Chapter 2: Point Estimation

We split this chapter into two parts: Part I: Finding estimators and Estimator evaluation, and Part II: Best Unbiased Estimator (UMVUE).

1 INTRODUCTION

We will first study how to find an estimator by two estimation principles and how to assess the goodness of estimators in this note.

Point Estimation:

The idea of point estimation is so simple that we just use a statistic $T(\boldsymbol{x})$ to estimate the unknown paramete of interest, say $g(\theta)$, where $\boldsymbol{x} = (x_1, \dots, x_n)'$ is a realization of the random sample $\boldsymbol{X} = (X_1, \dots, X_n)'$ **or** $\{X_i : i = 1, \dots, n\}$ of size n from a population with a pdf $f(\cdot\,|\theta)$ or pmf $p(\cdot\,|\theta)$ and θ is in the parameter space Θ.

In some cases, there is an obvious or natural point estimator of an unknown parameter. For instance, sample mean of a random sample is a natural point estimator of the population mean. However, when we leave such a simple case, we need a more methodical estimation techniques(s) that will at least give us some reasonable candidates for consideration. In the following section, we will study two most commonly used estimation approaches in statistics. They are (i) method of moments estimation and (ii) maximum likelihood estimation.

Remark:

[Parameter of interest] Most often, the parameter(s) of our interest to be estimated (called estimand) is a function of the unknown distribution parameter(s) θ, say $g(\theta)$. For instance, we may be interested in μ^2, instead of μ, or σ/μ, instead of μ or σ only, etc.

[Estimator? Estimate?] An estimator is a funciton of the random sample $\boldsymbol{X}$, wheil an estimate is the realized value of the estimator that is obtained when a sample of data is actually taken.

[Why is called 'point' estimation?] Note that the statistic T indeed is a 'point' in R^k, where $k \geq 1$ represents the number of unknown parameters to be estimated. We use it to estimate $g(\theta)$, which is also a 'point' in R^k. So, that's why $T(\boldsymbol{x})$ is called a point estimate of $g(\theta)$.

Caution: We ONLY estimate an UNKNOWN parameter(s). For any KNOWN parameter, there is no point for us to estimate it!!!

2 Methods of Finding Estimators

In practice, there are a lot of estimation techniques which can be used to estimate an unknown parameter(s), but we only detail two methods --- method of moments estimation and maximum likelihood estimation, partially because both of them are most popular in statistics and have some desirable properties, such as asymptotically unbiased and asymptotically normal.

2.1 Method of Moments Estimation

The method of moments estimation is historically one of the oldest estimation methods in statistics, dating back at least to Karl Pearson in the late 1800s.

Karl Pearson (1857-1936)

Pearson's thinking underpins many of the 'classical' statistical methods which are still in common use today. Examples of his contributions are correlation coefficient, p-value, Pearson's goodness-of-fit test, and so on.

As its name suggests, this method is related to moments. The motivation of this method is that in some situations, the parameter of interest can be written as a function of the population moments about zero.

Basic idea of the method of moments estimation:

If the function can be specified, then we replace the population moments by their corresponding sample moments. The function of these sample moments is called ***the method of moments estimator*** (MME) for the parameter of interest.

We also use the abbreviation MME to stand for the method of moments estimate when we are talking of the realized value of the estimator.

In general, if there are k unknown parameters to be estimated, then the FIRST k or more population moments (about zero), i.e. $\mu_i' = E(X^i), for\ i = 1, \dots, k,$ ..., are required to involve.

More formally, we have the following definition of MME.

Definition (MME): Suppose that there are k unknown parameters $\theta_1, \dots, \theta_k$. If we can rewrite them in them of the first k or more moments, i.e.

$$\begin{cases} \theta_1 = g_1(\mu_1', \mu_2', \dots, \mu_k', \dots) \\ \theta_2 = g_2(\mu_1', \mu_2', \dots, \mu_k', \dots) \\ \vdots \\ \theta_k = g_k(\mu_1', \mu_2', \dots, \mu_k', \dots) \end{cases},$$

then, the method of moments estimator (MME), denoted by $(\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_k)$, of $(\theta_1, \theta_2, \dots, \theta_k)$ is

$$\begin{cases} \tilde{\theta}_1 = g_1(\bar{X}, \overline{X^2}, \dots, \overline{X^k}, \dots) \\ \tilde{\theta}_2 = g_2(\bar{X}, \overline{X^2}, \dots, \overline{X^k}, \dots) \\ \vdots \\ \tilde{\theta}_k = g_k(\bar{X}, \overline{X^2}, \dots, \overline{X^k}, \dots) \end{cases}.$$

Remark:

- The method of moments estimation is "***quick and easy***", but MME obtained are often biased and it heavily relies on the existence of the required population moments.

- MME defined above ***may not be unique*** because the parameter can be written as different functions of moments, e.g. the parameter λ of a Poisson distribution is known to be the population mean μ_1' and population variance $\mu_2' - (\mu_1')^2$. ***One suggested way to fix this problem*** is to use fewer or lower moments to get MME. See the practice exercise for MME.

- **[Invariance property]** If $\tilde{\theta}_i$ is the MME for θ_i for $i = 1, \dots, k$, then $h(\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_k)$ is the MME for $h(\theta_1, \theta_2, \dots, \theta_k)$, where h is a known function.

- A sequence of $\{\tilde{\theta}_n \in R^k : n = 1,2, \dots\}$ or simply $\tilde{\theta}_n$ is ***consistent*** and ***asymptotically unbiased*** for θ. It is also ***asymptotically normally distributed***. To be more precise, under certain assumptions like $E|X|^{2k} < \infty$, we have

$$\sqrt{n}(\tilde{\theta}_n - \theta) \overset{d}{\to} N_k(\mathbf{0}, GHG'),$$

where G is a $k \times k$ matrix (suppose only the first k population moments are used in $g_1, \dots g_k$) with $\frac{\partial g_i}{\partial \mu_j'}$ as its $(i,j)^{th}$ entry and H is a $k \times k$ matrix with $\mu_{i+j}' - \mu_i'\mu_j'$ as its $(i,j)^{th}$ entry, for $i = 1, \dots, k$ and $j = 1, \dots, k$.

Example: Consider a r.s. of size n for X with $E|X|^4 < \infty$.

Take $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} = \begin{pmatrix} \mu_1' \\ \mu_2' - (\mu_1')^2 \end{pmatrix} = \begin{pmatrix} g_1(\mu_1', \mu_2') \\ g_2(\mu_1', \mu_2') \end{pmatrix}$.

Thus, we have

$$G = \begin{pmatrix} 1 & 0 \\ -2\mu_1' & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -2\mu & 1 \end{pmatrix}$$

and

$$H = \begin{pmatrix} \mu_2' - \mu_1'\mu_1' & \mu_{1+2}' - \mu_1'\mu_2' \\ \mu_{2+1}' - \mu_2'\mu_1' & \mu_{2+2}' - \mu_2'\mu_2' \end{pmatrix} = \begin{pmatrix} \sigma^2 & \mu_3' - \mu\mu_2' \\ \mu_3' - \mu\mu_2' & \mu_4' - (\mu_2')^2 \end{pmatrix}.$$

After some algebra, we have

$$GHG' = \begin{pmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{pmatrix}.$$

Note that

$$\tilde{\theta} = \begin{pmatrix} \bar{X}_n \\ \overline{X_n^2} - \bar{X}_n{}^2 \end{pmatrix} = \begin{pmatrix} \bar{X}_n \\ S_n^2 \end{pmatrix}.$$

Therefore, as n is large,

$$\sqrt{n}\left(\begin{pmatrix} \bar{X}_n \\ S_n^2 \end{pmatrix} - \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}\right) \xrightarrow{d} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{pmatrix}\right),$$

which implies that (by Multivariate Delta method, see the next page)

$$\sqrt{n}(S_n^2 - \sigma^2) \xrightarrow{d} N(0, \mu_4 - \sigma^4).$$

Using delta method with the above result and the condition that $\sigma^2 > 0$ yields

$$\sqrt{n}(S_n - \sigma) \xrightarrow{d} N\left(0, \frac{\mu_4 - \sigma^4}{4\sigma^2}\right).$$

Go to Practice Exercise for finding MME.

