Statistical Method

Univariate Data

Estimation for Statistical Modeling

Week08: October 23, 2022

Statistical inference

Statistical inference is the process of drawing conclusions about a certain population based on a subset of the data.

1. Assume an underlying model.
2. Use estimation methods to estimate the model parameters or a quantity of interest.
3. Construct the confidence intervals for the model parameters or the quantity of interest.
4. Perform the hypothesis testing and draw the statistical conclusion.

Methods of estimation in statistical method

Point estimation: (Chapter 5.3)

- Method of moments estimation (MME)
- Least square estimation (LSE)
- Maximum likelihood estimation (MLE)
- Bayesian estimation (BE)
  - Maximum a posteriori estimation (MAP)
  - (Optional) Expectation of the posterior distribution, which is obtained by Markov Chain Monte Carlo (MCMC)

Interval Estimation:

- By normal approximation (Chapter 5.3.5)
- By Bootstrapping method (Chapter 7.2)

Optimization problem

The estimation is an application of optimization.
The procedure is as follows:

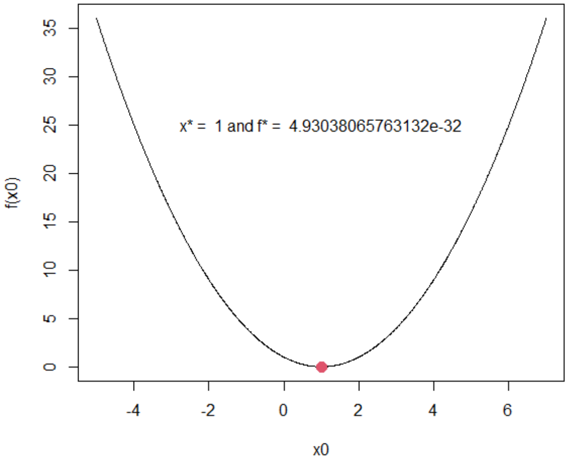
1. Problem definition and definition of decision variables
2. Specify the objective function
3. Theoretical derivation / efficient algorithms (numerical approaches) to obtain the possible solutions
4. Verify the solutions are appropriate and suitable for further conclusion.

Optimization problem: Example

Find the minimum of $(x-1)^2$.

1. Problem definition and definition of decision variables
   $\min f(x) = (x-1)^2$, and the decision variable is x.
2. Specify the objective function
   $\min_{x \in R} f(x) = (x-1)^2$.
3. Theoretical derivation / efficient algorithms (numerical approaches)
   - Theoretical method: $f'(x) = 2(x-1) =^{let} 0$, and $x = 1$.
   - Efficient algorithm: by the Brent optimization in optim(), the minimum is 4.930381e-32 at $x = 1$.
4. Verify the solutions are appropriate and suitable for further conclusion.
   - Theoretical method: $f''(x) = 2 > 0$ and $f(1) = 0$ is a minimum at $x = 1$.
   - Efficient algorithm: Draw the corresponding graph.

Verification



Estimation method: Purpose

Assume that the random samples $\boldsymbol{X}$ are from a the distribution $F(x; \boldsymbol{\theta})$.

- MME: Use information of data to fit the characteristic of the assumed distribution.
- LSE: Use the squared loss to minimize the error, $\sum_{i=1}^{n}(Obs_i - Fitted_i)^2$
- MLE: Likelihood function, which is a concept of the joint distribution of data with unknown parameters.
- BE: Posterior distribution of unknown parameters, which needs to give prior distributions for parameters first.
  - Maximum a posteriori estimation (MAP)
  - Expectation of the posterior distribution

Method	Purpose
MME	Match the theoretical moments with data.
LSE	Match data to the expected values with the minimum loss.
MLE	Let the joint possibilities of all data to be largest.
MAP	With prior knowledge, let the joint possibilities of data to be largest.

Estimation method: objective function

Assume that the random samples $\boldsymbol{X}$ are from a the distribution $F(x;\boldsymbol{\theta})$. The collected data is $\{x_i, i = 1, ..., \ldots\}$.

Method	Objective function
MME	$E_{\boldsymbol{\theta}}(X) = \frac{1}{n}\sum_{i=1}^n x_i$, $E_{\boldsymbol{\theta}}(X^2) = \frac{1}{n}\sum_{i=1}^n x_i^2$, ..., $E_{\boldsymbol{\theta}}(X^m) = \frac{1}{n}\sum_{i=1}^n x_i^m$
LSE	$\min_{\boldsymbol{\theta}} \sum_{i=1}^n (g(x_i) - Fitted_i)^2$.
MLE	$\max_{\boldsymbol{\theta}} f(x_1, \ldots, x_n; \boldsymbol{\theta}) =^{independent} \max_{\boldsymbol{\theta}} \Pi_{i=1}^n f(x_i; \boldsymbol{\theta})$.
MAP	Given prior knowledge $\pi(\boldsymbol{\theta})$, $\max_{\boldsymbol{\theta}} f(x_1, \ldots, x_n; \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta})$

