu = (\mu_1, \cdots, \mu_n)$, but because we don’t know the parameters and have only the data before us, we consider it to be a function of the parameters.

Remember: We want to find the parameter values that will make the data we see the “most likely.”

Therefore, we want to maximize $L_{\vec{x}}(\mu)$ as a function of $\mu$, and the resulting solutions will depend on the data $\vec{x}$.

The estimators $\hat{\mu}_1, \cdots, \hat{\mu}_n$ had better depend on the data $\vec{x} = (x_1, \cdots, x_n)$!
Finding maxima

Remember: the likelihood and log-likelihood functions depend on all the data \((\vec{x} = (x_1, \cdots, x_n))\) and all the parameters \(\mu = (\mu_1, \cdots, \mu_n)\),

but because we don’t know the parameters and have only the data before us, we consider it to be a function of the parameters.

Remember: We want to find the parameter values that will make the data we see the “most likely.”

**THEREFORE**, we want to maximize \(L_{\vec{x}}(\mu)\) as a function of \(\mu\), and the resulting solutions will depend on the data \(\vec{x}\).

The estimators \(\hat{\mu}_1, \cdots, \hat{\mu}_n\) had better depend on the data \(\vec{x} = (x_1, \cdots, x_n)\).
Finding maxima

Remember: the likelihood and log-likelihood functions depend on all the data \((\vec{x} = (x_1, \cdots, x_n))\) and all the parameters \(\mu = (\mu_1, \cdots, \mu_n)\),

but because we don’t know the parameters and have only the data before us, we consider it to be a function of the parameters.

Remember: We want to find the parameter values that will make the data we see the “most likely.”

\textit{THEREFORE}, we want to maximize \(L_{\vec{x}}(\mu)\) as a function of \(\mu\), and the resulting solutions will depend on the data \(\vec{x}\).

\textit{The estimators} \(\hat{\mu}_1, \cdots, \hat{\mu}_n\) \textit{had better depend on the data} \(\vec{x} = (x_1, \cdots, x_n)\)! 
Finding maxima, continued

Suppose the unknown parameter \( \mu \) was one-dimensional.

What is the first thing to try when we want to find maxima of a function of one variable?

Differentiate \( L_{\bar{x}}(\mu) \) with respect to the parameter \( \mu \), and solve the equation

\[
\frac{\partial L_{\bar{x}}(\mu)}{\partial \mu} = 0
\]

for \( \mu \).

Of course, in this case the partial derivative is just a first derivative; the symbol is used to distinguish derivatives with respect to \( \mu \) from derivatives with respect to \( x_i \).
Finding maxima, continued

Suppose the unknown parameter \( \mu \) was one-dimensional.

What is the first thing to try when we want to find maxima of a function of one variable?

Differentiate \( L_{\bar{x}}(\mu) \) with respect to the parameter \( \mu \), and solve the equation

\[
\frac{\partial L_{\bar{x}}(\mu)}{\partial \mu} = 0
\]

for \( \mu \).

Of course, in this case the partial derivative is just a first derivative; the symbol is used to distinguish derivatives with respect to \( \mu \) from derivatives with respect to \( x_i \).
Finding maxima, continued

Suppose the unknown parameter $\mu$ was one-dimensional.

What is the first thing to try when we want to find maxima of a function of one variable?

Differentiate $L_{\bar{x}}(\mu)$ with respect to the parameter $\mu$, and solve the equation

$$\frac{\partial L_{\bar{x}}(\mu)}{\partial \mu} = 0$$

for $\mu$.

Of course, in this case the partial derivative is just a first derivative; the symbol is used to distinguish derivatives with respect to $\mu$ from derivatives with respect to $x_i$. 
Finding maxima for higher-dimensional parameters

Suppose $\mu = (\mu_1, \cdots, \mu_n)$ is an unknown $n$-dimensional parameter.

Recall that we want to maximize $L_\bar{x}(\mu)$ with respect to $\mu_1, \cdots, \mu_n$.

First thing to try here as well: determine where the gradient of $L_\bar{x}(\mu)$ vanishes. This yields the equations:

$$\nabla_{\mu} (L_{\bar{x}}(\mu)) = 0$$

that is,

$$\frac{\partial L_{\bar{x}}(\mu)}{\partial \mu_1} = 0, \quad \frac{\partial L_{\bar{x}}(\mu)}{\partial \mu_2} = 0 \quad \cdots \quad \frac{\partial L_{\bar{x}}(\mu)}{\partial \mu_n} = 0$$
Finding maxima for higher-dimensional parameters

Suppose \( \mu = (\mu_1, \cdots, \mu_n) \) is an unknown \( n \)-dimensional parameter.

