We will talk about...

1. Properties of Maximum Likelihood Estimators
We will talk about...

1. Properties of Maximum Likelihood Estimators
2. Maximum Likelihood Estimators for multi-parametric models
Topics
We will talk about...

1. Properties of Maximum Likelihood Estimators
2. Maximum Likelihood Estimators for multi-parametric models
3. Sufficient Statistics
Invariance

Example: Lifetimes of electronic components

Example:

In the previous example about the lifetimes of electronic components, the parameter $\theta$ was interpreted as the failure rate of electronic components (ie. Number of failures per year).
The inverse $\psi = 1/\theta$ is the average lifetime of electronic components.

How can we calculate the MLE of $\psi$? Is there any relationship between the MLE of $\theta$ and the MLE of $\psi$?
The answer is True and it is explained in the following Theorem.
Invariance property of MLE’s

Theorem

Invariance Theorem

If \( \hat{\theta} \) is the maximum likelihood estimator of \( \theta \) and if \( g \) is a one-to-one function, then \( g(\hat{\theta}) \) is a MLE of \( g(\theta) \)

Example:
In the example about the lifetime components, we computed the observed MLE value of \( \hat{\theta} = 0.455 \). Following the theorem, the observed value of \( \hat{\psi} \) would be \( 1/0.455 \)

Note: This invariance property can be extended to functions that are not one-to-one.
Consistency property of MLE’s

Convergence of a sequence of MLE’s

Consistency property

Consider a random sample taken from a distribution with parameter $\theta$. Suppose that for every large sample of size $n$ greater than some given minimum, there exists a unique MLE of $\theta$. Then under certain conditions, the sequence of MLE’s is a consistent sequence of estimators of $\theta$. This means that the MLE sequence converges in probability to the unknown value of $\theta$ as $n \to \infty$

Note: Required conditions need to prove this results will not be detailed here but are typically satisfied for most practical problems.
Estimating MLE’s by numerical computations

Example: Sampling from a Gamma distribution

Suppose that \( X_1, X_2, \ldots, X_n \) is a random sample from a Gamma distribution with pdf:

\[
f(x|\alpha) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} \exp(-x)
\]

Suppose that \( \alpha \) is unknown (\( \alpha > 0 \)). The likelihood function is:

\[
f_n(x|\alpha) = \frac{1}{\Gamma^n(\alpha)} \left( \prod_{i=1}^{n} x_i \right)^{\alpha-1} \exp\left(- \sum_{i=1}^{n} x_i \right)
\]

The MLE of \( \alpha \) will satisfy the equation:

\[
\frac{\partial \log f_n(x|\alpha)}{\partial \alpha} = 0
\]
Estimating MLE’s by numerical computations

Example: Sampling from a Gamma distribution (Cont.)

When we take the derivative of the likelihood function with respect to $\alpha$ we get the following equation:

$$\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \frac{1}{n} \sum_{i=1}^{n} \log x_i$$

The function $\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$ is the **digamma function**.

The unique value of $\alpha$ that satisfies the above equation will be the MLE of $\alpha$. This value can be determined using a numerical method as the **Newton’s method** or by using the tables of the digamma function available in several mathematical packages.
Estimating MLE’s by numerical computations

Newton’s method

Let $f(\theta)$ be a real-valued function. Newton’s method is used to find the solution of the equation:

$$f(\theta) = 0$$

We get an initial value $\theta = \theta_0$. The Newton method works by updating the initial guess value with the equation:

$$\theta_1 = \theta_0 - \frac{f(\theta_0)}{f'(\theta_0)}$$

You continue iterating until the results stabilize to a given value.
Suppose we observed $n = 20$ random variables $X_1, X_2, \ldots, X_{20}$ from a Gamma distribution with parameters $\alpha$ and $\beta = 1$. Suppose that $\frac{1}{n} \sum_{i=1}^{n} \log x_i = 1.22$ and $\frac{1}{n} \sum_{i=1}^{n} x_i = 3.679$. Since $E[X_i] = \alpha/\beta = \alpha$, this suggests an initial value for $\alpha_0 = 3.679$. We want to find $\alpha$ such that $\psi(\alpha) = 1.22$ or $f(\alpha) = \psi(\alpha) - 1.22 = 0$. The derivative of $f(\alpha)$ is $\psi'(\alpha)$ which is the *trigamma function*. The first iterate in the Newton method is:

$$\alpha_1 = \alpha_0 - \frac{\psi(\alpha_0) - 1.22}{\psi'(\alpha_0)} = 3.679 - \frac{1.1607 - 1.22}{0.3120} = 3.871$$

After few iterations the value stabilizes at 3.876.
Suppose that $X_1, X_2, \ldots, X_n$ is a sample from a Normal distribution with mean $\mu$ and variance $\sigma^2$ both unknown. The parameter vector is $\theta = (\mu, \sigma^2)$. The likelihood function $f_n(x|\mu, \sigma^2)$ will be again given by the equation:

$$f_n(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right]$$

This function must be maximized for all possible values of $\mu$ and $\sigma^2$, where $-\infty < \mu < \infty$. Again it is easier to maximize log $f_n(x|\mu, \sigma^2)$
Multi-parametric models

Example: Normal Distribution with unknown mean and variance (Cont.)

The steps to maximize $L(\theta)$ are as follows:

1. Write the log likelihood:

   $$L(\theta) = \log f_n(x|\mu, \sigma^2)$$

   $$= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

2. Assume $\sigma^2$ is fixed and find $\hat{\mu}(\sigma^2)$ that maximizes $L(\theta)$.

3. Find the value $\hat{\sigma}^2$ of $\sigma^2$ that maximizes $L(\theta')$, where $\theta' = (\hat{\mu}(\sigma^2), \sigma^2)$

4. Set the MLE estimator as the random vector: $(\hat{\mu}(\hat{\sigma}^2), \hat{\sigma}^2)$
Multi-parametric models

Example: Normal Distribution with unknown mean and variance (Cont.)

- **Second Step:** The value of $\hat{\mu}(\sigma^2)$ was found when the likelihood was maximized for $\sigma^2$ fixed or known. The solution was $\hat{\mu}(\sigma^2) = \bar{x}_n$
- **Third Step:** We set $\theta' = (\bar{x}_n, \sigma^2)$ and maximize

$$L(\theta') = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$$

We take the derivative with respect to $\sigma^2$, makes it equal to 0 and solves for $\sigma^2$. The derivative is:

$$\frac{d}{d\sigma^2} L(\theta') = -\frac{n}{2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$$

Setting this derivative to zero and solving for $\sigma^2$ we get:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$$
Multi-parametric models

Example: Normal Distribution with unknown mean and variance (Cont.)

- Fourth Step: The maximum likelihood estimator is set as \( (\hat{\mu}(\hat{\sigma}^2), \hat{\sigma}^2) ):

\[
\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2) = (\bar{X}_n, \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2)
\]

Notes:

- Second derivative of \( L(\theta') \) is negative at the value of \( \sigma^2 \) found. Therefore a maximum has been reached.
- The first coordinate of the MLE is the sample mean.
- The second coordinate is the sample variance.
Definition of a Sufficient Statistic

Example: Lifetimes of electronic components

In this example the MLE of the mean lifetime of electronic components was found. Note that in this estimation procedure we made use of the data through the value of the statistics: $X_1 + X_2 + X_3$. This is an example of a sufficient statistic. This concept was introduced by Fisher in 1922. In this example there is a function $T = r(X_1, X_2, \ldots, X_n)$ that summarizes all the information in the random sample. Knowledge of the individual values of the sample are not always necessary to get a good estimator of $\theta$. A statistics $T$ having this property is called a sufficient statistic.
Definition of a Sufficient Statistic

**Sufficient statistics**

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution indexed by parameter $\theta$. Let $T$ be a statistics. Suppose that for every $\theta$ and every possible value $t$ of $T$, the conditional distribution of $X_1, X_2, \ldots, X_n$ given $T = t$ and $\theta$ depends on $t$ but not on $\theta$. This means that the conditional distribution of $X_1, X_2, \ldots, X_n$ given $T = t$ and $\theta$ is the same for all values of $\theta$. We say that $T$ is a *sufficient statistic* for the parameter $\theta$.