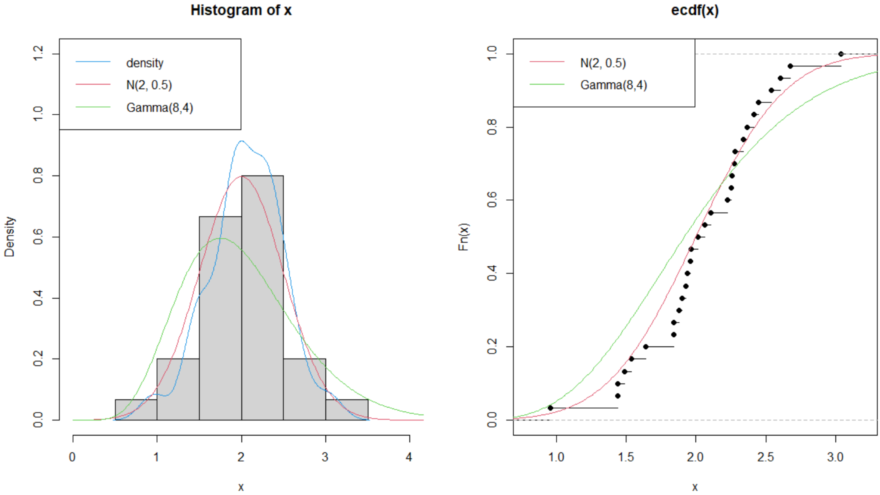
Model assumptions

Let X be a random variable with the following properties:

1. $F_\theta(x)$ and $f_\theta(x)$ are the cdf and pdf of random variable X.
2. Let $X_1, \ldots, X_n \sim F_\theta(\cdot)$, and they are independent and they have the identical distribution $F_\theta(\cdot)$. (iid)

Given a dataset, the values of θ in the distribution $F_\theta(\cdot)$ are unknown.

Example 1:


Histogram of x
density
N(2, 0.5)
Gamma(8,4)
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
1
2
3
4
x
ecdf(x)
N(2, 0.5)
Gamma(8,4)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.5
2.0
2.5
3.0
x

Example 1: MME (1)

Assume the data X is from $N(\mu, \sigma^2)$ or $Gamma(\alpha, \lambda)$.

1. Problem definition and definition of decision variables
   The decision variables for the normal distribution are (μ, σ^2) and for the gamma distribution are (α, λ).
2. Specify the objective function
   - Normal distribution: $E(X) = \mu$ and $E(X^2) = \sigma^2 + \mu^2$.
   - Gamma distribution: $E(X) = \alpha/\lambda$ and $E(X^2) = \alpha/\lambda^2 + \alpha^2/\lambda^2$.
3. Theoretical derivation / efficient algorithms (numerical approaches)
   - Normal distribution: $\hat{\mu}_{MME} = \frac{1}{n}\sum_{i=1}^n x_i$ and $\hat{\sigma}^2_{MME} = \frac{1}{n}\sum_{i=1}^n x_i^2 - \hat{\mu}^2_{MME}$.
   - Gamma distribution: $E(X) = \alpha/\lambda$ and $E(X^2) = \alpha/\lambda^2 + \alpha^2/\lambda^2$.
4. Verify the solutions are appropriate and suitable for further conclusion.
   Draw the corresponding graph.

Example 1: LSE (2)

Assume the data X is from $N(\mu, \sigma^2)$ or $Gamma(\alpha, \lambda)$.

1. Problem definition and definition of decision variables (μ, σ^2) and (α, λ).
2. Specify the objective function
   - Normal distribution: $\min_{\mu} \sum_{i=1}^{n} (x_i - \mu)^2$. (What about σ^2?)
   - Gamma distribution: $\min_{(\alpha,\lambda)} \sum_{i=1}^{n} (x_i - \alpha/\lambda)^2$.
3. Theoretical derivation / efficient algorithms (numerical approaches)
   - Normal distribution: $\hat{\mu}_{LSE} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $\hat{\sigma}^2_{LSE} =?$.
   - Gamma distribution: $\hat{\frac{\alpha}{\lambda}}_{LSE} = \frac{1}{n} \sum_{i=1}^{n} x_i$.
4. Verify the solutions are appropriate and suitable for further conclusion. Draw the corresponding graph.