In Chapter 1, we have studied **delta method** to get a limiting distribution of a function of a random variable/ an estimator. To be more precise, it is a univariate delta method. Using it with CLT can give us a limiting result of a function of sample mean, like $1/\bar{X}_n$. It is quite natural for us to extend it to mutivariate cases.

Theorem (Multivariate Delta Method): Let $\{\boldsymbol{X}_n \in R^k : n = 1,2, \dots\}$ be a sequence of random vectors that for a constant vector $\boldsymbol{a} \in R^k$,

$$\sqrt{n}(\boldsymbol{X}_n - \boldsymbol{a}) \xrightarrow{d} \boldsymbol{Y}, \qquad as\ n \to \infty,$$

where $\boldsymbol{Y}$ is a random vector in R^k.

If a function $h: R^k \to R$ has a derivative $\nabla h(\boldsymbol{a}) \neq \boldsymbol{0}$, then we have

$$\sqrt{n}\big(h(\boldsymbol{X}_n) - h(\boldsymbol{a})\big) \xrightarrow{d} \nabla h(\boldsymbol{a})\boldsymbol{Y}, \qquad as\ n \to \infty,$$

where $\nabla h = \left(\frac{\partial h(t_1,t_2,\dots,t_k)}{\partial t_1}, \frac{\partial h(t_1,t_2,\dots,t_k)}{\partial t_2}, \cdots, \frac{\partial h(t_1,t_2,\dots,t_k)}{\partial t_k}\right)$.

In particular, if $\boldsymbol{Y} \sim N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$, then we have

$$\sqrt{n}\big(h(\boldsymbol{X}_n) - h(\boldsymbol{a})\big) \xrightarrow{d} \nabla h(\boldsymbol{a}) N_k(\boldsymbol{0}, \boldsymbol{\Sigma}) = N_k\left(\boldsymbol{0}, \big(\nabla h(\boldsymbol{a})\big)\boldsymbol{\Sigma}\big(\nabla h(\boldsymbol{a})\big)'\right).$$

For the equality above, please refer to Lemma 1 in Chapter 1.

In the following, we will discuss (the details in class) the application of Multivariate Delta Method to get the Fisher z-transformation for Correlation Coefficient.

STRENGTH OF LINEARITY

When the relationship is shown to be ***linear***, then we can use the **Pearson's correlation coefficient** to quantify the strength/degree of ***linearity***.

Pearson's correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

It is a dimensionless value between -1 and 1.

r is used to estimate a population ***correlation coefficient***

$$\frac{E[(X - E(X))(Y - E(Y))]}{\sqrt{Var(X)Var(Y)}} \in [-1, 1]$$

(denoted by ρ) of two random variables, say X and Y.

- If ρ is positive, then X and Y are said to be ***positively (linearly) correlated***. If it is near to $+1$, then it means a strong linear relationship between X and Y.

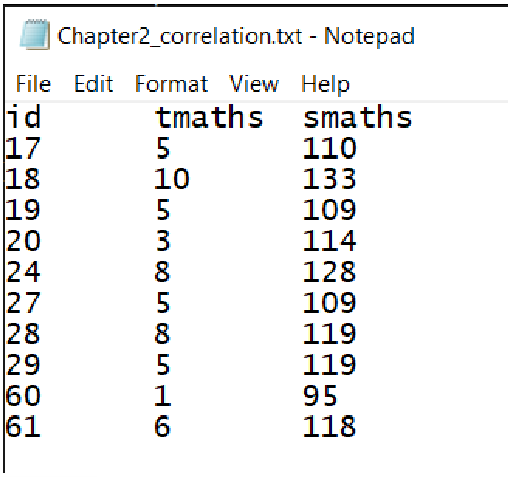
- If ρ is negative, then X and Y are said to be ***negatively (linearly) correlated***. If it is near to -1, then it also indicates a strong linear relationship between X and Y.

- If ρ is near to 0, then it ONLY indicates that the linearity between X and Y can be ignored, but it does NOT mean that X and Y have no relationship.

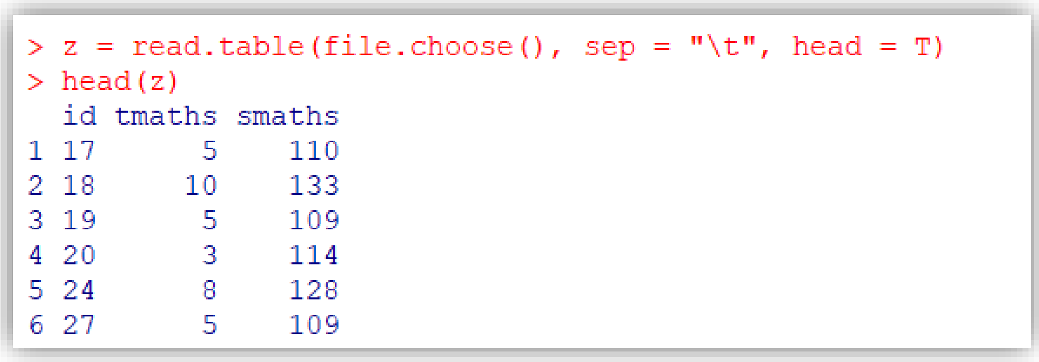
  If it is 0 exactly, then X and Y are said to be ***(linearly) UNcorrelated***.

R Corner: How do we find the Pearson’s correlation coefficient in R?

Consider the data (Chapter2_correlation.txt on canvas) for the students’ standardized maths attainment score, and teacher estimated score of students’ general maths ability based on a rating scale of 1 (well below average) to 10 (well above average), We would like to see if the standardized maths score (smaths) is related to the teacher estimate of maths ability (tmaths).


Chapter2_correlation.txt - Notepad
File Edit Format View Help

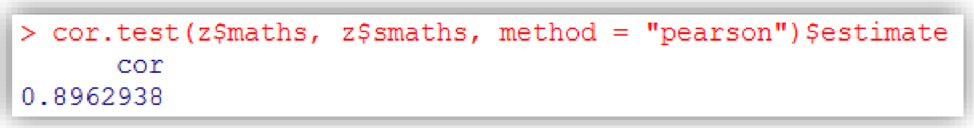
```
id      tmaths  smaths
17      5       110
18      10      133
19      5       109
20      3       114
24      8       128
27      5       109
28      8       119
29      5       119
60      1       95
61      6       118
```



