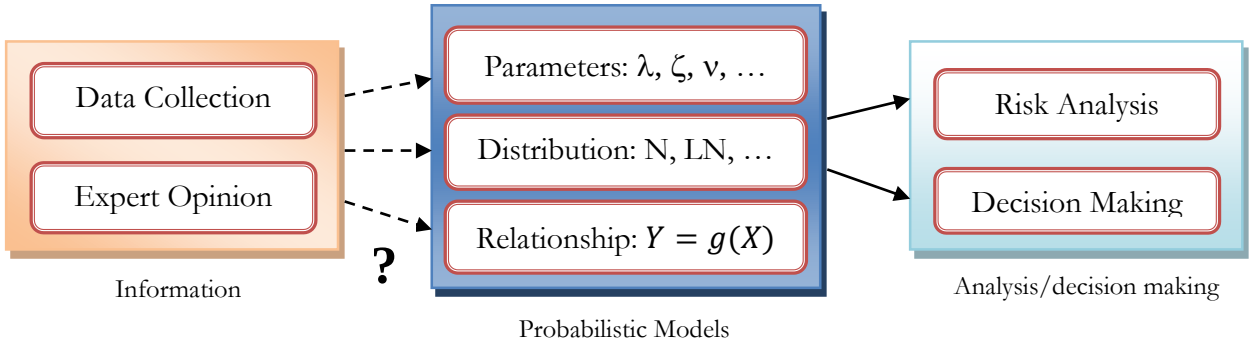
Seoul National University
Dept. of Civil and Environmental Engineering

Instructor: Junho Song
junhosong@snu.ac.kr

457.212 Statistics for Civil & Environmental Engineers
In-Class Material: Class 20
Point Parameter Estimation (1) (A&T: 6.1-6.2)

1. Statistical Inference



2. Point Estimation of Parameters

(a) Sample statistics $\hat{\theta}$: Estimator of true parameter θ

Example: Sample mean $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$ is an estimator of true mean μ, i.e. $\hat{\mu} = \bar{X}$

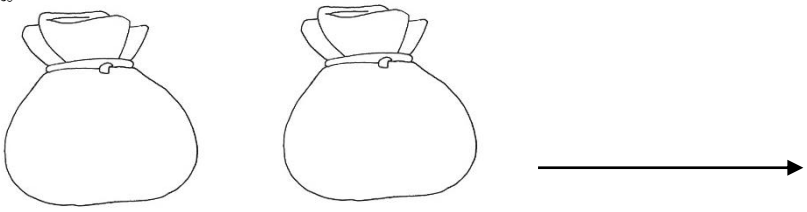
Sample standard deviation $\hat{\sigma} = s$

(b) Desirable properties of a point estimator

(i) **Unbiased**: $E[\hat{\theta}] = \theta$ ~ Average of point estimates is the same as the true parameter.

*Recall: "biased" (1/*N*) and "unbiased" (1/(*N*-1)) sample standard deviation

(ii) Consistent: $\lim_{N\to\infty} \mathrm{Var}[\hat{\theta}] = 0$ ~ As the size of the sample increases, $\hat{\theta}$ converges.



(iii) Efficient: $\mathrm{Var}[\hat{\theta}]$ as small as possible (for the same N).

Instructor: Junho Song
junhosong@snu.ac.kr

Example 1: Is the sample mean $\bar{X}$ an "unbiased" and "consistent" estimator?

3. Point Estimation by "Method of Moments"

Step 1: Find the relationship between the true parameter and moments

$$\theta = g(\mathrm{E}[X], \mathrm{E}[X^2], ...)$$

Step 2: Estimate the moments by

$$\hat{\mathrm{E}}[X^m] = \frac{1}{N}\sum_{i=1}^{N} x_i^m$$

Step 3: Substitute the estimated moments into the relationship in Step 1

$$\hat{\theta} = g(\hat{\mathrm{E}}[X], \hat{\mathrm{E}}[X^2], ...)$$

Example 2 (Modified A&T Example 6.2): Data for the fatigue life of 75 S-T Aluminum available. Its sample mean and variances are

$\bar{X} = 26.75$ million cycles
$s^2 = 360.0$ million cycles

If the fatigue life follows a Gumbel distribution (See Table 6.1 A&T), what are the point estimates on the distribution parameters u and α by M.M.?

Hint: $\mu = u + \frac{0.5772}{\alpha}$, $\sigma^2 = \frac{\pi^2}{6\alpha^2}$

Instructor: Junho Song
junhosong@snu.ac.kr

4. Point Estimation by Method of "Maximum Likelihood Estimation (MLE)"

(a) Known as "best" and "efficient" (i.e. minimum variance of $\hat{\theta}$ for the same sample size)

(b) Finds the values of parameters that _____ the likelihood of the available data set.

(c) Example: The available data set $\{x_1, x_2\} = \{-0.5, 0.5\}$

Suppose we want to estimate the mean of the random variable X, μ_X.