Example 1: MLE (3)

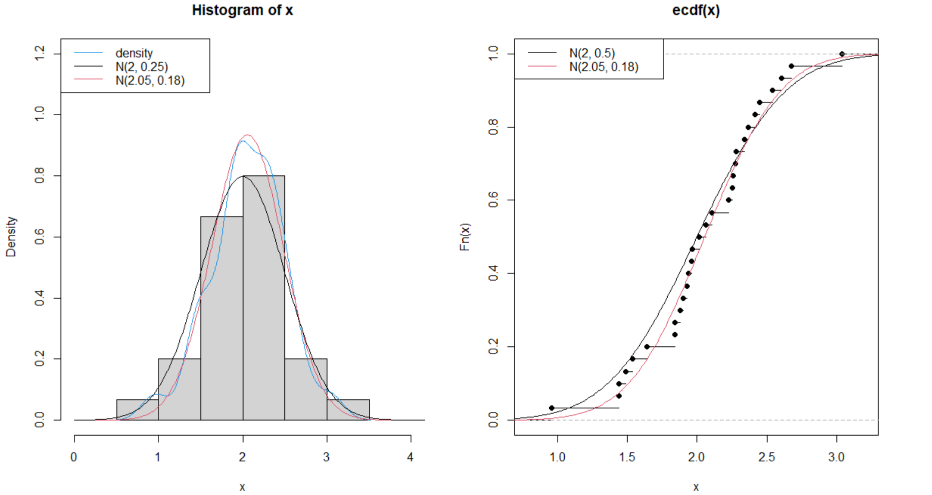
Assume the data X is from $N(\mu, \sigma^2)$ or $Gamma(\alpha, \lambda)$.

1. Problem definition and definition of decision variables (μ, σ^2) and (α, λ).
2. Specify the objective function
   - Normal distribution: $\max_{(\mu,\sigma^2)} \prod_{i=1}^{n} f(x_i; \mu, \sigma^2)$.
   - Gamma distribution: $\max_{(\alpha,\lambda)} \prod_{i=1}^{n} g(x_i; \alpha, \lambda)$.
3. Theoretical derivation / efficient algorithms (numerical approaches)
   - Normal distribution: $\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$ and $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
   - Gamma distribution: $\hat{\lambda}_{MLE} = \hat{\alpha}_{MLE}/\bar{x}$ and $\hat{\alpha}_{MLE}$ is obtained numerically.
4. Verify the solutions are appropriate and suitable for further conclusion. Draw the corresponding graph.

Example 1: MLE code for normal distribution

```
### Numerical method ###
likelihood.normal <- function(par, data){
  mu <- par[1]
  sig2 <- par[2]

  joint <- dnorm(data, mean = mu, sd = sqrt(sig2))

  return(prod(joint))
  return(sum(log(joint)))
  return(-sum(log(joint)))
}

likelihood.normal(c(1, 2), x)

optim(c(2, 1), likelihood.normal, data = x, control=list(fnscale=-1))
optim(c(2, 1), likelihood.normal, data = x)
```

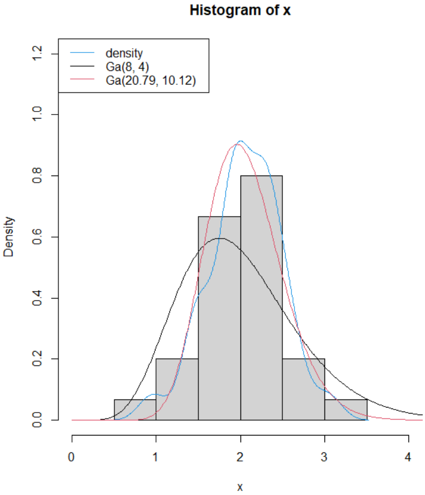
Example 1: MLE result for normal distribution

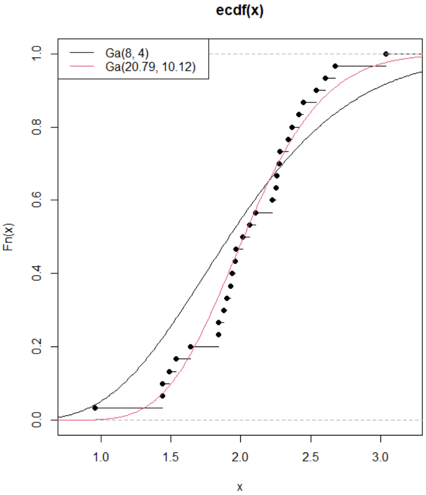


Example 1: MLE code for gamma distribution

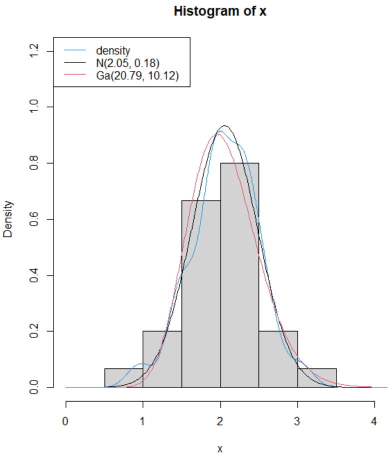
```
### MLE- gamma ###
likelihood.gamma <- function(par, data){
  alpha <- par[1]
  lambda <- par[2]

  joint <- dgamma(data, shape = alpha, rate = lambda)

  #return(prod(joint))
  #return(sum(log(joint)))
  return(-sum(log(joint)))
}

likelihood.gamma(c(8, 4), x)
mle.gamma <- optim(c(8, 4), likelihood.gamma, data = x)
```

