Chapter 1. Introduction to R, and Descriptive Data Analysis

What is *R*: an environment for data analysis and graphics based on *S* language

- a full-featured programming language
- freely available to everyone (with complete source code)
- Easier access to the means of handling BigData such as parallel computation, Hadoop, distributed computation.
- official homepage: `http://www.R-project.org`

1.1 Installation

Installing *R*: R consists of two major parts: the base system and a collection of (over 8.5K) user contributed add-on packages, all available from the above website.

To install the base system, Windows users may follow the link

```
http://CRAN.R-project.org/bin/windows/base/release.htm
```

Note. The base distribution comes with some high-priority add-on packages such as graphic systems, linear models etc.

After the installation, one may start R in the PC by going to `Start -> Statistics -> R`, or simply doubt-click the logo 'R' on your desktop. An R-console will pops up with a prompt character like '>'.

R may be used as a calculators. Of course it can do much more. Try out

```
> sqrt(9)/3 -1
```

To quit R, type at the prompt ‘q( )’.

It is strongly advised to use RStudio instead of R. You may find it with the link

https://www.rstudio.com/

From Wikipedia: RStudio is a free and open-source integrated development environment (IDE) for R, a programming language for statistical computing and graphics. RStudio was founded by JJ Allaire, creator of the programming language ColdFusion. Hadley Wickham is the Chief Scientist at RStudio.

We assume you use RStudio throughout the course.

To define a vector x consisting of integers $1, 2, \cdots, 100$

```
> x <- 1:100
> x
  [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
 [19]  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
 [37]  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
 [55]  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
 [73]  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
 [91]  91  92  93  94  95  96  97  98  99 100
> sum(x)
> [1] 5050
```

Or we may also try

```
> y <- (1:100)^2
> y
  [1]    1    4    9   16   25   36   49   64   81  100  121  144
 [13]  169  196  225  256  289  324  361  400  441  484  529  576
 [25]  625  676  729  784  841  900  961 1024 1089 1156 1225 1296
 [37] 1369 1444 1521 1600 1681 1764 1849 1936 2025 2116 2209 2304
 [49] 2401 2500 2601 2704 2809 2916 3025 3136 3249 3364 3481 3600
 [61]  3721  3844  3969  4096  4225  4356  4489  4624  4761  4900  5041  5184
 [73]  5329  5476  5625  5776  5929  6084  6241  6400  6561  6724  6889  7056
 [85]  7225  7396  7569  7744  7921  8100  8281  8464  8649  8836  9025  9216
 [97]  9409  9604  9801 10000
> y[14]     # print out the 14-th element of vector y
[1] 196
```

One may also try `x+y`, `(x+y)/(x+y)`, `help(log)`, `log(x)` etc.

Additional packages can be installed directly from the R prompt. Information on the available packages is available at

`http://cran.r-project.org/web/views/`
`http://cran.r-project.org/web/packages/`

For example, one may install HSAUR2 – *A Handbook of Statistical Analysis Using R (2nd edition)*:

```
> install.packages("HSAUR2")
> library("HSAUR2")  # To load all the objects in the package\\
                     #  into the current session
```

You may start an R help manual using command `help.start()`. By clicking `Packages` in the manual, you will see `HSAUR2` is listed among the installed packages.

1.2 Help and documentation

To start a manual page of R: `help.start()`

Alternatively we may access online manual at

```
http://cran.r-project.org/manuals.html
```

To access a manual for function ‘mean’: `help(mean)`, or `?mean`

To access the info on an added-on package: `help(package="HSAUR2")`

To access the info on a data set or a function in the installed package: `help(package="HSAUR2", men1500m)`

To load all the functions in an added-on package: `library("HSAUR2")`

To load a data set from the installed package into the current session: `data(men1500m, package="HSAUR2")`

Type `men1500m` to print out all the info in the data set ‘men1500m’.

Two other useful sites:

R Newsletter: `http://cran.r-project.org/doc/Rnews/`

R FAQ: `http://cran.r-project.org/faqs.html`

You may also simply follow the links on the main page of the R project http://www.R-project.org

Last but not least, `google` whatever questions often leads to most helpful answers

1.3 Data Import/Export

The easiest form of data to import into R is a simple text file. The primary function to import from a text file is `scan`. You may check out what ‘scan’ can do: `> ?scan`

Create a plain text file ‘simpleData’, in the folder ‘statsI’ in your Drive D, as follow:

```
This is a simple data file, created for illustration
of importing data in text files into R
1 2 3 4
5 6 7 8
9 10 11 12
```

It has two lines of explanation and 3 lines numbers. The R session below imports it into R as a vector `x` and 3×4 matrix `y`, perform some simple operations. Note the flag `skip=2` instructs R to ignore the first two lines in the file.

Note. R ignores anything after '#' in a command line.

```
> x <- scan("D:/statsI/simpleData.txt", skip=2)
> x                               # print out vector x
 [1]  1  2  3  4  5  6  7  8  9 10 11 12
> length(x)
[1] 12
> mean(x); range(x) # write 2 commands in one line to save space
[1] 6.5
[1]  1 12
> summary(x)                 # a very useful command!
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00    3.75    6.50    6.50    9.25   12.00 
```

```
> y <- matrix(scan("D:/statsI/simpleData.txt", skip=2), byrow=T,
            ncol=4)
> y                           # print out matrix y
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    5    6    7    8
[3,]    9   10   11   12
> dim(y)           # size of matrix y
[1] 3 4
> y[1,]           # 1st row of y
[1] 1 2 3 4
> y[,2]           # 2nd column of y
[1]  2  6 10
> y[2,4]          # the (2,4)-th element of matrix y
[1] 8
```

A business school sent a questionnaire to its graduates in past 5 years and received 253 returns. The data are stored in a plain text file 'Jobs' which has 6 columns:

C1: ID number
C2: Job type, 1 - accounting, 2 - finance, 3 - management, 4 - marketing and sales, 5 -others
C3: Sex, 1 - male, 2 - female
C4: Job satisfaction, 1 - very satisfied, 2 - satisfied, 3 - not satisfied
C5: Salary (in thousand pounds)
C6: No. of jobs after graduation

```
IDNo.  JobType  Sex    Satisfaction  Salary     Search
1        1       1       3              51      1
2        4       1       3              38      2
3        5       1       3              51      4
4        1       2       2              52      5
... ...
```

We import data into R using command `read.table`

```
> jobs <- read.table("D:/statsI/Jobs.txt"); jobs
         V1      V2  V3           V4     V5     V6
1      IDNo. JobType Sex Satisfaction Salary Search
2          1       1   1            3     51      1
3          2       4   1            3     38      2
4          3       5   1            3     51      4
   ... ...
> dim(jobs)
[1] 254   6
> jobs[1,]
     V1      V2  V3           V4     V5     V6
1 IDNo. JobType Sex Satisfaction Salary Search
```

We repeat the above again by taking the 1st row as the names of variables (`header=T`) and the entries in 1st column as the names of the rows (`row.names =1`).

```
> jobs <- read.table("D:/statsI/Jobs.txt", header=T, row.names=1)
> dim(jobs)
[1] 253   5
> names(jobs)
[1] "JobType"  "Sex"   "Satisfaction"   "Salary"  "Search"
> class(jobs)
[1] "data.frame"
> class(jobs[,1]); class(jobs[,2]); class(jobs[,3]);
   class(jobs[,4]); class(jobs[,5])
[1] "integer"
[1] "integer"
[1] "integer"
[1] "integer"
[1] "integer"
```

Since the first three variables are nominal, we may specify them as ‘factor’, while "Salary" can be specified as ‘numeric’:

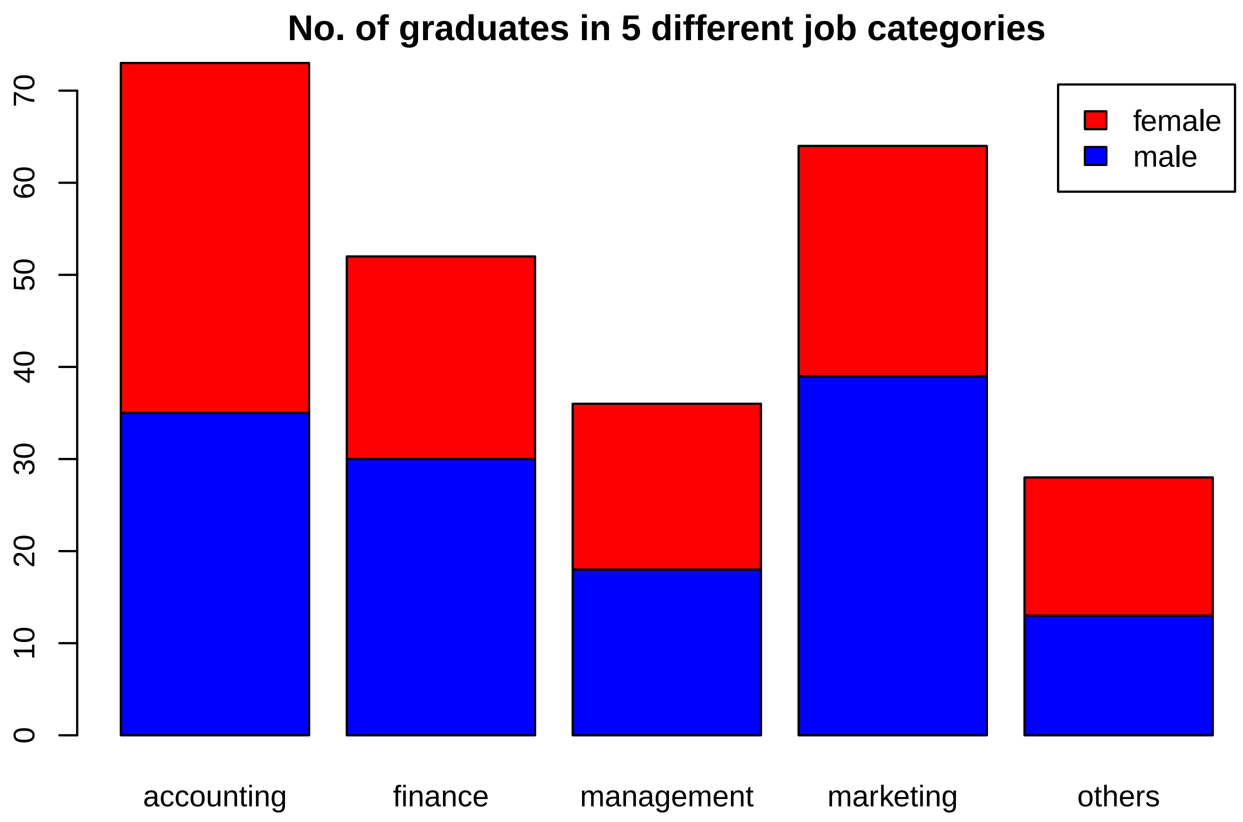
```
>  jobs <- read.table("D:/statsI/Jobs.txt", header=T, row.names=1,
            colClasses = c("factor", "factor", "factor",
            "numeric", "integer"))
> class(jobs[,1]); class(jobs[,2]); class(jobs[,3]);
       class(jobs[,4]); class(jobs[,5])
[1] "factor"
[1] "factor"
[1] "factor"
[1] "numeric"
[1] "integer"
```

Note. we need to specify the class for the row name variable (i.e. 1st column) as well.

Now we do some simple descriptive statistical analysis for this data.

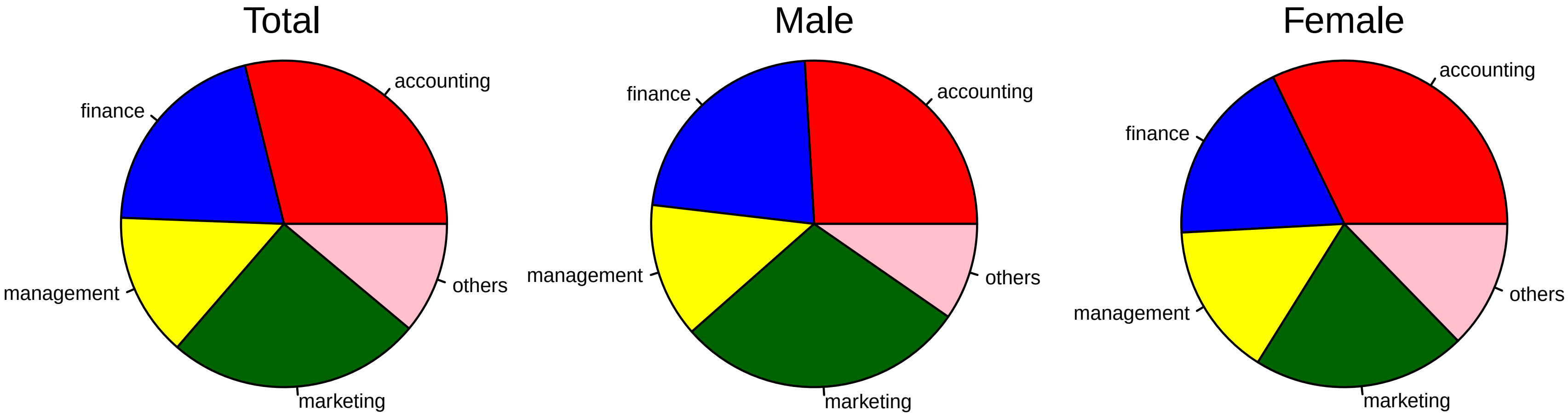
```
> table(jobs[,1])
 1  2  3  4  5
```

```
73 52 36 64 28   # No. of graduates with 5 different JobTypes
> t <-table(jobs[,2], jobs[,1], deparse.level=2)  # store table in t
> t
         jobs[, 1]
jobs[, 2]  1  2  3  4  5
        1 35 30 18 39 13           # No. of males with 5 different JobTypes
        2 38 22 18 25 15           # No. of females with 5 different JobTypes
> 100*t[1,]/sum(t[1,])
       1         2         3         4          5
25.92593  22.22222  13.33333  28.88889   9.62963
                        # Percentages of males with 5 different JobTypes
> 100*t[2,]/sum(t[2,])
       1         2         3         4          5
32.20339  18.64407  15.25424  21.18644  12.71186
                        # Percentages of females with 5 different JobTypes
> barplot(t, main="No. of graduates in 5 different job categories",
          legend.text=c("male", "female"), names.arg=c("accounting",
          "finance", "management", "marketing", "others"))  # draw a bar-plot
```


No. of graduates in 5 different job categories
female
male
0
10
20
30
40
50
60
70
accounting
finance
management
marketing
others

The barplot shows the difference in job distribution due to gender. We may also draw pie-plots, which are regarded as less effective.

```
> pie(t[1,]+t[2,],label=c("accounting","finance","management",
      "marketing","others")); text(0,1, "Total", cex=2)
> pie(t[1,],label=c("accounting","finance","management",
                "marketing","others")); text(0,1, "Male", cex=2)
> pie(t[2,],label=c("accounting","finance","management",
                "marketing","others")); text(0,1, "Female", cex=2)
```