Which of the estimates $\hat{\mu}_X = 0$ and $\hat{\mu}_X = 100$ makes more sense to you? (and why?)

(d) Consider a dataset: $\{x_1, x_2, ..., x_n\}$

Suppose we know the quantity follows a certain type of distribution (e.g. *N*, *LN*) and it requires a distribution parameter θ. Its marginal PDF is denoted as $f_X(x;\theta)$

Event E_1: $X = x_1$ ~ Probability $P(E_1) \propto f_X(x_1;\theta)$
Event E_2: $X = x_2$ ~ Probability $P(E_2) \propto f_X(x_2;\theta)$
…
Event E_n: $X = x_n$ ~ Probability $P(E_n) \propto f_X(x_n;\theta)$

Probability that we will get the particular dataset as an outcome?

$$P\left(\bigcap_{i=1}^{n} E_i\right) \propto \prod_{i=1}^{n} f_X(x_i;\theta) = L(x_1, ..., x_n;\theta)$$

$L(x_1, ..., x_n;\theta)$: "_____" function ~ a function of the parameter θ that is proportional to the probability that we observe the given dataset.

Want to find the value of θ that _____ the likelihood function.

(e) Point estimation by Method of Maximum Likelihood (MLE)

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(x_1, x_2, .., x_n;\theta)$$

Obtain the value of θ that maximizes the probability that the given data set would be observed.

(f) How? Solve

$$\frac{\partial L(x_1, ..., x_n;\theta)}{\partial \theta} =$$

Instructor: Junho Song
junhosong@snu.ac.kr

Example 3: Based on the given data set $\{x_1, x_2, ..., x_n\}$, estimate the parameter of Exponential distribution by MLE.

(Hint: PDF $f_X(x;\nu) = \nu \exp(-\nu x)$)

(g) It is much more convenient to find a value that maximize the natural logarithm of the likelihood function, "log-likelihood function" $\ln L(x_1, x_2, ..., x_n; \theta)$. The solution of the following equation is the same as the one in (f).

$$\frac{\partial \ln L(x_1, ..., x_n; \theta)}{\partial \theta} = 0$$

Why the same result?

Why more convenient? (1) Derivative of product vs. summation
(2) Exponential functions
(3) Products of terms in each PDF

Example 4 (Contd.): Estimate the parameter by MLE using the log-likelihood function

Instructor: Junho Song
junhosong@snu.ac.kr

(h) Multiple distribution parameters (e.g. (λ, ζ) for LN, (μ, σ) for N)

Solve the system equation

$$\frac{\partial \ln L(x_1, ..., x_n; \theta_1)}{\partial \theta_1} = 0, \ \ldots, \ \frac{\partial \ln L(x_1, ..., x_n; \theta_m)}{\partial \theta_m} = 0$$

Example 5: Given $\{x_1, x_2, ..., x_n\}$

Find the point estimates on λ and ζ by MLE.

Example 6: Show that MLE estimates of μ and σ^2 of a normal distribution are $\frac{1}{n}\sum_{i=1}^{n} x_i$ and $\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{MLE})^2$, respectively.

Example 7: Tossing an unfair coin n times and observed "HEAD" x times.

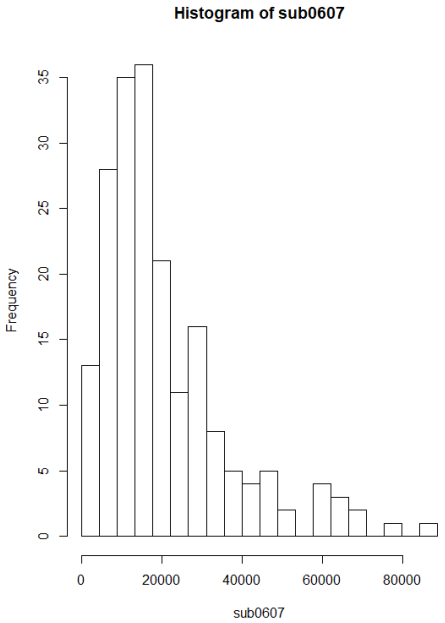
Find the MLE estimate on the probability of "HEAD" each time.

Instructor: Junho Song
junhosong@snu.ac.kr

(i) **[R Example]** MLE of multiple distribution parameters using R – Lognormal distribution

Download the dataset 'subway.txt' available from the eTL website

The column "use0607" represents the number of subway passengers from each DONG of Seoul from 6am to 7am.