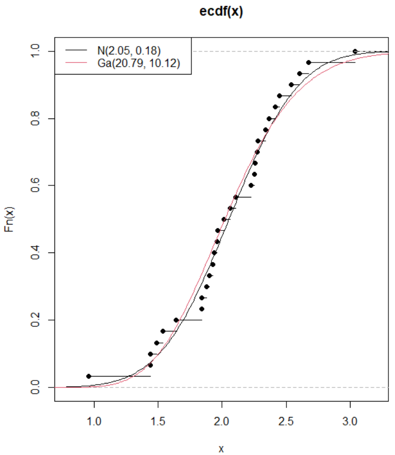
Example 1: MLE result for gamma distribution





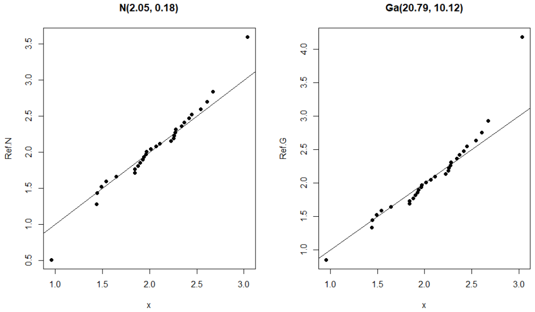
Example 1: Comparison between normal and gamma distribution





Example 1: Comparison between normal and gamma distribution

```
        Two-sample Kolmogorov-Smirnov test

data:  x and Ref.N
D = 0.1, p-value = 0.9259
alternative hypothesis: two-sided

> ks.G

        Two-sample Kolmogorov-Smirnov test

data:  x and Ref.G
D = 0.1331, p-value = 0.6644
alternative hypothesis: two-sided

> round(est.norm[1,],4)
    mu.hat sig2.hat obj.value
MLE 2.0555   0.1823   -17.037
> round(est.gamma[1,],4)
    alpha.hat lambda.hat obj.value
MLE   20.7964    10.1169   -18.176
>
```



Example: Lognormal distribution

$$X_1, \ldots, X_n \sim Log.Normal(\mu, \sigma)$$

How to write a code for obtaining the MLE?

1. function(...) + dlnorm(...) + optim (...)
2. function(...) + mle(...)
3. fitdistr(...)

Example: Lognormal distribution

```
> library(stats4)
> minuslogl <- function(mu, sigma) {
+   densities <- dlnorm(new_data$TV, meanlog=mu, sdlog=sigma)
+   -sum(log(densities))
+ }
> mle(minuslogl, start=list(mu=10, sigma=5))

Call:
mle(minuslogl = minuslogl, start = list(mu = 10, sigma = 5))

Coefficients:
       mu     sigma
2.4096583 0.7777092
```

```
> library(MASS)
> fitdistr(new_data$TV, densfun="lognormal")
     meanlog        sdlog
  2.409656712   0.777709043
 (0.003504225) (0.002477861)
```

Exercise

$$X_1, \ldots, X_n \sim Exp(\lambda) \text{ (rate parameter)}$$

- MME:
- LSE:
- MLE:

Short summary for Estimation

Objective function:

- LSE: $\sum_{i=1}^{n}(Obs_i - Fitted_i)^2$
  - linear: $y = a$, $y = a + bx$, ...
  - nonlinear: $y = ae^{bx}$,...
- MLE: Likelihood function, which is a concept of the joint distribution of data with unknown parameters.

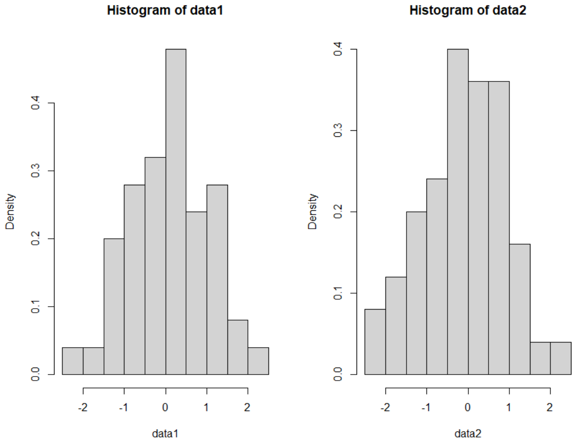
Statistical inference

Example:
Given the data, use the following steps to show if the mean of the population is 0.