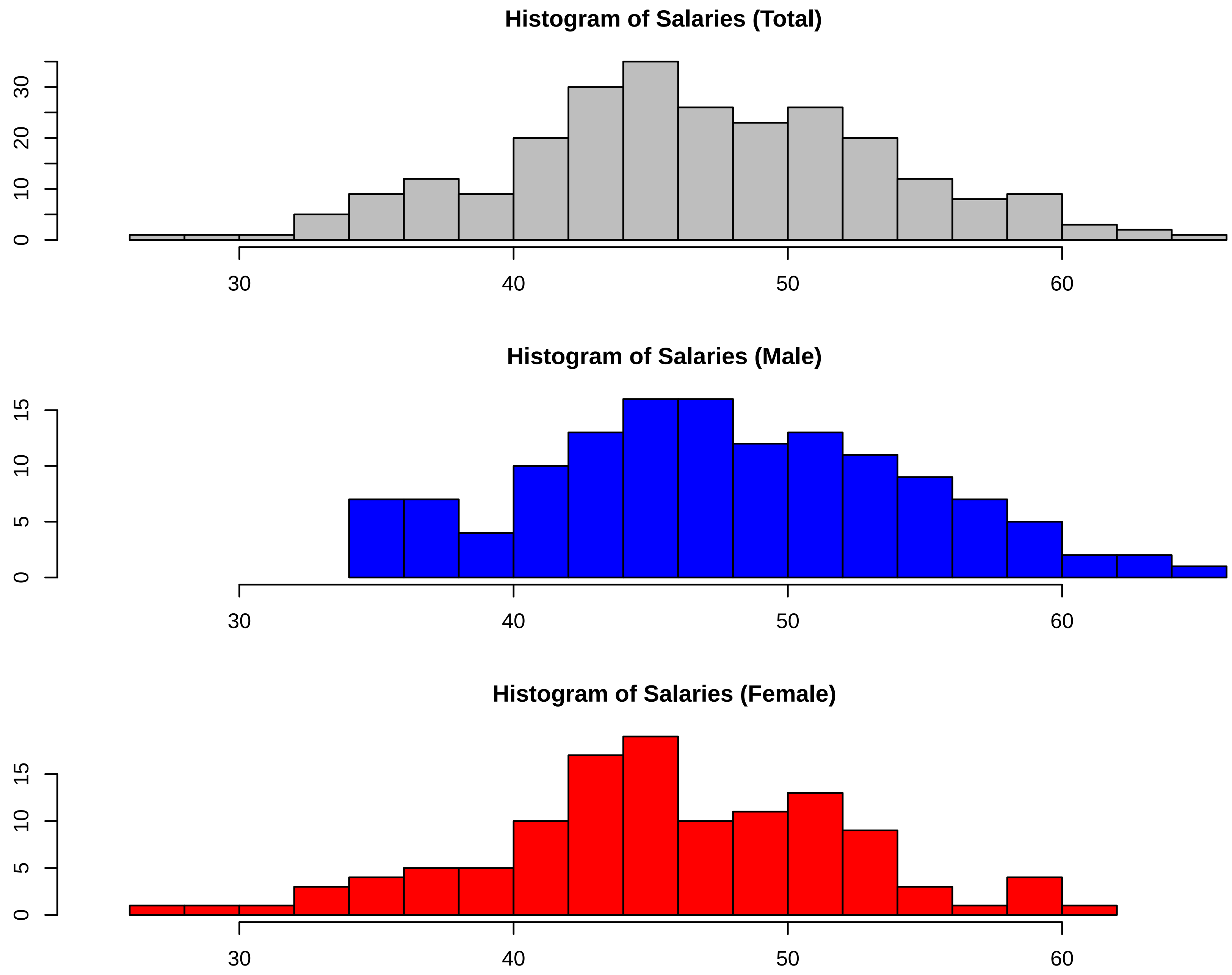

Total
accounting
finance
management
marketing
others
Male
accounting
finance
management
marketing
others
Female
accounting
finance
management
marketing
others

Now let look at the salary (`jobs[,4]`) distribution, and the impact due to gender.

```
> mSalary <- jobs[,4][jobs[,2]==1]
              # extract the salary data from male
> fSalary <- jobs[,4][jobs[,2]==2]
              # extract the salary data from female
> summary(jobs[,4]); summary(mSalary); summary(fSalary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  26.00   43.00   47.00   47.13   52.00   65.00
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  34.00   44.00   48.00   48.11   53.00   65.00
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  26.00   42.25   46.00   46.00   51.00   61.00
> hist(jobs[,4], col="gray", nclass=15, xlim=c(25,66),
              main="Histogram of Salaries (Total)")
      # plot the histogram of salary data
> hist(mSalary, col="blue", nclass=15, xlim=c(25,66),
              main="Histogram of Salaries (Male)")
```

```
> hist(fSalary, col="red", nclass=15, xlim=c(25,66),
                main="Histogram of Salaries (Female)")
```

You may also try stem-and-leaf plot: `stem(jobs[,4])`


Histogram of Salaries (Total)

Histogram of Salaries (Male)

Histogram of Salaries (Female)

To export data from R, use `write.table` or `write`.

To write `jobs` into a plain text file ‘Jobs1.txt’:

```
> write.table(jobs, "Jobs1.txt")
```

which retains both the row and column names. Note the different entries in the file are separated by spaces.

We may also use

```
> write.table(jobs, "Jobs2.txt", row.names=F, col.names=F),
> write.table(jobs, "Jobs3.txt", sep=",")
```

Compare the three output files.

Note that the values of factor variables are recorded with “ ". To record all the levels of factor variables as numerical values, we need to define a pure numerical data.frame first:

```
> t <- data.frame(as.numeric(jobs[,1]), as.numeric(jobs[,2]),
         as.numeric(jobs[,3]), jobs[,4], jobs[,5])
> write.table(t, "Jobs4.txt")
```

The file "Jobs4.txt" contains purely numerical values.

Note. (i) Working directory — all exported files are saved in ‘My Documents’ by default. You may change your working directory by clicking

```
    File -> Change dir...
```

in the RGui window. For example, I create on my laptop `D:\statsI` as my working directory for this course.

(ii) Saving a session — when you quit an R session `q()`, you will be offered an option to 'save workspace image'. By clicking on "yes", you will save all the objects (including data sets, loaded functions from added-on packages etc) in your R session. You may continue to work on this session by directly double-clicking on the image file in your working directory.

A useful tip: Create a separate working directory for each of your R projects.

1.4 Organising an Analysis

An R analysis typically consists of executing several commands. Instead of typing each of those commands on the R prompt, we may collect them into a plain text file. For example, the file "jobsAnalysis.r" in my working directory reads like:

```
jobs <- read.table("Jobs.txt", header=T, row.names=1)
        # File "Jobs.txt" is in the working directory now
mSalary <- jobs[,4][jobs[,2]==1]
fSalary <- jobs[,4][jobs[,2]==2]
summary(jobs[,4])
summary(mSalary)
summary(fSalary)
par(mfrow=c(3,1))   # display 3 figures in one column
hist(jobs[,4], col="gray", nclass=15, xlim=c(25,66),
         main="Histogram of Salaries (Total)")
hist(mSalary, col="blue", nclass=15, xlim=c(25,66),
        main="Histogram of Salaries (Male)")
hist(fSalary, col="red", nclass=15, xlim=c(25,66),
        main="Histogram of Salaries (Female)")
```

You may carry out the project by sourcing the file into an R session:

```
> source("jobAnalysis.r", echo=T)
```

Also try `source("jobAnalysis.r")`.

1.5 Writing functions in R

For some repeated task, it is convenient to define a function in R. We illustrate this idea by an example.

Consider the famous ‘Birthday Coincidences’ problem: *In a class of k students, what is the probability that at least two students have the same birthday?*

Let us make some assumptions to simplify the problems:

(i) only 365 days in every year,

(ii) every day is equally likely to be a birthday,

(iii) students' birthdays are independent with each other.

With k people, the total possibilities is $(365)^k$.

Consider the complementary event: all k birthdays are different. The total such possibility is

$$365 \times 364 \times 363 \times \cdots \times (365 - k + 1) = \frac{365!}{(365 - k)!}$$

So the probability that at least two students have the same birthday is

$$p(k) = 1 - \frac{365!}{(365 - k)!(365)^k}.$$

We may use R to compute $p(k)$. Unfortunately factorials are often too large, e.g. $52! = 8.065525e + 67$, and often cause overflow in computer. We adopt the alternative formula

$$p(k) = 1 - \exp\{\log(365!) - \log((365 - k)!) - k \log(365)\}.$$

We define a R-function pBirthday to perform this calculation for different k.

```
> pBirthday <- function(k)
+ 1 - exp(lfactorial(365) - lfactorial(365-k) - k*log(365))
                 # lfactorial(n) returns log(n!)
> pBirthday(100)
[1] 0.9999997   # probability with a class of 100 students
> x <- c(20, 30, 40, 50, 60)
> pBirthday(x)
[1] 0.4114384 0.7063162 0.8912318 0.9703736 0.9941227
```

With 20 students in class, the probability of having overlapping birthdays is about 0.41. But with 60 students, the probability is almost 1, i.e. *it is almost always true that at least 2 out of 60 students have the same birthday.*

Note. The expression in a function may have several lines. In this case the expression is enclosed in curly braces { }, and the final line determines the return value.

Another Example — **The capture and recapture problem**

To estimate the number of whitefish in a lake, 50 whitefish are caught, tagged and returned to the lake. Some time later another 50 are caught and only 3 are tagged ones. Find a reasonable estimate for the number of whitefish in the lake.