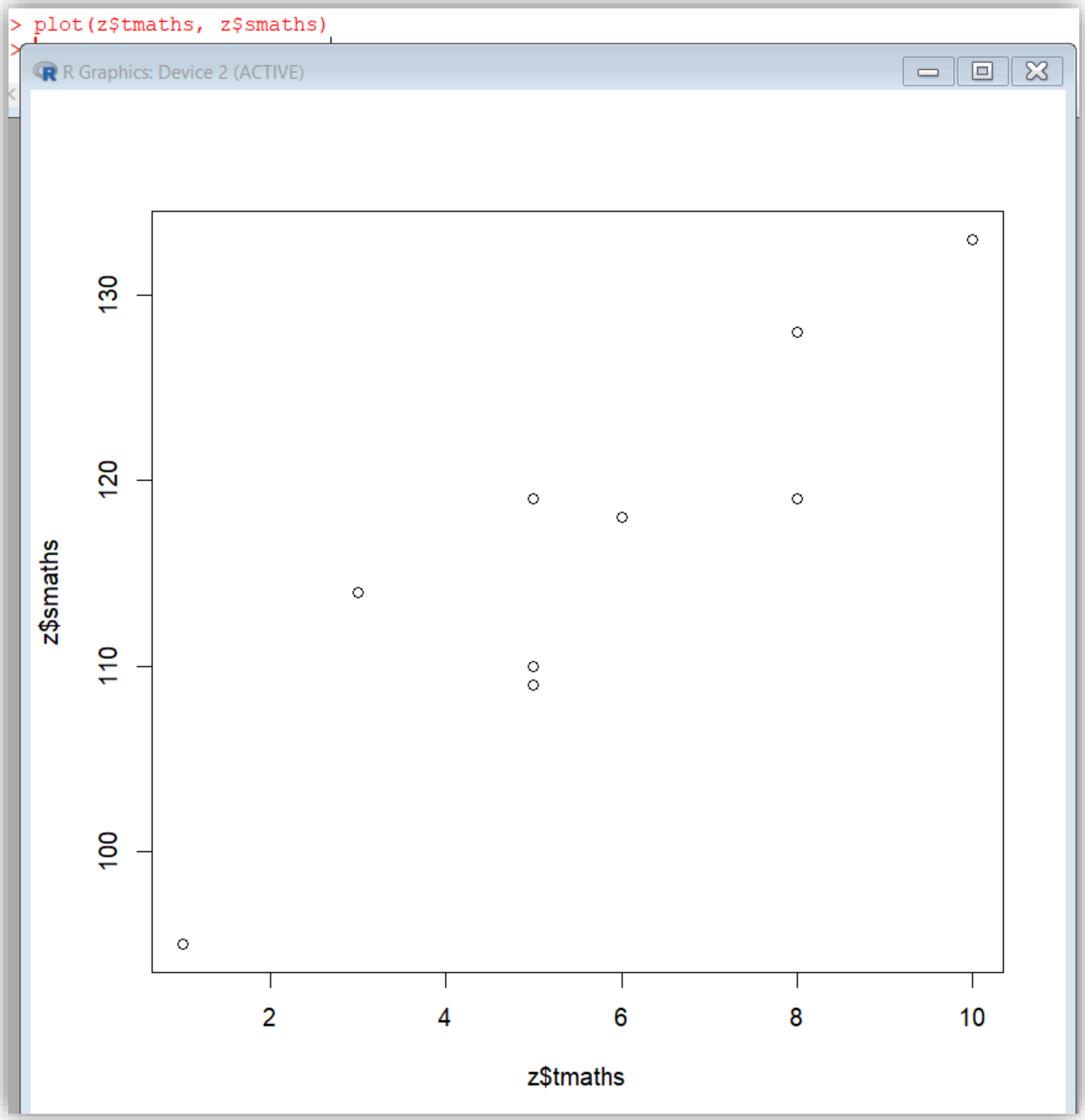
```
> z = read.table(file.choose(), sep = "\t", head = T)
> head(z)
  id tmaths smaths
1 17      5    110
2 18     10    133
3 19      5    109
4 20      3    114
5 24      8    128
6 27      5    109
```

Note that the scatter plot below indicates a positive and linear relationship between the two variables. So, we can now use the Pearson’s correlation coefficient to measure the strength of the linearity.

In R, ***cor.test*** can be used to get the result.



```
> cor.test(z$maths, z$smaths, method = "pearson")$estimate
      cor
0.8962938
```



```
> plot(z$tmaths, z$smaths)
>
```

Test For Population Correlation Coefficient

As mentioned before, the target parameter indeed is the population correlation coefficient ρ between X and Y.

How do we construct the confidence interval for ρ?

We need a point estimator and its exact (or approximated) distribution.

Undoubtedly, the point estimator we use here is just the random variable of r by replacing all (x_i, y_i), $\bar{x}$ and $\bar{y}$ by (X_i, Y_i), $\bar{X}$ and $\bar{Y}$, respectively.

Since it is not easy to get the exact distribution of the point estimator, R.A. Fisher in 1915 proposed a so-called ***Fisher z-transformation*** to a transformed estimator and its approximated normal distribution, and then make a back-transformation to produce the confidence interval for ρ.

Fisher z-transformation:

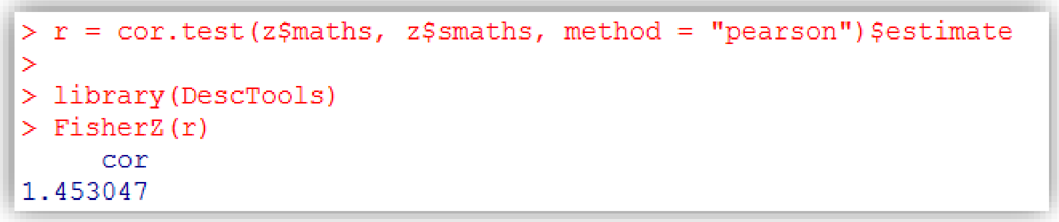
$$z_{Fisher} = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right),$$

where r is the sample Pearson's correlation coefficient. Taking the randomness of data into account, Fisher showed that the random variable Z_{Fisher} follows an approximated normal distribution with mean $\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$ and (after adjusted) variance $\frac{1}{n-3}$.

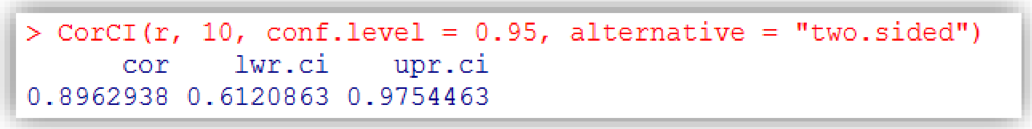
In R, we can use a library "*DescTools*" to convert r to a z or z to r using the Fisher transformation and its back transformation, and find the confidence intervals for a population correlation by using the following R scripts.

```
install.packages("DescTools") ← you may need to install this package first


library(DescTools)

FisherZ(r)

FisherZInv(z)

CorCI(r, n, conf.level = 0.95, alternative = c("two.sided", "less", "greater"))
```

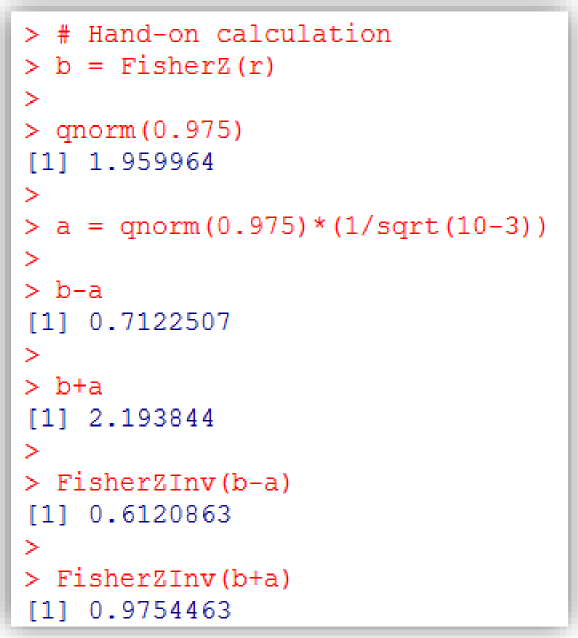
For instance,



```
> r = cor.test(z$maths, z$smaths, method = "pearson")$estimate
>
> library(DescTools)
> FisherZ(r)
     cor
1.453047
```