Recall that we want to maximize \( L_{\vec{x}}(\mu) \) with respect to \( \mu_1, \cdots, \mu_n \).

First thing to try here as well: determine where the gradient of \( L_{\vec{x}}(\mu) \) vanishes. This yields the equations:

\[
\nabla_{\mu} (L_{\vec{x}}(\mu)) = 0
\]

that is,

\[
\frac{\partial L_{\vec{x}}(\mu)}{\partial \mu_1} = 0, \quad \frac{\partial L_{\vec{x}}(\mu)}{\partial \mu_2} = 0 \quad \cdots \quad \frac{\partial L_{\vec{x}}(\mu)}{\partial \mu_n} = 0
\]
Finding maxima for higher-dimensional parameters

Suppose $\mu = (\mu_1, \cdots, \mu_n)$ is an unknown $n$-dimensional parameter.

Recall that we want to maximize $L_{\vec{x}}(\mu)$ with respect to $\mu_1, \cdots, \mu_n$.

First thing to try here as well: determine where the gradient of $L_{\vec{x}}(\mu)$ vanishes. This yields the equations:

$$\nabla_\mu (L_{\vec{x}}(\mu)) = 0$$

that is,

$$\frac{\partial L_{\vec{x}}(\mu)}{\partial \mu_1} = 0, \quad \frac{\partial L_{\vec{x}}(\mu)}{\partial \mu_2} = 0 \quad \cdots \quad \frac{\partial L_{\vec{x}}(\mu)}{\partial \mu_n} = 0$$
Invariance principle

Maximum likelihood estimators satisfy a very useful invariance principle:

**Theorem** (Invariance principle). Let $(\hat{\mu}_1, \cdots, \hat{\mu}_n)$ be the maximum likelihood estimators for the parameters $(\mu_1, \cdots, \mu_n)$. Then the maximum likelihood estimator of any function $h(\mu_1, \cdots, \mu_n)$ of these parameters is the function $h(\hat{\mu}_1, \cdots, \hat{\mu}_n)$ of the maximum likelihood estimators.

From this invariance principle, for instance, if we know the MLE for $\sigma^2 = V(X)$, then we can get the MLE for the standard deviation $\sqrt{\sigma^2}$. 
Invariance principle

Maximum likelihood estimators satisfy a very useful *invariance principle*:

*Theorem* (Invariance principle). Let \((\hat{\mu}_1, \cdots, \hat{\mu}_n)\) be the maximum likelihood estimators for the parameters \((\mu_1, \cdots, \mu_n)\). Then the maximum likelihood estimator of any function \(h(\mu_1, \cdots, \mu_n)\) of these parameters is the function \(h(\hat{\mu}_1, \cdots, \hat{\mu}_n)\) of the maximum likelihood estimators.

From this invariance principle, for instance, if we know the MLE for \(\sigma^2 = V(X)\), then we can get the MLE for the standard deviation \(\sqrt{\sigma^2}\).
Large-sample properties of the MLE: Efficiency

Under very general conditions on the random sample, the MLE for a parameter \( \mu \) is approximately the minimum variance unbiased estimator for \( \mu \).

That is, for large samples, the MLE is nearly unbiased and has variance that is nearly as small as possible (i.e. the variance of the MLE nearly achieves the Cramér-Rao lower bound).
Under very general conditions on the random sample, the MLE for a parameter $\mu$ is approximately the minimum variance unbiased estimator for $\mu$.

That is, for large samples, the MLE is nearly unbiased and has variance that is nearly as small as possible (i.e. the variance of the MLE nearly achieves the Cramér-Rao lower bound).
Examples: compute the MLE by hand!

Compute the MOM estimator and the MLE for a normal distribution with mean $\mu$ and variance $\sigma^2$. How do they compare?

Can you use the invariance principle to find the MLE for the standard deviation?

Compute the MLE for data that is exponential with parameter $\lambda$.

Compute the MLE for data that is uniform on $[0, \theta]$, $\theta$ unknown. What issues arise? Can you always rely on differentiation alone to find maxima?
Examples: compute the MLE by hand!