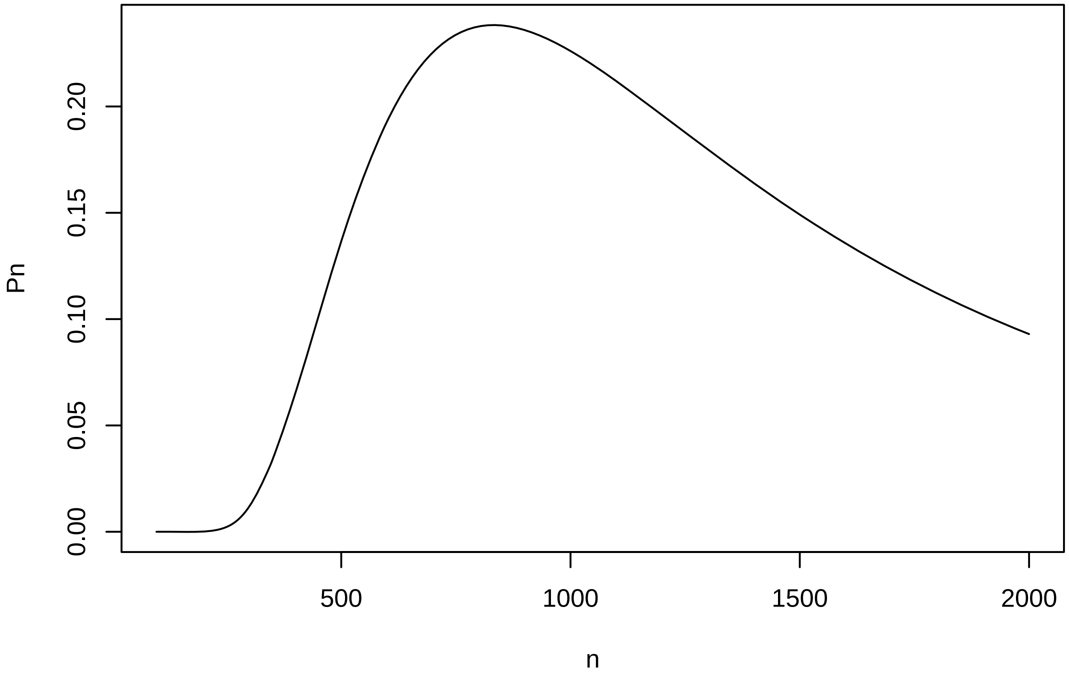
Suppose there are n whitefish in the lake. Catching 50 fish can be done in $\left(\begin{array}{c} n \\ 50 \end{array} \right) = \frac{n!}{50!(n-50)!}$ ways, while catching 3 tagged ones and 47 untagged can be done in $\left(\begin{array}{c} 50 \\ 3 \end{array} \right) \left(\begin{array}{c} n-50 \\ 47 \end{array} \right)$ ways. Therefore the probability for the latter event to occur is

$$P_n = \left(\begin{array}{c} 50 \\ 3 \end{array} \right) \left(\begin{array}{c} n-50 \\ 47 \end{array} \right) \Big/ \left(\begin{array}{c} n \\ 50 \end{array} \right).$$

Therefore, a reasonable estimate for n should be the value at which P_n obtains its maximum. We use R to compute P_n and to find the estimate.

```
> Pn <- function(n) {
+      tmp <- choose(50,3)*choose(n-50,47)
+      tmp/choose(n,50)
+ }            # Definition for function Pn ends here
> n <- 97:2000    # as there are at least 97 fish in the lake
> plot(n, Pn(n), type='l')
```

It produces the plot of P_n against n:



To find the maximum:

```
> m <- max(Pn(n)); m
[1] 0.2382917
> n[Pn(n)==m]
[1] 833
```

Hence the estimated number of fish in the lake is 833.

1.6 Control structure: loops and conditionals

An `if` statement has the form

```
if (condition) expression1 else expression2
```

It executes ‘expression1’ if ‘condition’ is true, and ‘expression2’ otherwise. When ‘condition’ contains several lines, they should be enclosed in curly braces { }. The same applies to expressions.

The above statement can be compactly written in the form

```
ifelse(condition, expression1, expression2)
```

When the else-part is not present:

```
if (condition) expression
```

It executes ‘expression’ if ‘condition’ is true, and does nothing otherwise.

A `for` loop allows a statement to be iterated as a variables assumes values in a specified sequence. It has the form:

```
for(variable in sequence) statement
```

A `while` loop does not use an explicit loop variable:

```
while (condition) expression
```

It repeats 'expression' as long as 'condition' holds. This makes it differently from the "if-statement" above.

We illustrate those control commands by examining a simple 'doubling' strategy in gambling.

You go to a casino to play a simple 0-1 game: you bet x dollars and flip a coin. You win $2x$ dollars and keep your bet if 'Head', and lose x dollars if 'Tail'. You start 1 dollar in first game, and double your bet in each new games, i.e. you bet 2^{i-1} dollars in the i-th game, $i = 1, 2, \cdots$.

With this strategy, once you win, say, at the $(k+1)$-th game, you will recover all your losses in your previous games plus a profit of $2^k + 1$ dollars, as

$$2 \times 2^k > \sum_{i=1}^{k} 2^{i-1} = 2^k - 1.$$

Hence as long as (i) the probability p of the occurrence of 'Head' is positive (no matter how small), and (ii) you have enough capital to keep you in the games, you may win handsomely at the end — is it really true?

Condition (ii) is not trivial, as the maximum loss in 20 games is $2^{20} - 1 = 1,048,575$ dollars!

Plan A: Suppose you could afford to lose maximum n games and, therefore, decide to play n games. We define the R-function `nGames` below to simulate your final earning/loss (after n games).

```
nGames <- function(n,p) {
      # n is the No. of games to play
      # p is the prob of winning each game
x <- 0 # earning after each game
for(i in 1:n)  ifelse(runif(1)<p, x <- x+2^i, x <- x-2^(i-1))
      # runif(1) returns a random number from uniform dist on (0, 1)
x     # print out your final earning/loss
}
```

To play $n = 20$ games with $p = 0.1$:

```
> nGames(20, 0.1)
[1] -999411
> nGames(20, 0.1)
[1] -1048575
> nGames(20, 0.1)
[1] 524289
> nGames(20, 0.1)
```

```
[1] -655263
> nGames(20, 0.1)
[1] -1016895
```

We repeated the experience 5 times above, with 5 different results.

One way to assess this gameplan is to repeat a large number of times and look at the average earning/loss:

```
> x = vector(length=5000)
> for(i in 1:5000) x[i] <- nGames(20, 0.1)
> mean(x)
[1] -733915
```

In fact, this mean -733915 is stable measure reflecting the average loss of this gameplan.

Plan B: Play the maximum *n*, but quit as soon as winning one game. The *R*-function `winStop` simulates the earning/loss.

```
winStop <- function(n,p) {
      # n -- maximum No. of games, p -- prob of winning each game
i <- 1
ifelse((runif(1)<p), x<- 2, x<- -1)  # play 1st game
while((x<0)&(i<n)){ i <- i+1      # i records the no. of games played
                 ifelse(runif(1)<p, x <- x+2^i, x <- x-2^(i-1))
               }
x
}
```

Set $n = 20$, $p = 0.1$, we repeat the experience a few times:

```
>winStop(20, 0.1)
[1] 2
```

```
> winStop(20, 0.1)
[1] 17
> winStop(20, 0.1)
[1] 129
> winStop(20, 0.1)
[1] -1048575
> winStop(20, 0.1)
[1] 16385
```

To assess the gameplan:

```
> x<- 1:5000
> for(i in 1:5000) x[i] <- winStop(20, 0.1)
> mean(x)
[1] -112672.9   # This indicates "Plan B" is better than "Plan A"
> for(i in 1:5000) x[i] <- winStop(80, 0.1)
          # the maximum no. of games is 80 now
> mean(x)
```

```
[1] -7.22886e+20
> for(i in 1:5000) x[i] <- winStop(90, 0.1)
          # the maximum no. of games is 90 now
> mean(x)
3.790896e+18
```

With p as small as 0.1, you need a huge capital in order to play about 90 games to generate the positive returns in average.

The best and the most effective way to learn R: use it!

Hands-on experience is the most illuminating.

Chapter 2. Probability

Probability: a number between 0 and 1 to quantifying uncertainty in a mathematical manner.