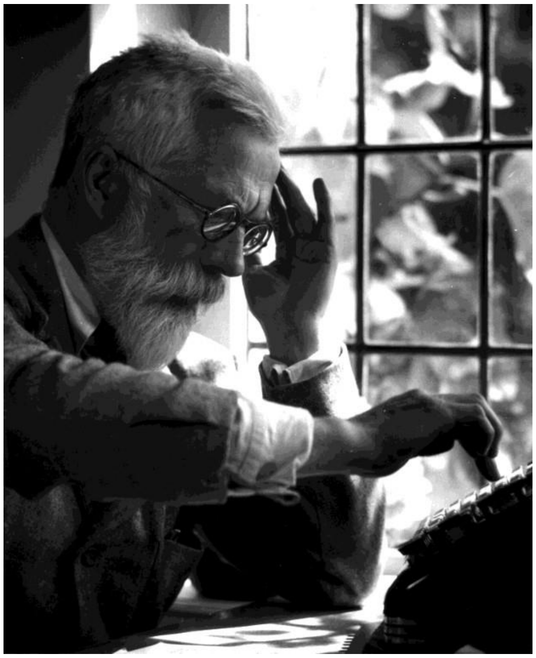
```
> CorCI(r, 10, conf.level = 0.95, alternative = "two.sided")
      cor     lwr.ci     upr.ci
0.8962938 0.6120863 0.9754463
```



```
> # Hand-on calculation
> b = FisherZ(r)
>
> qnorm(0.975)
[1] 1.959964
>
> a = qnorm(0.975)*(1/sqrt(10-3))
>
> b-a
[1] 0.7122507
>
> b+a
[1] 2.193844
>
> FisherZInv(b-a)
[1] 0.6120863
>
> FisherZInv(b+a)
[1] 0.9754463
```

2.2 Maximum Likelihood Estimation

The method of maximum likelihood is, by far, the most popular technique for deriving estimators. It was popularized in mathematical statistics by Ronald Aylmer Fisher in 1922. Nowadays, there are still a lot of research studying the properties of this estimation method.



Ronald Aylmer Fisher (1890-1962)

Fisher is one of the most prominent statisticians of the 19th-20th century. Other examples of his contributions are sufficiency, consistency, efficiency, Fisher information, genetical statistics, etc.

More details about him can be found in the article "How Ronald Fisher became a mathematical statistician" by Stephen M. Stigler.

Before showing how to find this estimator, let's first understand what the 'likelihood' is.

What is 'Likelihood'?

Consider a r.s. of size n from a population with a pdf $f(\cdot\ |\theta)$ or pmf $p(\cdot\ |\theta)$. After collection, we have the realization $\boldsymbol{x} = (x_1, \dots, x_n)'$. **The likelihood function is then defined by**

$L(\theta) = L(\theta_1, \theta_2, \dots, \theta_k|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\theta)$ for continuous cases and $L(\theta) = \prod_{i=1}^{n} p(x_i|\theta)$ for discrete cases. Note that the likelihood function can be used to quantify how the observed data is likely to occur.

Remark that $L(\theta)$ is a function of θ, with $\boldsymbol{x}$ held fixed. That is, the role of $\boldsymbol{x}$ and the parameter θ are interchanged between $L(\theta)$ and the joint pdf/pmf.

Basic idea of the maximum likelihood estimation:

It comes from **the statistical belief** that there is a highest chance of getting our current particular set of data. Thus, for each realization $\boldsymbol{x}$**, we want to find a value of** θ, denoted by $\hat{\theta}$, in Θ **at which** $L(\theta)$ **attains its maximum**. That is, we find a value --- ***maximum likelihood estimate*** (MLE) --- such that **our observed data is the most likely to occur**.

More formally, we have the following definition of MLE.

Definition (MLE): The maximum likelihood estimate is $\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmax}} L(\theta)$, which means

$$L(\hat{\theta}) = \max_{\theta\in\Theta} L(\theta),$$

where $\max\limits_{\theta\in\Theta}$ means the maximum over the parameter space Θ.

We also use the abbreviation MLE to stand for the maximum likelihood estimator when we are talking the random counterpart of the estimate.

In some cases, especially when differentiation is used, it is easier to work with a natural logarithm of $L(\theta)$, i.e. $l(\theta) = \log L(\theta)$, called ***log likelihood***, than it is to work with $L(\theta)$ directly. This is possible because the log function is strictly increasing, which implies that the maxima of $L(\theta)$ and $l(\theta)$ coincide.

Remark:

- MLE may be biased and it may not exist in Θ, especially when Θ is an open set, so for more general cases, we would define MLE in the closure $\overline{\Theta}$ of Θ. For instance, $\Theta = (0,1)$ and $\overline{\Theta} = [0,1]$. However, the MLE taking a value outside Θ is not a reasonable estimator.

- MLE defined above may not be unique. See the practice exercise for MLE.

- **[Invariance property]** If $\hat{\theta}_i$ is the MLE for θ_i for $i = 1, \dots, k$, then $h(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k)$ is the MLE for $h(\theta_1, \theta_2, \dots, \theta_k)$, where h is a known function.

- For $\theta \in R^k$, $\hat{\theta}_n$ is ***consistent, asymptotically unbiased, asymptotically efficient*** and ***asymptotically normally distributed***. To be more precise, under regularity assumptions, we have

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} N_k\left(\mathbf{0}, I_X^{-1}(\theta)\right),$$

where $I_X(\theta)$ is known as ***Fisher Information matrix*** (More details about this matrix will be discussed later) and it is a $k \times k$ matrix with the $(i,j)^{th}$ entry defined as

$$E\left[\left(\frac{\partial}{\partial\theta_i}\log f_X(X|\theta)\right)\left(\frac{\partial}{\partial\theta_j}\log f_X(X|\theta)\right)\right]$$

for $i = 1, \dots, k$ and $j = 1, \dots, k$.

There are three standard approaches to find MLE. **Our job is to find a global maximum!!!**

(i) If the parameter space Θ contains finitely many points, then an MLE can always be obtained by simply comparing finitely many value of (log) $L(\theta)$, for all $\theta \in \Theta$.

(ii) If $L(\theta)$ is differentiable on the interior of Θ, then one possible way of finding an MLE is to consider the values of $\theta = (\theta_1, \theta_2, \dots, \theta_k)'$ in the interior that solve the first-order/ likelihood/ log likelihood equations

$$\frac{\partial}{\partial \theta_i} L(\theta) = 0 \ \ or \ \frac{\partial}{\partial \theta_i} l(\theta) = 0, for\ i = 1, \dots, k.$$

However, this is just a necessary condition for a maximum (or minimum), not a sufficient condition. To be more precise, the solutions to the above equations are just the critical points, which may or may not be extrema. Furthermore, the zeros of the first derivative only locate the critical points in the interior of the domain of the (log) $L(\theta)$. If the maximum occur on the boundary, the first derivative may not be zero. Thus, the boundary must be checked separately for MLE.

[A special case for one-parameter cases] When $k = 1$, there is a case that we can get a global maximum easily. If there is a **<u>unique</u> critical point** and it has **a negative second derivative** of (log) $L(\theta)$, then it must be a global maximum. Note that for this case, we do not have to check any boundary point!!

Example: Consider a random sample of size n from $N(\theta, 1)$. Then,

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2}} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{\sum(x_i-\theta)^2}{2}}.$$

The first derivative of $\log L(\theta)$ being 0 is

$$0 = \frac{d}{d\theta} l(\theta) = \frac{d}{d\theta}\left[\frac{-n}{2}\log(2\pi)\right] + \frac{d}{d\theta}\left[-\frac{\sum(x_i-\theta)^2}{2}\right] = \sum(x_i - \theta),$$

which yields the solution $\hat{\theta} = \bar{x}$. To verify that it is, in fact, a global maximum of $\log L(\theta)$ (or $L(\theta)$), we first note that it is the unique solution to the first-order equation. Second, we can check that

$$0 = l''(\theta) = \frac{d^2}{d\theta^2} l(\theta)\bigg|_{\theta=\bar{x}} = -n < 0.$$

Therefore, $\hat{\theta} = \bar{x}$ is a global maximum --- MLE.