1. Assume an underlying model.
2. Use estimation methods to estimate the model parameters or a quantity of interest.
3. Construct the confidence intervals for the model parameters or the quantity of interest. (by theory or bootstrapping)
4. Perform the hypothesis testing and draw the statistical conclusion. $H_0: \ \mu = 0$.

Why interval estimation?

Let the data be from $N(0,1)$. We draw the two data sets with size 50, and the histograms are as follows:

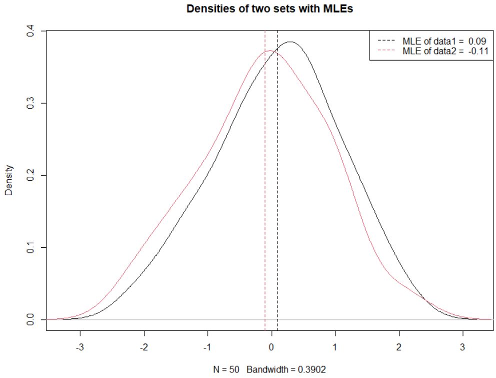


Why interval estimation?

Under $\sigma = 1$, use the two sets to estimate the value of μ. (how to code the likelihood and the estimation?)

$$\hat{\mu}_1 = 0.09, \;\; \hat{\mu}_2 = -0.11.$$

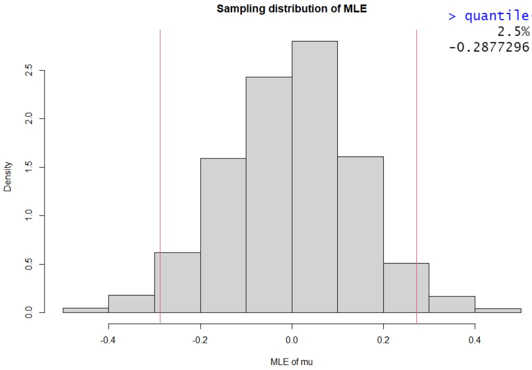
Which is a good guess, 0.09 or -0.11? (the estimates depend on samples.)



Why interval estimation?

Let's draw 50 observations from $N(0, 1)$ 1000 times, and obtain 1000 MLEs of μ. (Note: the true value of μ is 0.)

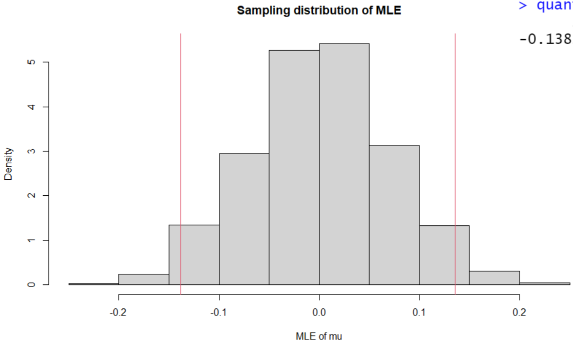
The interval estimation could be defined by an interval constructed by $100 \times (1 - \alpha)\%$ cases of sampling. We say that the true value of μ is located between the interval. $\mu \in \mathrm{CI}_{50} = [-0.29, 0.27]$.




```
> quantile(mu.hat, c(0.025, 0.975))
      2.5%      97.5%
-0.2877296  0.2724782
```

Why interval estimation?

Let's draw 200 observations from $N(0,1)$ 1000 times, and obtain 1000 MLEs of μ. (Note: the true value of μ is 0.)
The interval estimation is $\mu \in \text{CI}_{200} = [-0.14, 0.14]$.
(comparison with $\text{CI}_{50} = [-0.29, 0.27]$.)
Which is better?




```
> quantile(mu.hat, c(0.025, 0.975))
      2.5%      97.5%
-0.1382798  0.1356078
```

The formulation of confidence interval (CI)

- Model assumption:
  Let $X_1, \ldots, X_n \sim F_\theta(\cdot)$, and they are independent and they have the identical distribution $F_\theta(\cdot)$.
- If there exist sample statistics $L_n = g_L(X_1, \ldots, X_n)$ and $U_n = g_U(X_1, \ldots, X_n)$, then (L_n, U_n) is called the $100(1-\alpha)\%$ confidence interval for θ, which satisfying

$$Pro\{L_n < \theta < U_n\} = 1 - \alpha.$$

- A general (approximated) form is expressed as

$$\theta \in [L_n, U_n] = [t - c\sigma_T, t + c\sigma_T],$$

where $\hat{\theta} = h(x_1, \ldots, x_n)$ is a function of data and $\sigma_{\hat{\theta}}$ is the standard error of $\hat{\theta}$.