2.1 Sample space and events

Sample Space Ω: a set of possible outcomes of an experiment.

Sample outcome, realization or **element**: a point in a sample space, denoted by $\omega \in \Omega$.

Event or **random event**: a subset of Ω, i.e. an assemble of some sample outcomes

Example 1. Experiment – Toss a coin two times.

Sample space = $\{HH, HT, TH, TT\}$.

$A \equiv \{HH, HT\} = \{\text{1st toss is head}\}$ is an event.

What is the sample space if we toss a coin for ever? — *the Bernoulli trial*.

Background of tossing a coin: success or failure, up or down, better or worse, boy or girl, 1 or 0 and etc.

Example 2. Find the sample space in each of the following cases

- number of insect damaged leaves on a plant
- lifetime (in hours) of a light bulb
- weight of a 10-hour old infant
- exchange rate of pounds sterling to US-dollars today next year
- directional movement S&P500 index price tomorrow

Complement of event A: $A^c = \{\omega \in \Omega : \omega \notin A\}$. Obviously $\Omega^c = \emptyset$ (the empty set).

Union of events A and B: $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$. Then $A \cup B = B \cup A$, $A \cup A^c = \Omega$.

Intersection of events A and B: $A \cap B \equiv AB = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$. Then $A \cap B = B \cap A$, $A \cap A^c = \emptyset$.

If $A_1, A_2, \cdots$ is a sequence of events,

$$\bigcup_{i=1}^{\infty} A_i = \{\omega \in \Omega : \omega \in A_i \text{ for at least one } i\},$$

$$\bigcap_{i=1}^{\infty} A_i = \{\omega \in \Omega : \omega \in A_i \text{ for all } i\}.$$

Difference of events A and B: $A - B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \notin B\}$. Obviously $A - B \neq B - A$.

Inclusion: occurrence of event A implies that of B, we say $A \subset B$.

Summary of Terminology

Ω	Sample space, true event (always true)
$\varnothing$	null event (always false)
ω	outcome, realization or element
A^c	complement of A (not A)
$A \cup B$	union (A or B)
$A \cap B$ or AB	intersection (A and B)
$A - B$ or $A \backslash B$	set difference
$A \subset B$	set inclusion

Mutually exclusive or **disjoint**: A and B are mutually exclusive if $A \cap B = \varnothing$. Obviously A and A^c are mutually exclusive.

Partition of Ω: a sequence $\underline{\text{disjoint}}$ events $A_1, A_2, \cdots$ such that

$$\bigcup_{i=1}^{\infty} A_i = \Omega.$$

Indicator of A: $I_A \equiv I_A(\omega)$ — a function defined on $\omega \in \Omega$:

$$I_A = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{otherwise,} \end{cases} \quad \text{or equivalently} \quad I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases}$$

Limits of a sequence of monotonic events:
(i) A sequence $A_1, A_2, \cdots$ is monotone increasing if $A_1 \subset A_2 \subset \cdots$. We define $\lim_{n\to\infty} A_n = \cup_{i=1}^{\infty} A_i$.
(ii) A sequence $A_1, A_2, \cdots$ is monotone decreasing if $A_1 \supset A_2 \supset \cdots$. We define $\lim_{n\to\infty} A_n = \cap_{i=1}^{\infty} A_i$.
In both cases, we may write $A_n \to A$, where A denotes its limit.

Example 3. Let $\Omega = (-\infty, \infty)$, $A_i = [0, 1/i)$. Then

$$\cup_{i=1}^{\infty} A_i = [0, 1), \qquad \cap_{i=1}^{\infty} A_i = \{0\}.$$

If we change to $A_i = (0, 1/i)$, then $\cup_{i=1}^{\infty} A_i = (0, 1)$ and $\cap_{i=1}^{\infty} A_i = \emptyset$.

For $A_i = (-i, i)$, $\cup_{i=1}^{\infty} A_i = \Omega$.

Example 4. Let Ω be the salaries earned by the graduates from a Business School. We may choose $\Omega = [0, \infty)$. Based on the dataset "Jobs.txt", we extract some *interesting* events/subsets. Recall the info of the dataset:

C1: ID number

C2: Job type, 1 - accounting, 2 - finance, 3 - management, 4 - marketing and sales, 5 -others

C3: Sex, 1 - male, 2 - female

C4: Job satisfaction, 1 - very satisfied, 2 - satisfied, 3 - not satisfied

C5: Salary (in thousand pounds)

C6: No. of jobs after graduation

We have defined salaries of male and female graduates respectively as follows:

```
> jobs <- read.table("Jobs.txt", header=T, row.names=1)
> mSalary <- jobs[,4][jobs[,2]==1]
> fSalary <- jobs[,4][jobs[,2]==2]
```

Similarly we may extract the salaries from finance sector or accounting:

```
> finSalary <- jobs[,4][jobs[,1]==2]; summary(finSalary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  37.00   46.00   53.00   52.08   58.00   65.00 
> accSalary <- jobs[,4][jobs[,1]==1]; summary(accSalary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  40.00   47.00   51.00   50.45   54.00   62.00 
```

According to this dataset, finance pays slightly higher than accounting. We may also extract the salaries for males (females) in accounting:

```
> maccSalary <- jobs[,4][(jobs[,1]==1) & (jobs[,2]==1)]
     # &' stands for logic operation and'
> summary(mfinSalary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  44.00   48.00   51.00   51.31   55.00   62.00
> faccSalary<- jobs[,4][(jobs[,1]==1) & (jobs[,2]==2)]
> summary(ffinSalary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  40.00   45.25   49.50   49.66   53.00   61.00
```

To extract the salaries for males in both finance and accounting:

```
>  mfinaccSalary <- jobs[,4][ (jobs[,2]==1) & ( (jobs[,1]==1) |
     (jobs[,1]==2) ) ]     # |' stands for logic operation or'
```

To remove (unwanted) objects:

```
> rm(mSalary, fSalary, accSalary, finSalary, maccSalary,
           faccSalary, mfinaccSalary)
```

2.2 Probability

Definition. Probability a function P that assigns a real number P(A) to each event in a sample space, which satisfies the three conditions:

(i) $P(A) \geq 0$ for any event A,
(ii) $P(\Omega) = 1$, and
(iii) For disjoint events $A_1, A_2, \cdots,$ $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$.

Let $A_1 = \Omega,\ A_2 = A_3 = \cdots = \varnothing$. By (iii) and (ii), $P(\varnothing) = 0$.
Hence for any disjoint A and B, $P(A \cup B) = P(A) + P(B)$.

More properties of probability:

1. $P(A^c) = 1 - P(A)$.
2. If $A \subset B$, $P(B) = P(A) + P(B - A) \geq P(A)$.

3. $P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B)$.
4. Boole inequality: $P(\bigcup_{i=1}^n A_i) \leq \sum_{i=1}^n P(A_i)$.
5. If $A_n \to A$, $P(A_n) \to P(A)$.

Proof. 1. $P(A) + P(A^c) = P(A \cup A^c) = P(\Omega) = 1$.

3. $A \cup B = (AB) \cup (AB^c) \cup (A^cB)$, and the 3 events on the RHS are disjoint. Hence

$$P(A \cup B) = P(AB) + P(AB^c) + P(A^cB).$$

Since $A = (AB) \cup (AB^c)$, $P(A) = P(AB) + P(AB^c)$. Similarly $P(B) = P(AB) + P(A^cB)$. Therefore
$P(A \cup B) = P(AB) + \{P(A) - P(AB)\} + \{P(B) - P(AB)\} = P(A) + P(B) - P(AB)$.

4. is obtained by applying 3. repeatedly.

The proof of 5. is a bit more involved, we refer to p.7 of Wasserman (2004).

Example 5. Toss a fair 6-sided die, there are 6 possible outcomes each with probability 1/6. If we toss it two times, the sample space is $\Omega = \{(i, j) : i, j = 1, \cdots, 6\}$. Since each outcome is equally likely,

$$P(A) = \frac{\text{No. of elements in } A}{36}, \qquad A \in \Omega.$$

For example, $P(A) = 2/36 = 1/18$ for $A = \{\text{the sum is 3}\}$, and $P(A) = 3/36 = 1/12$ for $A = \{\text{the sum is 4}\}$.