(iii) Another way to find an MLE is to abandon differentiation and proceed with a direct maximization. One general technique is to find a global upper bound on (log) $L(\theta)$ and then establish that there is a unique point for which the upper bound is attained.

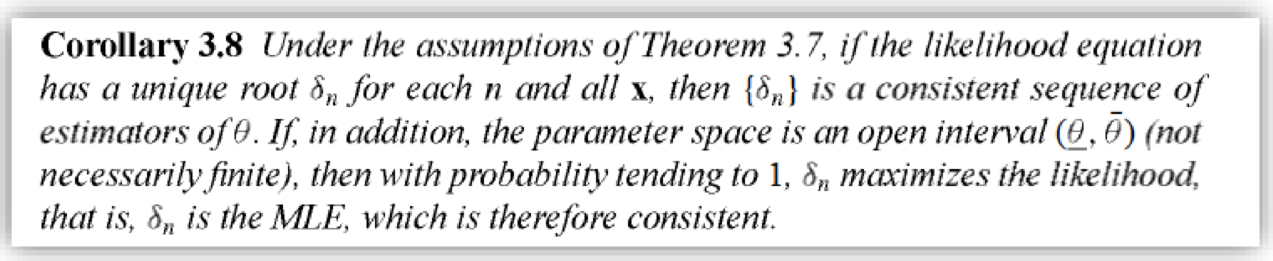
Example (cont'): Instead of using calculus, we can also show that $\hat{\theta} = \bar{x}$ is MLE algebraically. Note that $\sum(x_i - \theta)^2 \geq \sum(x_i - \bar{x})^2$ for any θ, where they are equal if and only if $\theta = \bar{x}$. Thus, for any $\theta \in \Theta$,

$$L(\theta) \leq L(\bar{x})$$

with equality if and only if $\theta = \bar{x}$. Hence, the MLE for θ is $\bar{x}$.

Remark that the global maximum finding problem in the above case can be solved for large n situations when some regularity conditions are required.

For instance,



Corollary 3.8 *Under the assumptions of Theorem 3.7, if the likelihood equation has a unique root δ_n for each n and all $\mathbf{x}$, then $\{\delta_n\}$ is a consistent sequence of estimators of θ. If, in addition, the parameter space is an open interval $(\underline{\theta}, \bar{\theta})$ (not necessarily finite), then with probability tending to 1, δ_n maximizes the likelihood, that is, δ_n is the MLE, which is therefore consistent.*

More details can be found in the book "Theory of Point Estimation" written by E. L. Lehmann and George Casella.

Go to Practice Exercise for finding MLE.

Practice Exercises for finding MME and MLE

1. Consider a r.s. of size n from $N(\mu_0, \sigma^2)$, where $\sigma^2 \in (0, \infty)$ is unknown and μ_0 is known. Use the method of moments to estimate σ^2.

2. Consider a r.s. with size n of $X \sim Poisson(\lambda)$, where $\lambda \in (0, \infty)$. Find the MME of λ.

3. Consider a r.s. with size n of $X \sim Binomial(1, \theta)$, where $\theta \in [0,1]$. Find the MLE of θ, and then get the MLE when $n = 10$ and $\sum_{i=1}^{n} x_i = 4$.

4. Consider a r.s. of size n of X from $Gamma(\alpha, \beta)$, where $0 < \alpha < \infty$ and $0 < \beta < \infty$.
   a. Find MME when
      i. β is unknown but α is known, say α_0.
      ii. α is unknown but β is known, say β_0.
      iii. Both are unknown.
      iv. Both are unknown, but the mean is known to be K_0.
   b. Find MLE when
      i. β is unknown but α is known, say α_0.
      ii. α is unknown but β is known, say β_0.
      iii. Both are unknown.
      iv. Both are unknown, but the mean is known to be K_0.

5. Consider a r.s. of size n from $N(\mu, \sigma^2)$, where $\sigma^2 \in (0, \infty)$ is unknown and $\mu \in (-\infty, \infty)$ is unknown. Find the MME and MLE of $\theta = (\mu, \sigma^2)'$.

6. Suppose X is from $U[0, \theta]$, where $0 < \theta < \infty$. Find the MLE of θ if a r.s. with size n of X is considered.

7. Suppose X is from $U[\theta - 1, \theta + 1]$, where $-\infty < \theta < \infty$. Find the MLE of θ if a r.s. with size n of X is considered.

8. Consider a r.s. of size n from $N(\mu, 1)$, where $0 \leq \mu < \infty$ is unknown. Find the MME and MLE of μ.

R corner: Practical skill of finding MLE with R

Except for a few cases, typically we are only able to write down (log) $L(\theta)$ but cannot maximize it analytically because there are no explicit solutions to the likelihood equation. However, there is still some hope of maximizing it ***numerically*** by R or other statistical packages and, hence, finding MLE. Note that when this is done, there is still always the question of whether a local or global maximum is found.

PRINCIPLE OF THE NUMERICAL SOLUTION TO LIKELIHOOD EQUATIONS

Example: Consider a r.s. with size n of $X \sim Cauchy(\theta)$. Find an MLE of θ.

First, try to get (log) $L(\theta)$. Since the pdf of X is $f_X(x|\theta) = \pi^{-1}[1+(x-\theta)^2]^{-1}$, the likelihood is $L(\theta) = \pi^{-n}\prod_{i=1}^{n}[1+(x_i-\theta)^2]^{-1}$ and

$$l(\theta) = -n\log\pi - \sum_{i=1}^{n}\log[1+(x_i-\theta)^2].$$

Setting

$$l'(\theta) = \frac{d}{d\theta}l(\theta) = \sum_{i=1}^{n}\frac{2(x_i-\theta)}{1+(x_i-\theta)^2} = 0$$

yields the MLE (Again, we then also have to check if it is a global maximum.) Note that the (solution) MLE cannot be solved explicitly in this case, but we can obtain/ approximate it by numerical method like ***Newton-Raphson Algorithm***.

According to Taylor, we have the following result:

$$0 = \frac{1}{n}l'(\hat{\theta}) \approx \frac{1}{n}l'(\theta) + (\hat{\theta}-\theta)\frac{1}{n}l"(\theta).$$