Suppose you know $T$ and can simulate random variables $X'_1, X'_2, \ldots, X'_n$ such that for every $\theta$, the joint distribution of $X'_1, X'_2, \ldots, X'_n$ is the same as the joint distribution of $X_1, X_2, \ldots, X_n$. The statistics $T$ is sufficient in the sense that one can use $X'_1, X'_2, \ldots, X'_n$ in the same way as $X_1, X_2, \ldots, X_n$. 
Factorization criterion

How to find a sufficient statistic?

This method was developed by R.A. Fisher in 1922, J. Neyman in 1935 and P.R. Halmos and L.J. Savage in 1949.

**Theorem: Factorization criterion**

Let $X_1, X_2, \ldots, X_n$ form a random variable from either a continuous or a discrete distribution for which the pdf or the pf is $f(x|\theta)$. The value of $\theta$ is unknown and belongs to a parameter space $\Omega$. A statistic $T = r(X_1, X_2, \ldots, X_n)$ is a sufficient statistic *if and only if* the joint pdf or pf $f_n(x|\theta)$ can be factored as: $f_n(x|\theta) = u(x)v[r(x, \theta)]$ for all values of $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ and $\theta \in \Omega$

Functions $u$ and $v$ are non-negatives. $u$ depends on $x$ but it does not depend on $\theta$. $v$ depends on $\theta$ and depends on $x$ only through the value of the statistics $r(x)$. 
Factorization criterion

Example 1: Sampling from a Poisson distribution

Suppose that \( X = (X_1, X_2, \ldots, X_n) \) is a random sample from a Poisson distribution for which the true value of the mean is unknown. (\( \theta > 0 \)). Let \( r(x) = \sum_{i=1}^{n} x_i \). We can prove that \( T = r(X) = \sum_{i=1}^{n} X_i \) is a sufficient statistic for \( \theta \). The joint pf \( f_n(x|\theta) \) is:

\[
f_n(x|\theta) = \prod_{i=1}^{n} \frac{e^{-\theta x_i} \theta^{x_i}}{x_i!} = \left( \prod_{i=1}^{n} \frac{1}{x_i!} \right) e^{-n\theta} \theta^{r(x)}
\]

Let \( u(x) = \prod_{i=1}^{n} \frac{1}{x_i!} \) and \( v(t, \theta) = e^{-n\theta} \theta^{t} \). This means that \( f_n(x|\theta) \) has been factored as in the theorem. It follows that \( T = \sum_{i=1}^{n} X_i \) is a sufficient statistic for \( \theta \).
**Factorization criterion**

*Example 2: Sampling from a continuous distribution*

Suppose that $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from a continuous distribution:

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

We will show that $T = \prod_{i=1}^{n} X_i$ is a sufficient statistic for $\theta$. For $0 < x_i < 1 (i = 1, \ldots, n)$ the joint pdf is:

$$f_n(x|\theta) = \theta^n (\prod_{i=1}^{n} x_i)^{\theta-1} = \theta^n [r(x)]^{\theta-1}$$

If at least one value of $x_i$ is outside the interval $0 < x_i < 1$, $f_n(x|\theta) = 0$. If we set $u(x) = 1$ and $v(t, \theta) = \theta^n t^{\theta-1}$, $f_n(x|\theta)$ is factored as in the Theorem and $T = \prod_{i=1}^{n} X_i$ is a sufficient statistics for $\theta$. 
Thanks for your attention ...