2.3 Independence

Definition. k events $A_1, \cdots, A_k$ are independent if

$$P(A_{i_1} A_{i_2} \cdots A_{i_j}) = P(A_{i_1}) P(A_{i_2}) \cdots P(A_{i_j})$$

for any $1 \leq i_1 < i_2 < \cdots < i_j \leq k$ and $2 \leq j \leq k$.

Intuition. If A and B are independent, the occurrence of A has nothing to do with the occurrence of B. For example, two persons toss two coins: two outcomes are independent with each other.

Example 6. Toss a fair coin 10 times. Let A = “at least one head". Define T_j be the event that tail occurs on the j-th toss. Then

$$\begin{aligned} P(A) &= 1 - P(A^c) = 1 - P(T_1 \cdots T_{10}) = 1 - P(T_1)P(T_2) \cdots P(T_{10}) \\ &= 1 - (0.5)^{10} \approx 0.9999. \end{aligned}$$

Example 7. John and Peter play each other in the final of a tennis tournament. Whoever wins 2 out of 3 games will win the tournament. Suppose that John is higher ranked player who beats Peter in a single game with probability 0.6, and each game will be played independently. Find the probability that John will win the tournament.

Let A_i =“John wins the i-th game", and A =“John wins the tournament". Then

$$A = (A_1 A_2) \cup (A_1 A_2^c A_3) \cup (A_1^c A_2 A_3),$$

and the 3 events on the RHS are disjoint. Hence

$$\begin{aligned} P(A) &= P(A_1A_2) + P(A_1A_2^cA_3) + P(A_1^cA_2A_3) \\ &= (0.6)^2 + 2 \times (0.6)^2 \times 0.4 = 0.648, \end{aligned}$$

which is greater than the probability for John to win a single game.

Question. Would John prefer to play the maximum 5 (instead of 3) games in the final?

2.4 Conditional Probability

Example 8. Five people take one ball each out of a bag containing 4 white balls and one red ball.

Obviously the Probability for the 1st person to take the red ball is 1/5. What is the probability for the 2nd, 3rd, 4th or the last person to take the red ball?

Definition. If $P(B) > 0$, the conditional probability of A given B is

$$P(A|B) = P(AB)/P(B),$$

which is the probability of event A given the condition that event B occurs already.

Remark. (i) If A and B are independent, $P(A|B) = P(A)$.

(ii) In general $P(AB) = P(A|B)P(B)$.

Example 8. (Continue)

$$\begin{aligned} &P(\text{2nd person takes R}) = P(\text{1st person takes W, 2nd Person take R}) \\ &= P(\text{1st person takes W}) \times P(\text{2nd person takes R}|\text{1st person takes W}) \\ &= \frac{4}{5} \times \frac{1}{4} = 1/5, \end{aligned}$$

which is the same as the probability for the 1st person to take the red.

Let $A_1, \cdots, A_k$ be a partition of Ω.

Law of Total Probability. For any event B,

$$P(B) = P(BA_1) + \cdots + P(BA_k).$$

Proof. $B = B\Omega = B(\cup_i A_i) = \cup_i (BA_i)$. Since $BA_1, \cdots, BA_k$ are disjoint, the law holds.

Bayes' Formula. Let $P(B) > 0$ and $P(A_i) > 0$ for $i = 1, \cdots, k$. Then

$$P(A_j|B) = P(B|A_j)P(A_j) \Big/ \sum_{i=1}^{k} P(B|A_i)P(A_i).$$

Proof. $P(A_j|B) = P(A_jB)\big/P(B) = P(B|A_j)P(A_j)\big/P(B)$. Replacing $P(B)$ using the law of total probability, we obtain Bayes' Formula.

Example 9. Larry divides his emails into 3 categories: A_1 ="spam", A_2 ="low priority" and A_3 ="high priority". From previous experience he concludes

$$P(A_1) = 0.7, \quad P(A_2) = 0.2, \quad P(A_3) = 0.1.$$

Let B be the event that an email contains the word 'free'. Again based on previous experience,

$$P(B|A_1) = 0.9, \quad P(B|A_2) = 0.1, \quad P(B|A_3) = 0.1.$$

He receives a new email with word 'free'. What is the probability that it is spam?

By Bayes' theorem,

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{\sum_{i=1}^{3} P(B|A_i)P(A_i)} = 0.955.$$

Chapter 3. Random Variables and Distributions

Basic idea of introducing random variables: represent outcomes and/or random events by numbers.

3.1 Random variables and Distributions.

Definition. A random variable is a function defined on the sample space Ω, which assigns a real number $X(\omega)$ to each outcome $\omega \in \Omega$.

Example 1. Flip a coin 10 times. We may define random variables (r.v.s) as follows:

X_1 = no. of heads,
X_2 = no. of flips required to have the first head,

X_3 = no. of ‘HT’-pairs,
X_4 = no. of tails.

For $\omega = HHTHHHTHHTT$, $X_1(\omega) = 6,\ X_2(\omega) = 1,\ X_3(\omega) = 3,\ X_4(\omega) = 4$. Note $X_1 \equiv 10 - X_4$.

Remark. The values of a r.v. varies and cannot be pre-determined before an outcome occurs.

Definition. For any r.v. X, its (cumulative) distribution function (CDF) is defined as $F_X(x) = P(X \leq x)$.

Example 2. Toss a fair coin twice and let X be the number of heads. Then

$$P(X = 0) = P(X = 2) = 1/4, \quad P(X = 1) = 1/2.$$

Hence its CDF is $F_X(x) = \begin{cases} 0 & x < 0, \\ 1/4 & 0 \leq x < 1, \\ 3/4 & 1 \leq x < 2, \\ 1 & x \geq 2. \end{cases}$

Note. (i) $F_X(x)$ is right continuous, non-decreasing, and defined for all $x \in (-\infty, \infty)$. For example, $F_X(1.1) = 0.75$.

(ii) The CDF is a non-random function.

(iii) If $F(\cdot)$ is the CDF of r.v. X, we simply write $X \sim F$.

Properties of CDF. A function $F(\cdot)$ is a CDF iff

(i) F is non-decreasing: $x_1 < x_2$ implies $F(x_1) \leq F(x_2)$,
(ii) F is normalized: $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$,
(iii) F is right continuous: $\lim_{y\downarrow x} F(y) = F(x)$.

Probabilities from CDF

(a) $P(X > x) = 1 - F(x)$

(b) $P(x < X \leq y) = F(y) - F(x)$

(c) $P(X < x) = \lim_{h \downarrow 0} F(x - h) \equiv F(x-)$

(d) $P(X = x) = F(x) - F(x-)$.

Note. It is helpful for understanding (c) & (b) to revisit Example 2.

3.2 Discrete random variables

If r.v. X only takes some isolated values, X is called a discrete r.v. Its CDF is called a discrete distribution.

Definition. For a discrete r.v. X taking values $\{x_1, x_2, \cdots\}$, we define the probability function (or probability mass function) as

$$f_X(x_i) = P(X = x_i), \quad i = 1, 2, \cdots .$$

Obviously, $f_X(x_i) \geq 0$ and $\sum_i f_X(x_i) = 1$.

It is often more convenient to list a probability function in a table:

X	x_1	x_2	$\cdots\cdots$
Probability	$f_X(x_1)$	$f_X(x_2)$	$\cdots\cdots$

Example 2 (continue). The probability function is be tabulated:

X	0	1	2
Probability	1/4	1/2	1/4

Expectation or Mean EX or $E(X)$: a measure for the 'center', 'average value' of a r.v. X, and is often denoted by μ.

For a discrete r.v. X with probability function $f_X(x)$,

$$\mu = EX = \sum_i x_i f_X(x_i).$$

Variance Var(X): a measure for variation, uncertainty or 'risk' of a r.v. X, is often denoted by σ^2, while σ is called **standard deviation** of X.

For a discrete r.v. X with probability function $f_X(x)$,

$$\sigma^2 = \mathrm{Var}(X) = \sum_i (x_i - \mu)^2 f_X(x_i) = \sum_i x_i^2 f_X(x_i) - \mu^2.$$

The k-th moment of X: $\mu_k \equiv E(X^k) = \sum_i x_i^k f_X(x_i)$, $k = 1, 2, \cdots$.

Obviously, $\mu = \mu_1$, and $\sigma^2 = \mu_2 - \mu_1^2$.