Thus,

$$\hat{\theta} \approx \theta - l'(\theta)[l"(\theta)]^{-1}.$$

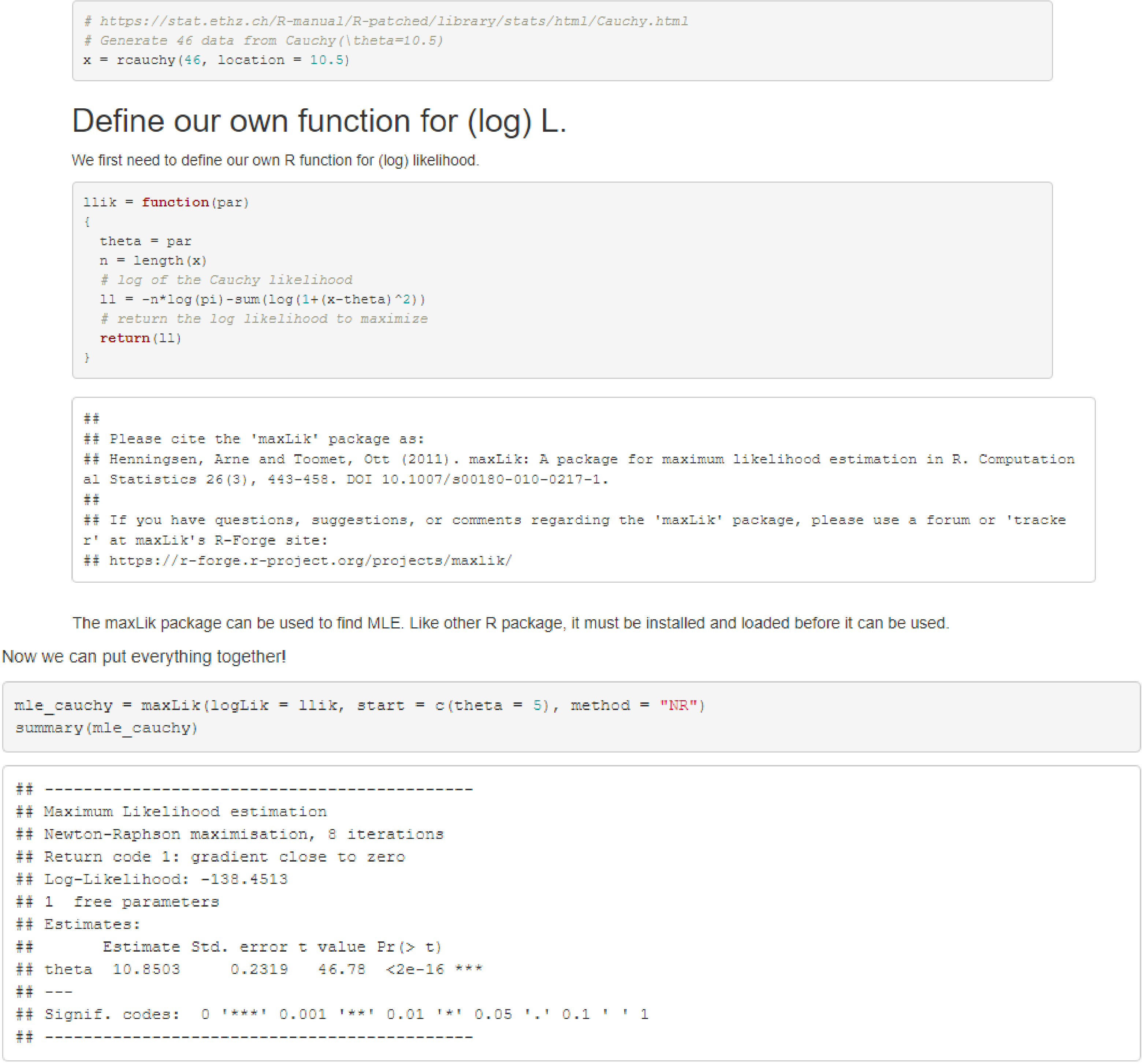
Newton-Raphson Algorithm:

$$\theta_{j+1} \approx \theta_j - l'(\theta_j)[l"(\theta_j)]^{-1}, j = 0, 1, 2, \ldots,$$

R

We would use the R package ***maxLik*** to maximize (log) $L(\theta)$ in the following. Other R functions like ***optim*** can also be used.

[Case 1: One unknown parameter]



```
# https://stat.ethz.ch/R-manual/R-patched/library/stats/html/Cauchy.html
# Generate 46 data from Cauchy(\theta=10.5)
x = rcauchy(46, location = 10.5)
```

Define our own function for (log) L.

We first need to define our own R function for (log) likelihood.

```
llik = function(par)
{
  theta = par
  n = length(x)
  # log of the Cauchy likelihood
  ll = -n*log(pi)-sum(log(1+(x-theta)^2))
  # return the log likelihood to maximize
  return(ll)
}
```

```
##
## Please cite the 'maxLik' package as:
## Henningsen, Arne and Toomet, Ott (2011). maxLik: A package for maximum likelihood estimation in R. Computation
al Statistics 26(3), 443-458. DOI 10.1007/s00180-010-0217-1.
##
## If you have questions, suggestions, or comments regarding the 'maxLik' package, please use a forum or 'tracke
r' at maxLik's R-Forge site:
## https://r-forge.r-project.org/projects/maxlik/
```

The maxLik package can be used to find MLE. Like other R package, it must be installed and loaded before it can be used.

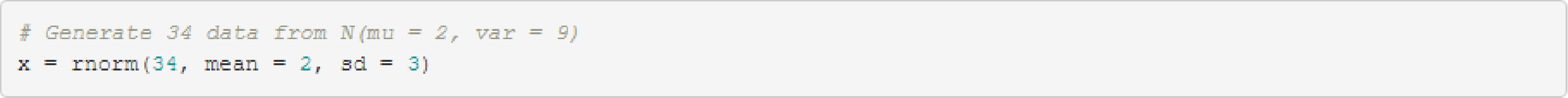
Now we can put everything together!

```
mle_cauchy = maxLik(logLik = llik, start = c(theta = 5), method = "NR")
summary(mle_cauchy)
```

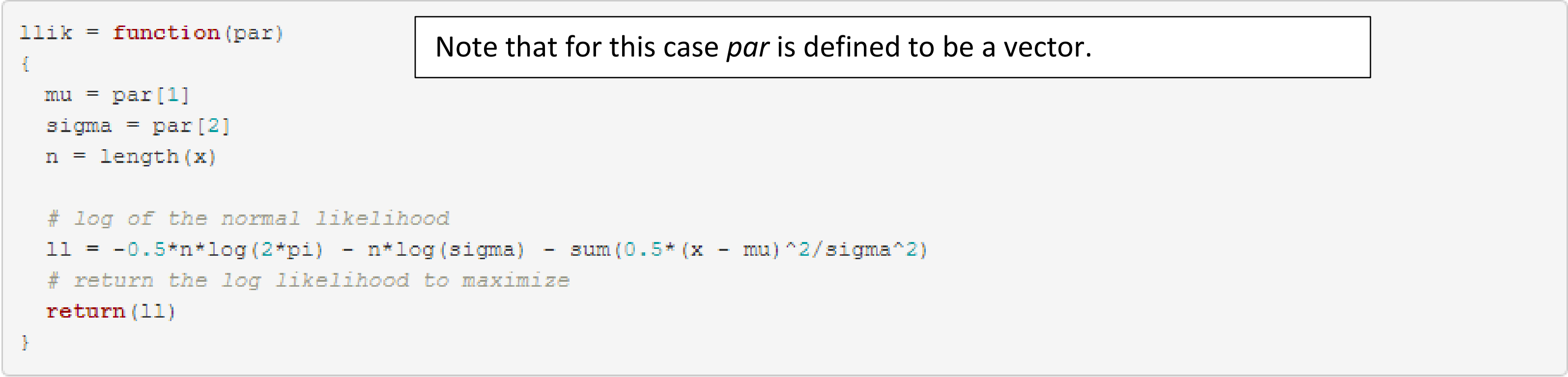
```
## --------------------------------------------
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 8 iterations
## Return code 1: gradient close to zero
## Log-Likelihood: -138.4513
## 1  free parameters
## Estimates:
##       Estimate Std. error t value Pr(> t)
## theta  10.8503     0.2319   46.78  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

[Case 2: Multi unknown parameters]

Next, we give an example of a normal distribution with two unknown parameters



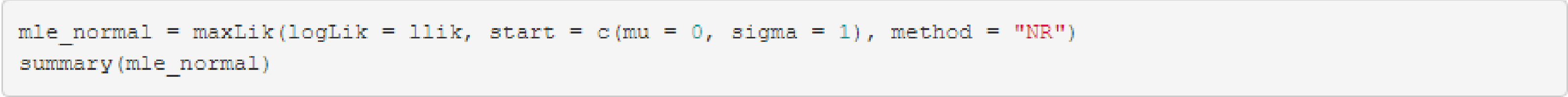
```
# Generate 34 data from N(mu = 2, var = 9)
x = rnorm(34, mean = 2, sd = 3)
```



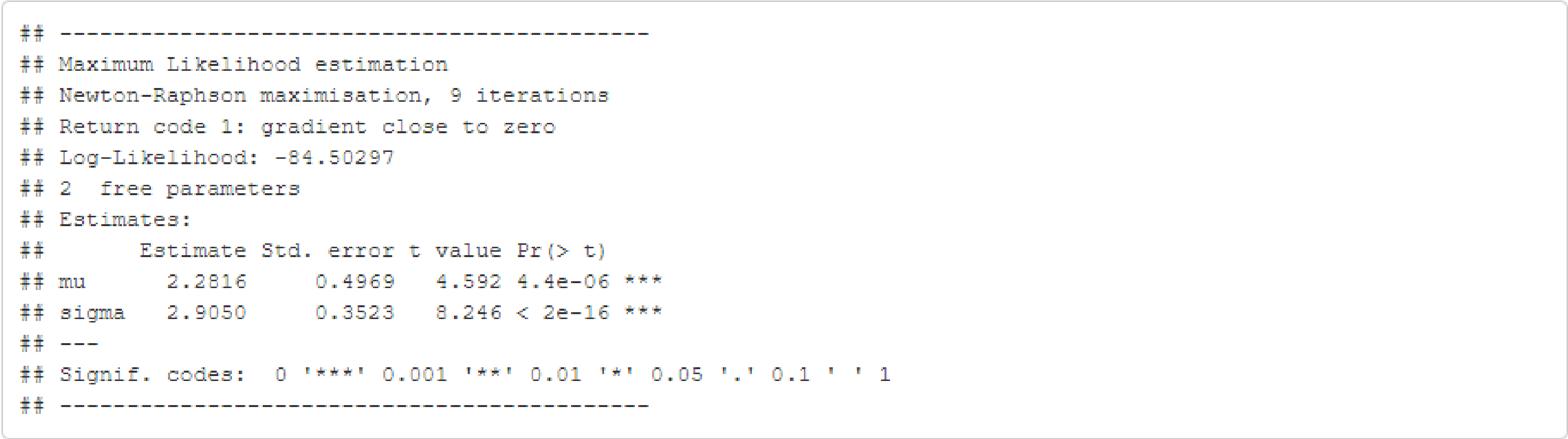
```
llik = function(par)
{
  mu = par[1]
  sigma = par[2]
  n = length(x)