Simple example (σ^2 is known)

Let

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

where σ^2 is known (ex. $\sigma = 1$).

- (Theory) The estimator of μ is $\bar{X} = \sum_{i=1}^{n} X_i / n$.
- The **sampling distribution** of $\hat{\mu}$ is

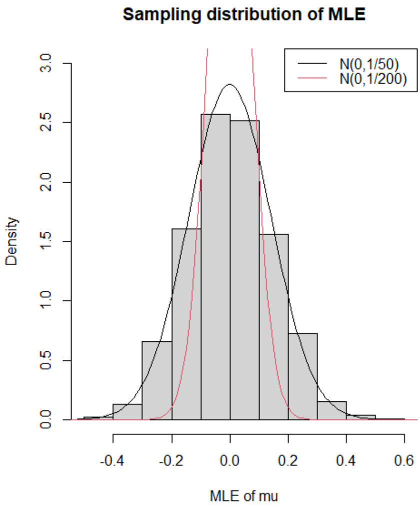
$$\hat{\mu} = \bar{X} \sim N(\mu, \sigma^2/n).$$

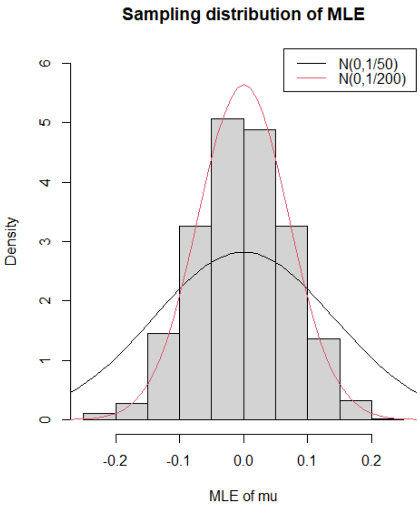
- The CI is constructed as

$$1 - \alpha = Pro\{L_n < \mu < U_n\} = Pro\left\{-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right\},$$

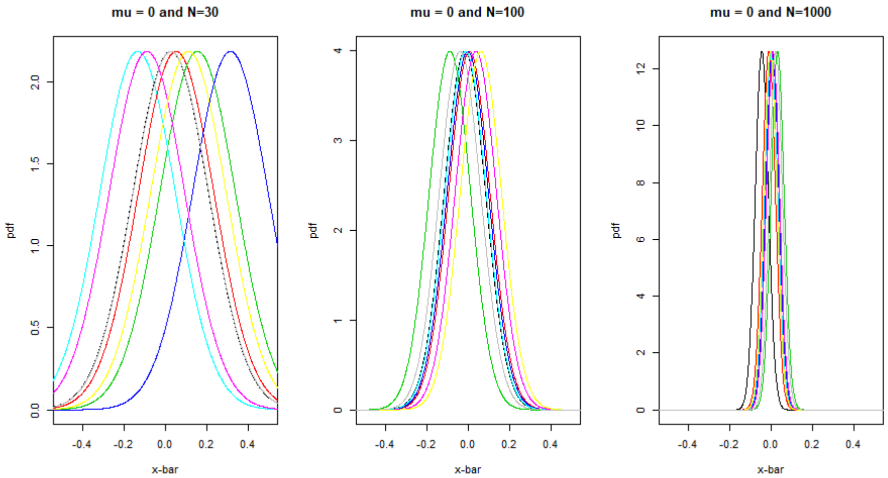
where $L_n = \bar{X} - z_{\alpha/2}\sigma/\sqrt{n}$ and $U_n = \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}$.

Sampling distribution (example)





Effect of sample size on CIs



Simulation code (True value is in the CI)

```
> N <- 30
> include <- rep(NA, 10000)
> for(j in 1:10000){
+   Y <- rnorm(N)
+   ybar <- mean(Y)
+   se <- 1/sqrt(N)
+   Ln <- ybar - qnorm(0.975, 0 ,1) * se
+   Un <- ybar + qnorm(0.975, 0 ,1) * se
+   idx <- sum(c((Ln < 0),(Un > 0)))
+   include[j] <- ifelse(idx == 2, 1, 0)
+ }
> mean(include)
[1] 0.9496
```

However,...

The **exact** sampling distribution of an estimator is not easily derived. Two possible ways are as follows:

- Bootstrapping.
- Need some approximated or asymptotic theory.

Bootstrapping (non-parametric way)

Given a dataset, the procedure of the non-parametric Bootstrapping:

1. Re-sample the dataset with the same length.
2. Calculate the value of the estimator.
3. Repeat Steps 1-2 B times.
4. Draw and find the quantiles of the bootstrapped values of the estimator.