Some important discrete distributions

Convention. We often use upper case letters $X, Y, Z, \cdots$ to denote r.v.s, and lower case letters $x, y, z, \cdots$ to denote the values of r.v.s. In contrast letters a, b or A, B are often used to denote (non-random) constants.

Degenerate distribution: $X \equiv a$, i.e. $F_X(x) = 1$ for any $x \geq a$, and 0 otherwise.

It is easy to see that $\mu = a$ and $\sigma^2 = 0$.

Bernoulli distribution: X is binary, $P(X = 1) = p$ and $P(X = 0) = 1 - p$, where $p \in [0, 1]$ is a constant. It represents the outcome of flipping a coin.

$$\mu = 1 \cdot p + 0 \cdot (1 - p) = p, \quad \sigma^2 = p(1 - p).$$

Note. Bernoulli trial refers to an experiment of flipping a coin repeatedly.

Binomial distribution Bin(n, p): X takes values 0, 1, $\cdots$, n only with the probability function

$$f_X(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \quad x = 0, 1, \cdots, n.$$

Theorem. If we toss a coin n times, let X be the number of heads. Then $X \sim$ Bin(n, p), where p is the probability that head occurs in tossing the coin once.

Proof. Let $\omega = HTHHT \cdots H$ denote an outcome of n tosses. Then $X = k$ iff there are k 'H' and $(n-k)$ 'T' in ω. Therefore the probability of such a ω is $p^k(1-p)^{n-k}$. Since those k H's may occur in any n positions of the sequence, there are $\left(\begin{array}{c} n \\ k \end{array} \right)$ such ω's. Hence

$$P(X = k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}, \quad k = 0, 1, \cdots, n.$$

Let us check if the probability function above is well defined. Obviously $P(X = k) \geq 0$, furthermore

$$\sum_{k=0}^{n} P(X = k) = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} = \{p + (1-p)\}^n = 1^n = 1.$$

Let us work out the mean and the variance for $X \sim \mathrm{Bin}(n, 1-p)$.

$$\begin{aligned} \mu &= \sum_{k=0}^{n} k \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = \sum_{k=1}^{n} k \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\ &= \sum_{j=0}^{n-1} np \frac{(n-1)!}{j!(n-1-j)!} p^j (1-p)^{n-1-j} \\ &= np \sum_{j=0}^{m} \frac{m!}{j!(m-j)!} p^j (1-p)^{m-j} = np. \end{aligned}$$

Note that $\sigma^2 = E(X^2) - \mu^2 = E\{X(X-1)\} + \mu - \mu^2$. We need to work out

$$E\{X(X-1)\} = \sum_{k=0}^{n} k(k-1)\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

$$= \sum_{k=2}^{n} n(n-1)p^2 \frac{(n-2)!}{(k-2)!\{(n-2)-(k-2)\}}p^{k-2}(1-p)^{\{(n-2)-(k-2)\}}$$

$$= n(n-1)p^2 \sum_{j=0}^{n-2} \frac{(n-2)!}{j!\{(n-2)-j)\}}p^j(1-p)^{\{(n-2)-j\}} = n(n-1)p^2.$$

This gives $\sigma^2 = n(n-1)p^2 + np - (np)^2 = np(1-p)$.

By the above theorem, we can see immediately

(i) If $X \sim \mathrm{Bin}(n,p)$, $n - X \sim \mathrm{Bin}(n, 1-p)$.
(ii) If $X \sim \mathrm{Bin}(n,p)$, $Y \sim \mathrm{Bin}(m,p)$, and X and Y are independent, then $X + Y \sim \mathrm{Bin}(n+m,p)$.

Furthermore, let $Y_i = 1$ if the i-th toss yields H, and 0 otherwise. Then $Y_1, \cdots, Y_n$ are *independent* Bernoulli r.v.s with mean p and variance $p(1-p)$. Since $X = Y_1 + \cdots + Y_n$, we notice

$$EX = \sum_{i=1}^n EY_i = np, \quad \text{Var}(X) = \sum_{i=1}^n \text{Var}(Y_i) = np(1-p).$$

This is a much easier way to derived the means and variances for binomial distributions, which is based on the following general properties.

(i) For any r.v.s $\xi_1, \cdots, \xi_n$, and any constants $a_1, \cdots, a_n$,

$$E\Big(\sum_{i=1}^n a_i \xi_i\Big) = \sum_{i=1}^n a_i E(\xi_i).$$

(ii) If, in addition, $\xi_1, \cdots, \xi_n$ are *independent*,

$$\mathrm{Var}\Big(\sum_{i=1}^{n} a_i\xi_i\Big) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(\xi_i).$$

Independence of random variables. The r.v.s $\xi_1, \cdots, \xi_n$ are independent if

$$P(\xi_1 \le x_1, \cdots, \xi_n \le x_n) = P(\xi_1 \le x_1) \times \cdots \times P(\xi_n \le x_n)$$

for any $x_1, \cdots, x_n$.

Moment generate function (MGF) of r.v. X:

$$\psi_X(t) = E(e^{tX}), \quad t \in (-\infty, \infty).$$

(i) It is easy to see that $\psi'_X(0) = E(X) = \mu$. In general $\psi_X^{(k)}(0) = E(X^k) = \mu_k$.

(ii) If $Y = a + bX$, $\psi_Y(t) = E(e^{(a+bX)t}) = e^{at}\psi_X(bt)$.

(iii) If $X_1, \cdots, X_n$ are independent, $\psi_{\sum_i X_i}(t) = \prod_{i=1}^n \psi_{X_i}(t)$, and vice versa

If X is discrete, $\psi_X(t) = \sum_i e^{x_i t} f_X(x_i)$.

To generate a r.v. from Bin(n, p), we can flip a coin (with p-probability for H) n times, and count the number of heads. However R can do the flipping for us much more efficiently:

```
> rbinom(10, 100, 0.1)  # generate 10 random numbers from \Bin(100, 0.1)
 [1]  8 11  9  9 18  7  5  5  3  7
> rbinom(10, 100, 0.1)  # do it again, obtain different numbers
 [1] 11 13  6  7 11  9  9  9 12 10
> x <- rbinom(10, 100, 0.7); x; mean(x)
 [1] 66 77 67 66 64 68 70 68 72 72
[1] 69                # mean close to np=70
> x <- rbinom(10, 100, 0.7); x; mean(x)
 [1] 70 73 72 70 68 69 70 66 79 71
[1] 70.8
```

Note that `rbinom(10000, 1, 0.5)` is equivalent to toss a fair coin 10000 times:

```
> y <- rbinom(10000, 1, 0.5); length(y); table(y)
[1] 10000
y
   0    1
4990 5010   # about a half times with head
```

You may try with smaller sample size, such as

```
> y <- rbinom(10, 1, 0.5); length(y); table(y)
[1] 10
y
0 1
3 7  # 7 heads and 3 tails
```

Also try out `pbinom` (CDF), `dbinom` (probability function), `qbinom` (quantile) for Binomial distributions.

Geometric Distribution Geom(p): X takes all positive integer values with probability function

$$P(X = k) = (1 - p)^{k-1} p, \qquad k = 1, 2, \cdots .$$

Obviously, X is the number of tosses required in a Bernoulli trial to obtain the first head.

$$\mu = \sum_{k=1}^{\infty} k(1 - p)^{k-1} p = -p \frac{d}{dp} \sum_{k=1}^{\infty} (1 - p)^{k} = -p \frac{d}{dp} (1/p) = 1/p,$$

and it can be shown that $\sigma^2 = (1 - p)/p^2$.

Using the MGF provides an alternative way to find mean and variance: for

$t < -\log(1-p)$ (i.e. $e^t(1-p) < 1$),

$$\begin{aligned}\psi_X(t) &= E(e^{tX}) = \sum_{i=1}^{\infty} e^{ti}(1-p)^{i-1}p = \frac{p}{1-p}\sum_{i=1}^{\infty}\{e^t(1-p)\}^i \\ &= \frac{p}{1-p}\frac{e^t(1-p)}{1-e^t(1-p)} = \frac{pe^t}{1-e^t(1-p)} = \frac{p}{e^{-t}-1+p}.\end{aligned}$$

Now $\mu = \psi'_X(0) = \left[\frac{pe^{-t}}{(e^{-t}-1+p)^2}\right]_{t=0} = 1/p$, and

$$\mu_2 = \psi''_X(0) = \left[\frac{2pe^{-2t}}{(e^{-t}-1+p)^3} - \frac{pe^{-t}}{(e^{-t}-1+p)^2}\right]_{t=0} = 2/p^2 - 1/p.$$

Hence $\sigma^2 = \mu_2 - \mu^2 = (1-p)/p^2$.