  # log of the normal likelihood
  ll = -0.5*n*log(2*pi) - n*log(sigma) - sum(0.5*(x - mu)^2/sigma^2)
  # return the log likelihood to maximize
  return(ll)
}
```

Note that for this case *par* is defined to be a vector.

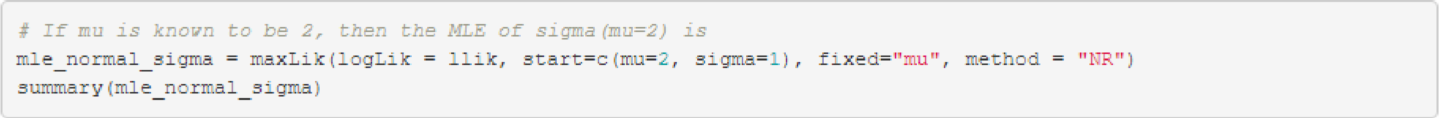
Thus, we have



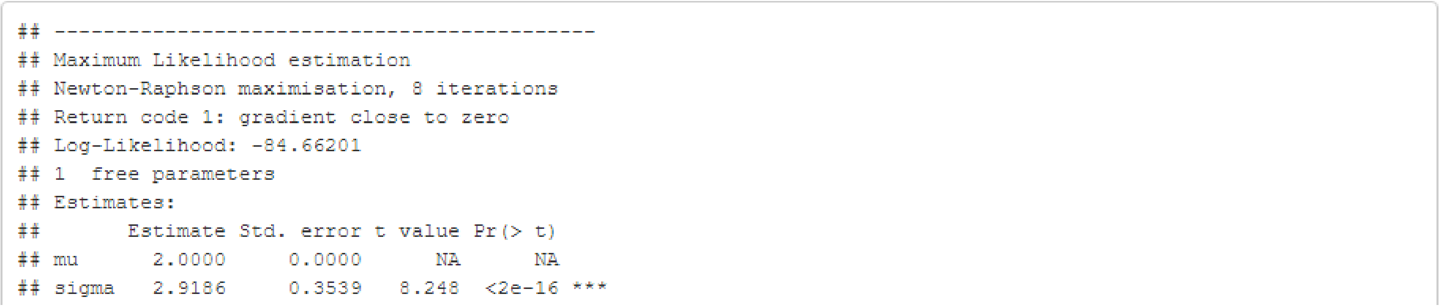
```
mle_normal = maxLik(logLik = llik, start = c(mu = 0, sigma = 1), method = "NR")
summary(mle_normal)
```



```
## --------------------------------------------
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 9 iterations
## Return code 1: gradient close to zero
## Log-Likelihood: -84.50297
## 2  free parameters
## Estimates:
##       Estimate Std. error t value Pr(> t)
## mu      2.2816     0.4969   4.592 4.4e-06 ***
## sigma   2.9050     0.3523   8.246 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```



```
# If mu is known to be 2, then the MLE of sigma(mu=2) is
mle_normal_sigma = maxLik(logLik = llik, start=c(mu=2, sigma=1), fixed="mu", method = "NR")
summary(mle_normal_sigma)
```



```
## --------------------------------------------
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 8 iterations
## Return code 1: gradient close to zero
## Log-Likelihood: -84.66201
## 1  free parameters
## Estimates:
##       Estimate Std. error t value Pr(> t)
## mu      2.0000     0.0000      NA      NA
## sigma   2.9186     0.3539   8.248  <2e-16 ***
```

3 Estimator Evaluation

In addition to using MME and MLE, we can also have a whole bunch of other estimators to estimate the parameter(s) of interest. Thus, our next problem about the point estimation is how to evaulate the goodness of the estimator, so that we can comapre different estimators and then get the best estimator in a class of estimators under consideration.

3.1 Mean Square Error (MSE)

For the evaluation of the goodness of an estimator, we consider the "closeness" of an estimator $\hat{\theta}(\boldsymbol{X})$, or simply $\hat{\theta}$, to the true unknown parameter θ. (Note that $\hat{\theta}$ used in this section 3 represents any estimator, it is not necessary to be MLE.)

So, it is reasonable to use a distance function to measure the closeness. Here we consider the squared error norm (or L_2 norm), $\left(\hat{\theta}-\theta\right)^2$, because of its easy calculation and nice properties.

Note that $\left(\hat{\theta}-\theta\right)^2$ is random, so we need to find a way to remove its randomness to get a numerical quantity for the comparison of different estimators. Conventionally, we fix this problem by taking the expectation. More precisely, we have

Definition (MSE): The mean squared error (MSE) of $\hat{\theta}$ for θ is defined by $E\left(\hat{\theta}(\boldsymbol{X})-\theta\right)^2$.

Note that MSE is a function of θ. For any two estimators, say $\hat{\theta}_1$ and $\hat{\theta}_2$, if for all $\theta \in \Theta$,

$$E\left(\hat{\theta}_1(\boldsymbol{X})-\theta\right)^2 \leq E\left(\hat{\theta}_2(\boldsymbol{X})-\theta\right)^2,$$

and the inequality is strict for at least one θ, then $\hat{\theta}_1$ is uniformly better than $\hat{\theta}_2$. Consider a class M of all estimators for θ, if there exists an estimator $\hat{\theta}^{**}$ in M that is uniformly better than any other estimators in M, then $\hat{\theta}^{**}$ is said to be a uniform minimum MSE estimator for θ in M.