Bootstrapping (parametric way)

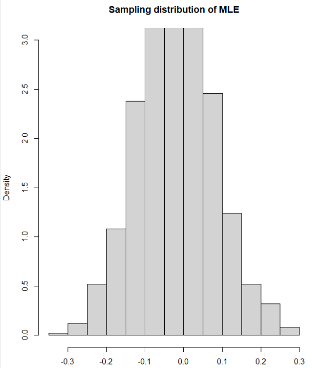
Given a dataset, the procedure of the non-parametric Bootstrapping:

1. Generate the dataset from the "given distribution" with the same length as the collected data.
2. Calculate the value of the estimator.
3. Repeat Steps 1-2 B times.
4. Draw and find the quantiles of the bootstrapped values of the estimator.

Simple example by the non-parametric Bootstrapping method (σ^2 is known)

```r
### Non-parametric Bootstrapping sampling distribution ###
set.seed(1024)
data.collected <- rnorm(100, 0, 1)
est.collected <- optim(0, likelihood.mu, sd = 1, data = data.collected,
                       method = "Brent", lower = -10, upper = 10)
est <- est.collected$par

np.mu.hat.100 <- rep(NA, 1000)
for(j in 1:1000){
  dataY <- sample(data.collected, length(data.collected), replace = TRUE)
  estY <- optim(0, likelihood.mu, sd = 1, data = dataY,
                method = "Brent", lower = -10, upper = 10)
  np.mu.hat.100[j] <- estY$par
}

quantile(np.mu.hat.100, c(0.025, 0.975))
Ln <- est + qnorm(0.025) * sqrt(var(np.mu.hat.100))
Un <- est + qnorm(0.975) * sqrt(var(np.mu.hat.100))
c(Ln, Un)
```

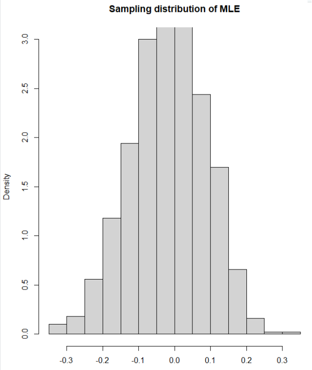
```
> quantile(np.mu.hat.100, c(0.025, 0.975))
      2.5%      97.5%
-0.2088754  0.1878363
> Ln <- est + qnorm(0.025) * sqrt(var(np.mu.hat.100))
> Un <- est + qnorm(0.975) * sqrt(var(np.mu.hat.100))
> c(Ln, Un)
[1] -0.2122330  0.1732376
```



Simple example by the parametric Bootstrapping method (σ^2 is known)

```r
### Parametric Bootstrapping sampling distribution ###

p.mu.hat.100 <- rep(NA, 1000)
for(j in 1:1000){
  dataY <- rnorm(100, est, 1)
  estY <- optim(0, likelihood.mu, sd = 1, data = dataY,
                method = "Brent", lower = -10, upper = 10)
  p.mu.hat.100[j] <- estY$par
}

quantile(p.mu.hat.100, c(0.025, 0.975))
Ln <- est + qnorm(0.025) * sqrt(var(p.mu.hat.100))
Un <- est + qnorm(0.975) * sqrt(var(p.mu.hat.100))
c(Ln, Un)

par(mfrow = c(1, 1))
hist(p.mu.hat.100, probability = TRUE, xlab = "MLE of mu",
     main = "Sampling distribution of MLE", ylim = c(0, 3))
```

```
> quantile(p.mu.hat.100, c(0.025, 0.975))
      2.5%      97.5%
-0.2281404  0.1733398
> Ln <- est + qnorm(0.025) * sqrt(var(p.mu.hat.100))
> Un <- est + qnorm(0.975) * sqrt(var(p.mu.hat.100))
> c(Ln, Un)
[1] -0.2182050  0.1792096
```



Statistical inference

Statistical inference is the process of drawing conclusions about a certain population based on a subset of the data.

1. Assume an underlying model.
2. Use estimation methods to estimate the model parameters or a quantity of interest.
3. Construct the confidence intervals for the model parameters or the quantity of interest.
4. Perform the hypothesis testing and draw the statistical conclusion.

Exercise: σ^2 is unknown

Assume a dataset be random samples from a normal distribution $N(0, 1)$.

- The estimator of μ is $\bar{X} = \sum_{i=1}^{n} X_i/n$.
- The **sampling distribution** of $\hat{\mu}$ is

$$\hat{\mu} = \bar{X} \sim N(\mu, S^2/n),$$

  where S^2 is the sample variance.
- The CI is constructed as

$$L_n = \bar{X} - z_{\alpha/2} S/\sqrt{n},$$

  and

$$U_n = \bar{X} + z_{\alpha/2} S/\sqrt{n}.$$

- Try to use the bootstrapping method for construct the CIs for μ.