* Image source: 서울연구데이터서비스(The Seoul Research Data Service, http://data.si.re.kr/node/103)

* Data source: 서울열린데이터광장(http://data.seoul.go.kr/dataList/datasetView.do?infId=OA-12252&srvType=S&serviceKind=1)

```
library(stats4) # package which contains MLE: comes with base R
subway = read.table('subway.txt', header=TRUE)
sub0607 = subway$use0607 # number of subway passengers from each DONG
in Seoul (6-7am)
hist(sub0607, breaks=seq(0,max(sub0607),max(sub0607)/20))

# User-defined function to compute log-likelihood function
LL = function(meanlog, sdlog) {
  Probs = suppressWarnings(dlnorm(sub0607, meanlog, sdlog))
  # ignore warning message during optimization: not important
  -sum(log(Probs)) # Log-Likelihood function
}

MLE_result = mle(LL, start=list(meanlog=10, sdlog=10))
# parameter estimation by MLE; start: setting initial value of the
search
lambda = unname(MLE_result@coef['meanlog'])
# unname : delete name of a list
zeta = unname(MLE_result@coef['sdlog'])
mean_MLE = exp(lambda+0.5*zeta^2) # obtain mean and sd by MLE
sd_MLE = mean_MLE*sqrt(exp(zeta^2)-1)
mean_MLE; sd_MLE
mean(sub0607); sd(sub0607) # sample mean and std devition for
comparison
```

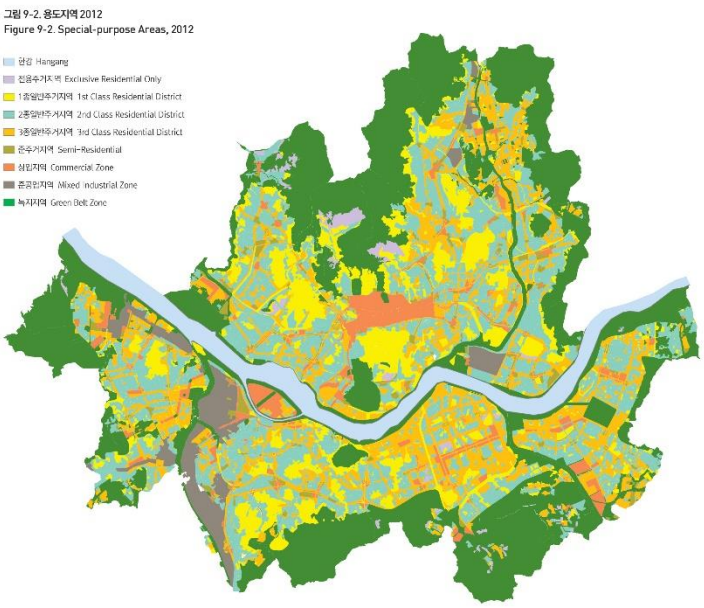
If we assume the number of subway passengers follows a Lognormal distribution, what are the MLE estimates of λ and ζ? (Corresponding) MLE estimates on the mean and standard deviation?

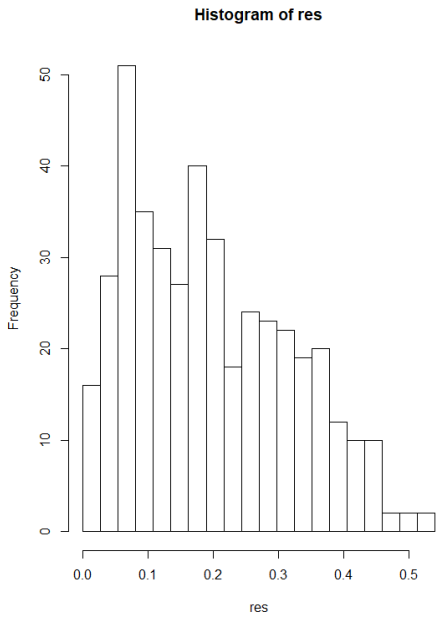
Instructor: Junho Song
junhosong@snu.ac.kr

[R Example 2] MLE of multiple distribution parameters using R – Gamma distribution

Download the dataset 'residentialarea.txt' available from the eTL website

The column "`ResidentialRate`" represents the percentage of residential areas for each DONG of Seoul.





* Image source: 서울정책아카이브(Seoul Solution, https://seoulsolution.kr/ko/seoul-map)

* Data source: 환경공간정보서비스(https://egis.me.go.kr/main.do)

```
resid = read.table('residentialarea.txt', header=TRUE)
res = resid$ResidentialRate
# Percentage of residentia area for each DONG
hist(res, breaks=seq(0,max(res),max(res)/20))

LL = function(shape, rate) {
  Probs = suppressWarnings(dgamma(res, shape, rate))
  -sum(log(Probs))
}

MLE_result = mle(LL, start=list(shape=1, rate=1))
k = unname(MLE_result@coef['shape'])
nu = unname(MLE_result@coef['rate'])
mean_MLE = k/nu
sd_MLE = sqrt(k/nu^2)
mean_MLE; sd_MLE
mean(res); sd(res)
```

If we assume the residential area rate follows a Gamma distribution, what are the MLE estimates of k and ν? (Corresponding) MLE estimates on the mean and standard deviation?

- The development of the R examples above were made possible by the help of the Transportation Engineering Laboratory (http://trlab.kr/) at Seoul National University (Faculty advisor: Prof. Dong-Kyu Kim).