Compute the MOM estimator and the MLE for a normal distribution with mean $\mu$ and variance $\sigma^2$. How do they compare?

Can you use the invariance principle to find the MLE for the standard deviation?

Compute the MLE for data that is exponential with parameter $\lambda$.

Compute the MLE for data that is uniform on $[0, \theta]$, $\theta$ unknown. What issues arise? Can you always rely on differentiation alone to find maxima?
Examples: compute the MLE by hand!

Compute the MOM estimator and the MLE for a normal distribution with mean $\mu$ and variance $\sigma^2$. How do they compare?

Can you use the invariance principle to find the MLE for the standard deviation?

Compute the MLE for data that is exponential with parameter $\lambda$.

Compute the MLE for data that is uniform on $[0, \theta]$, $\theta$ unknown. What issues arise? Can you always rely on differentiation alone to find maxima?
Examples: compute the MLE by hand!

Compute the MOM estimator and the MLE for a normal distribution with mean $\mu$ and variance $\sigma^2$. How do they compare?

Can you use the invariance principle to find the MLE for the standard deviation?

Compute the MLE for data that is exponential with parameter $\lambda$.

Compute the MLE for data that is uniform on $[0, \theta]$, $\theta$ unknown. What issues arise? Can you always rely on differentiation alone to find maxima?
MLEs for Weibull data

Set up the equations that would yield the MLE for the Weibull distribution with parameters \( \alpha, \beta \). Can you solve them explicitly?

Recall the Weibull distribution with parameters \( \alpha, \beta \) has density

\[
f(t) = \begin{cases} 
\frac{\alpha}{\beta^\alpha} t^{\alpha-1} e^{-\left(\frac{t}{\beta}\right)^\alpha} & t \geq 0 \\
0 & t < 0 
\end{cases}
\]

In particular, the mean and variance are given by

\[
E(X) = \beta \Gamma \left(1 + \frac{1}{\alpha}\right)
\]

\[
V(X) = \beta^2 \left\{ \Gamma \left(1 + \frac{2}{\alpha}\right) - \left[ \Gamma \left(1 + \frac{1}{\alpha}\right) \right]^2 \right\}
\]
MLEs for Weibull using R

To find MLEs using R, we need the `optim` function, which minimizes the negative of the log-likelihood function.

First, we need to write the negative of the log likelihood function for Weibull-distributed data.

Next, we need to use the `optim` function to determine the minimizers.
Some sample code: first write negative-log-likelihood function

```r
neg.log.like<-function(theta,Y){

    shape1<-theta[1]
    scale1<-theta[2]

    n<-length(Y) # number of samples
    f<-rep(NA,n)

    for(i in 1:n){
        f[i]<-dweibull(Y[i],shape=shape1,scale=scale1,log=TRUE)
    }
    return(-sum(f))
}
```
Next, generate samples from Weibull with certain shape and scale parameters:

```r
set.seed(10)
n<-30
True.Shape<-2
True.Scale<-5
sample<-rweibull(n,True.Shape,True.Scale)
```
# More sample code: using `optim`

```r
### Estimate MLE
theta0 <- c(1,1)
fit1 <- optim(theta0, 
neg.log.like, 
Y = sample, 
hessian = TRUE)
fit1
avar1 <- solve(fit1$hessian)
result1 <- rbind 
  (mle = fit1$par, 
   se = sqrt(diag(avar1)), 
   lower = fit1$par - 1.96 * sqrt(diag(avar1)), 
   upper = fit1$par + 1.96 * sqrt(diag(avar1))))
colnames(result1) <- c("Shape","Scale")
print(result1)
```
MLEs for the Gamma distribution

Suppose $X_1, \cdots, X_n$ form a random sample from a Gamma distribution with parameters $\alpha$ and $\beta$.

Calculate the maximum likelihood estimator for $\mu = \alpha \beta$.

Derive the equations whose solution yields the MLEs for $\alpha$ and $\beta$, respectively. Can they be solved explicitly? How can you use R to find the MLEs?
MLEs for the Gamma distribution

Suppose $X_1, \cdots, X_n$ form a random sample from a Gamma distribution with parameters $\alpha$ and $\beta$.

Calculate the maximum likelihood estimator for $\mu = \alpha \beta$.

Derive the equations whose solution yields the MLEs for $\alpha$ and $\beta$, respectively. Can they be solved explicitly? How can you use R to find the MLEs?