However, such an estimator $\hat{\theta}^{**}$ in general does not exist because (i) we are too ***G***reedy to get a uniform 'best' estimator over all θ, and (ii) we are too ***G***enerous to consider too many (all) estimators for θ, even some of them are poor or not reasonable (like $\hat{\theta}=3423$).

For (i), to remove the dependence of MSE on θ, we can

1) Replace MSE by its maximum, and then compare estimators by looking at their respective maximum MSE, naturally preferring the one with the smallest maximum MSE over M. Such an estimator is said to be ***minimax***.

2) Average out θ, just as we average out the dependence on samples when going from $\left(\hat{\theta}-\theta\right)^2$ to $E\left(\hat{\theta}(\boldsymbol{X})-\theta\right)^2$. Then, a natural question then is

"How should θ be average out?"

The answer is based on "Bayesian statistics".

For (ii), we can restrict us to consider a particular class of estimators. Is it reasonable? Yes, it is because sometimes a "very poor" estimator can be a locally best estimator. For instance, $\hat{\theta}=3423$ is undoubtedly a poor estimator because no information of data is used, i.e. 3423 is always used to estimate an unknown parameter θ no matter what the observed data are. However, it is the best if the true value of θ is really equal to 3423. Thus, at least, we have to shrink a class of estimator to kick such a poor estimator out.

Obviously, we want to keep estimators with some nice properties. In this course, we keep mean-unbiased estimators.

Definition (Unbiasedness): If an estimator $\hat{\theta}$ satisfies $E\left(\hat{\theta}\right)=\theta$ for all $\theta\in\Theta$, then it is said to be mean-unbiased or unbiased for θ; otherwise, it is biased.

Interpretation:

The statement "$\hat{\theta}$ is unbiased for θ" means that in repeating sampling, $\hat{\theta}$ equals θ <u>on average</u>. That is, in the long run, the amounts by which $\hat{\theta}$ overestimates and underestimates θ will balance.

Note that $E\left(\hat{\theta}(\boldsymbol{X})-\theta\right)^2=Var\left(\hat{\theta}(\boldsymbol{X})\right)+bias^2$, where $bias=E\left(\hat{\theta}(\boldsymbol{X})\right)-\theta$.

So, if $\hat{\theta}$ is unbiased for θ, then its MSE is just its variance! In other words, we fix the bias to be zero, and then look for an estimator with the smallest variance (or the most efficient!!). Such an estimator is called a *UMVUE* --- ***uniform minimum variance unbiased estimator***.

In the whole 2nd part, we would learn different possible ways to 'catch' UMVUE.

Remarks:

1) According to Lemma 1 in Chapter 1, we know that the k^{th} sample moment (about 0) is unbiased for the k^{th} population moment (about 0), and sample variance S_{n-1}^2 is unbiased for θ, but S_n^2 is not. (Recall that S_n^2 is the MME and MLE for σ^2 when μ is unknown.)

2) Mean-unbiasedness is just one of the criteria we can use. ***It is NOT optimal***. Other criteria like Median-unbiasedness (the frequency, not the amount of over- and under-estimation, is considered) can also be used.

3) The **biased estimator is NOT always bad** because a bias estimator can have a smaller MSE than an unbiased estimator.

4) It is possible to have ***infinitely many diferent*** **or** ***NO*** **unbiased estimators** for θ.
   a. [Infinitely many] Consider a r.s. of size n from a distribution with a finite mean θ. All estimators in form of $\frac{\sum_{i=1}^{n}(a_i X_i)}{\sum_{i=1}^{n} a_i}$ are unbiased for θ, where $a_1, \dots, a_n \in R$ and $\sum_{i=1}^{n} a_i \neq 0$.
   b. [No] Suppose that we have a r.s. from Binomial(1, θ) with $g(\theta) = \frac{\theta}{1-\theta}$ as the parameter to be estimated. Note that there does not exist an unbiased estimator for $g(\theta)$.

5) An unbiased estimator ***may be a poor estimate***.
   For instance, consider a situation at which a telephonist has to leave the switch board for a short time. Let X be the number of telephone calls per 10 minutes. Suppose that $X \sim Poisson(\theta)$ and the telephonist is absent for 20 minutes. We want to estimate the probability that there is no calls during his absence, i.e. $e^{-2\theta}$.

   If we can only observe the number of calls received in the preceding 10 minutes, say X_1, and want to use a function h of X_1 to get an unbiased estimator for $e^{-2\theta}$, then we would have the result that $h(X_1) = (-1)^{X_1}$. Note that if x_1 is any odd integer, then -1 will be obtained to be an estimator of the above probability $e^{-2\theta}$. It is very Poor!!!

6) ***Unbiasedness does not have an invariance property***. That is, if $\hat{\theta}$ is unbiased for θ, then $h(\hat{\theta})$ may NO be unbiased for $h(\theta)$. For instance, $\bar{X}$ is unbiased for μ, but $\bar{X}^2$ is not unbiased for μ^2 when $\sigma > 0$.

3.2 LARGE-SAMPLE PROPERTIES OF A POINT ESTIMATOR

In additional to considering the above finite and fixed n property to evaluate the performance of a point estimator, most often we are interested in large-sample/large-n/ asymptotic properties of a sequence of estimators, when $n \to \infty$, because the calculations/ procedures are relatively simple. So, the evaluations that are difficuly to check or impossible in the finite-sample case will become routine.

1. Asymptotic Unbiaedness

Unbiasedness is nice, but in many cases our estimators, say MME or MLE, may not be unbiased. Luckily, when we assess the performance of a sequence of estimators asymptotically, we find that the biases of most biased estimators will disappear. If an estimator whose bias tends to 0 as $n \to \infty$, then it is said to be *asymptotically unbiased*. More formally, we have

Definition (Asymptotic Unbiasedness): A sequence of estimator, $\{\hat{\theta}_n : n = 1,2, \dots, \}$, based on a r.s. of size n is said to be ***asymptotically unbiased*** if $\lim_{n\to\infty} E(\hat{\theta}_n) = \theta$, for all $\theta \in \Theta$.

Note that S_n^2, the MLE and MME of σ^2, is asymptotically biased, although it is biasd in a finite-sample case. That is, in the asymptotic sense, S_n^2 can also enjoy a nice property.

2. Consistency

For consistency, $\hat{\theta}_n$ has to be arbitrarily close to θ with a high probability, i.e. $\hat{\theta}_n \xrightarrow{p} \theta$.

Definition (Consistency): A sequence of estimator, $\{\hat{\theta}_n : n = 1,2, \dots, \}$, based on a r.s. of size n is said to be ***consistency*** if, for any $\epsilon > 0$, $\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| \le \epsilon) = 1$, for all $\theta \in \Theta$.

Note that although asymptotically unbiasedness and consistency look similar, but they cannot imply each other in general. However, an estimator will be consistent if it is asymptotical unbiased AND its variance $Var(\hat{\theta}_n)$ tends to 0 as $n \to \infty$.

3. Asymptotic Normality

This property is important when we want to use $\hat{\theta}_n$ to make more statistical inference about θ, say find a confidence intercal or do a hypothesis testing.

Definition (Asymptotic Normality): A sequence of estimator, $\{\hat{\theta}_n : n = 1,2, \dots, \}$, based on a r.s. of size n is said to be ***asymptotically normal*** if $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma_\theta^2)$.

Note that the asymptotic variance of $\hat{\theta}_n$ is $\frac{\sigma_\theta^2}{n}$.

Recall that MME and MLE are both consistent, asymptotically unbiased, and asymptotically normal. So, are they asymptotically good? No, in the asymptotic sense, MLE is the optimal since its aymptotic variance is the smallest. To be more preicse, it can achieve the Cramér-Rao lower bound!! More details about this lower bound will be discussed in the 2nd part of this chapter.