Exercise: Sepal width of Iris

Given the sepal width of iris data, use the following steps to show if the mean of sepal width is 3.

1. Assume an underlying model.
2. Use estimation methods to estimate the model parameters or a quantity of interest.
3. Construct the confidence intervals for the model parameters or the quantity of interest. (by theory or bootstrapping)
4. Perform the hypothesis testing and draw the statistical conclusion. $H_0: \ \mu = 3$.

Constructed by the asymptotic theory

The asymptotic theory for the **maximum likelihood estimator** of a model parameter is as follow:

$$\hat{\theta}_{MLE} \sim^d N(\theta, I^{-1}(\theta)) \text{ as } n \to \infty,$$

where

$$I(\theta) = E\left\{\left[\frac{dl(\theta)}{d\theta}\right]^2\right\} = E\left\{-\frac{dl^2(\theta)}{d\theta^2}\right\},$$

$l(\theta)$ is the log-likelihood function of θ and $I(\theta)$ is called Fisher information.

Example: (σ^2 is known)

$$I(\mu) = n/\sigma^2,$$

$$\hat{\mu}_{MLE} = \bar{X} \sim N(\mu, \sigma^2/n).$$

CI: $(L_n, U_n) = \left(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}\right)$.

Constructed by the asymptotic theory

The asymptotic theory for the **maximum likelihood estimators** of parameters is as follow:

$$\hat{\boldsymbol{\theta}}_{MLE} \sim^{d} N(\boldsymbol{\theta}, I^{-1}(\boldsymbol{\theta})) \text{ as } n \to \infty,$$

where

$$I_{ij}(\boldsymbol{\theta}) = E\left\{\left[\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_i}\right]\left[\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_j}\right]\right\} = E\left\{-\frac{\partial l^2(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right\},$$

$l(\boldsymbol{\theta})$ is the log-likelihood function of $\boldsymbol{\theta}$ and $I(\boldsymbol{\theta})$ is called Fisher information matrix.

Constructed by the asymptotic theory

Example: (σ^2 is unknown)

$$I(\mu, \sigma^2) = \begin{bmatrix} n/\sigma^2 & 0 \\ 0 & n/2\sigma^4 \end{bmatrix}$$

$$\left(\hat{\mu}_{MLE}, \hat{\sigma}^2_{MLE}\right)' \sim N((\mu, \sigma^2)', \Sigma),$$

where

$$\Sigma = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{bmatrix}$$

CI: $\mu \in (L_n, U_n) = \left(\hat{\mu}_{MLE} - z_{\alpha/2}\hat{\sigma}_{MLE}/\sqrt{n}, \hat{\mu}_{MLE} + z_{\alpha/2}\hat{\sigma}_{MLE}/\sqrt{n}\right).$

$\sigma^2 \in (L_n, U_n) = \left(\hat{\sigma}^2_{MLE} - z_{\alpha/2}\hat{\sigma}^2_{MLE}\sqrt{2/n}, \hat{\sigma}^2_{MLE} + z_{\alpha/2}\hat{\sigma}^2_{MLE}\sqrt{2/n}\right).$

Using observed information matrix to approximate Fisher information matrix

Let X be the data from the distribution $f_\theta(x)$. Then, the observe information can be used to approximate the Fisher information:

$$I(\theta) = E\left\{-\frac{dl^2(\theta|\boldsymbol{X})}{d\theta^2}\right\} \approx \left\{-\frac{dl^2(\theta|\boldsymbol{x})}{d\theta^2}\right\},$$

where the size of data is large.

Approximated Fisher information matrix

$$I(\theta) = E\left\{-\frac{dl^2(\theta|\boldsymbol{X})}{d\theta^2}\right\} = \int_R -\frac{dl^2(\theta|\boldsymbol{X})}{d\theta^2} f_\theta(\boldsymbol{x})d\boldsymbol{x}.$$

We can use the Monte Carlo integration to approximate the integral.

1. Draw a dataset $(\boldsymbol{x}^{(i)})$ from the distribution $f_\theta(\boldsymbol{x})$.
2. For the dataset, evaluate the value of $-\frac{dl^2(\theta|\boldsymbol{x}^{(i)})}{d\theta^2}$.
3. Repeat Steps 1 and 2 K times, and

$$I(\theta) = E\left\{-\frac{dl^2(\theta|\boldsymbol{X})}{d\theta^2}\right\} = \int_R -\frac{dl^2(\theta|\boldsymbol{X})}{d\theta^2} f_\theta(\boldsymbol{x})d\boldsymbol{x} \approx \frac{1}{K}\sum_{i=1}^{K} -\frac{dl^2(\theta|\boldsymbol{x}^{(i)})}{d\theta^2}.$$