The R functions for Geom(p): `rgeom`, `dgeom`, `pgeom` and `qgeom`.

Poisson Distribution Poisson(λ): X takes all non-negative integers with probability function

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k = 0, 1, 2, \cdots ,$$

where $\lambda > 0$ is a constant, called parameter.

The MGF $X \sim$ Poisson(λ):

$$\psi_X(t) = \sum_{k=0}^{\infty} e^{kt} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^t\lambda)^k}{k!} = e^{-\lambda} e^{e^t\lambda} = \exp\{\lambda(e^t - 1)\}.$$

Hence

$$\mu = \psi'_X(0) = [\exp\{\lambda(e^t - 1)\}\lambda e^t]_{t=0} = \lambda,$$

$$\mu_2 = \psi''_X(0) = [\exp\{\lambda(e^t - 1)\}\lambda e^t + \exp\{\lambda(e^t - 1)\}(\lambda e^t)^2]_{t=0} = \lambda + \lambda^2.$$

Therefore $\sigma^2 = \mu_2 - \mu^2 = \lambda$.

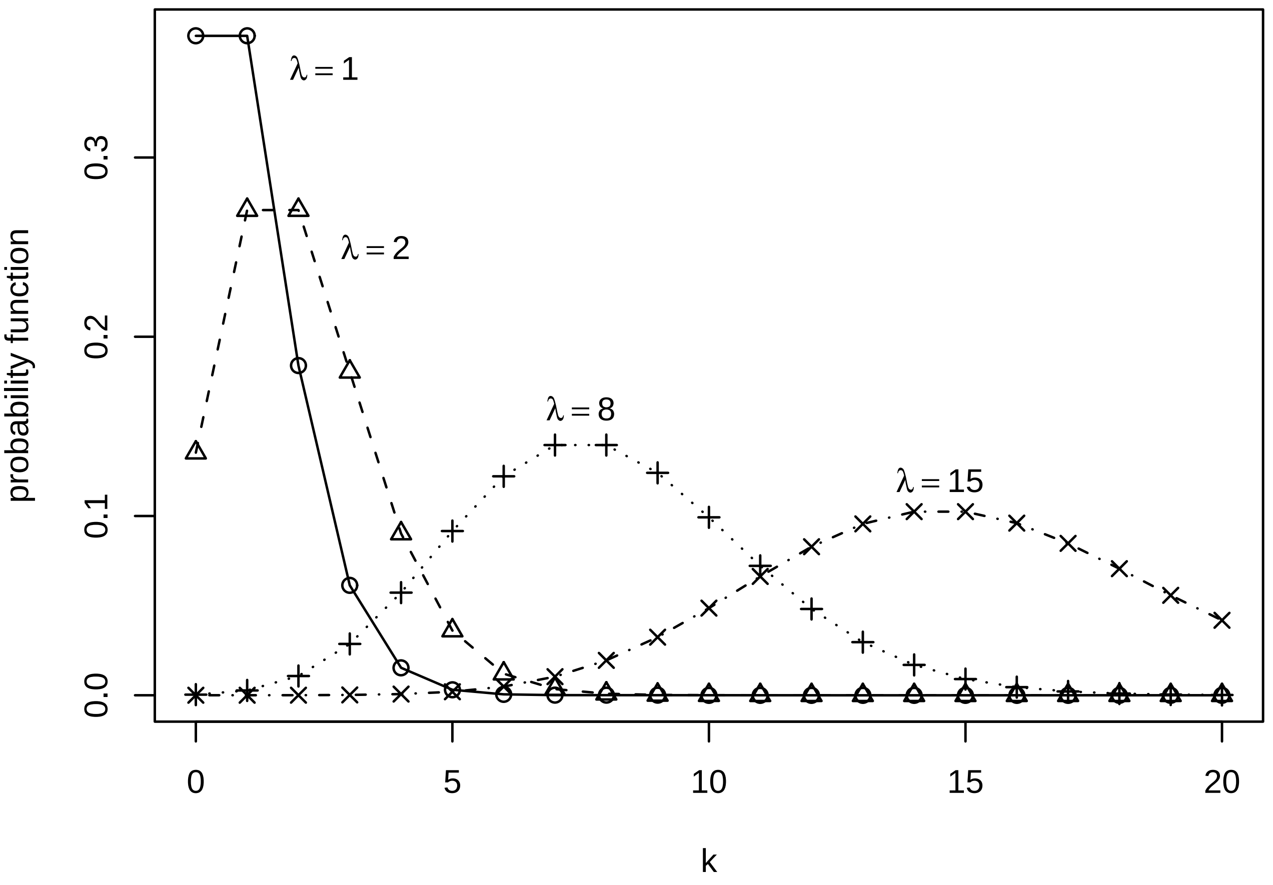
Remark. For Poisson distributions, $\mu = \sigma^2$.

The R functions for Poisson(λ): `rpois, dpois, ppois` and `qpois`.

To understand the role of the parameter λ, we plot the probability function of Poisson(λ) for different values of λ.

```
> x <- c(0:20)
> plot(x,dpois(x,1),type='o',xlab='k', ylab='probability function')
> text(2.5,0.35, expression(lambda==1))
> lines(x,dpois(x,2),typ='o',lty=2, pch=2)
> text(3.5,0.25, expression(lambda==2))
> lines(x,dpois(x,8),typ='o',lty=3, pch=3)
> text(7.5,0.16, expression(lambda==8))
> lines(x,dpois(x,15),typ='o',lty=4, pch=4)
> text(14.5, 0.12, expression(lambda==15))
```

Plots of $\lambda^k e^{-\lambda}/k!$ against k



Three ways of computing probability and distribution functions:

- calculators — for simple calculation
- statistical tables — for, e.g. the final exam
- R — for serious tasks such as real application

3.2 Continuous random variables

A r.v. X is *continuous* if there exists a function $f_X(\cdot) \geq 0$ such that

$$P(a < X < b) = \int_a^b f_X(x)dx, \quad \forall a < b.$$

We $f_X(\cdot)$ the *probability density function* (PDF) or, simply, density function. Obviously

$$F_X(x) = \int_{-\infty}^x f_X(u)du.$$

Properties of continuous random variables

(i) $F_X(x) = P(X \leq x) = P(X < x)$, i.e. $P(X = x) = 0 \neq f_X(x)$.

(ii) The PDF $f_X(\cdot) \geq 0$, and $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

(iii) $\mu = E(X) = \int_{-\infty}^{\infty} x f_X(x)dx$,

$$\sigma^2 = \mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)dx = \int x^2 f_X(x)dx - \mu^2.$$

Furthermore the MGF of X is equal to

$$\psi_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx.$$

Some important continuous distributions

Uniform distribution $U(a,b)$: X takes any values between a and b equally likely. Its PDF is

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases}$$

Then the CDF is

$$F(x) = \int_{-\infty}^{x} f(u)du = \begin{cases} 0 & x < a, \\ \frac{1}{b-a}\int_a^x du = \frac{x-a}{b-a} & a \le x \le b, \\ 1 & x > b. \end{cases},$$

and

$$\mu = \int_a^b \frac{x dx}{b-a} = \frac{a+b}{2}, \quad \mu_2 = \int_a^b \frac{x^2 dx}{b-a} = \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3}$$

Hence $\sigma^2 = \mu_2 - \mu^2 = (b-a)^2/12$.

R-functions related to uniform distributions: `runif`, `dunif`, `punif`, `qunif`.

Quantile. For a given CDF $F(\cdot)$, its quantile function is defined as

$$F^{-1}(p) = \inf\{x : F(x) \ge p\}, \quad p \in [0, 1]$$

```r
> x <- c(1, 2.5, 4)
> punif(x, 2, 3)     # CDF of U(2, 3) at 1, 2.5 and 4
[1] 0 0.5 1
> dunif(x, 2, 3)     # PDF of U(2, 3) at 1, 2.5 and 4
[1] 0 1 0
> qunif(0.5, 2, 3)    # quantile of U(2, 3) at p=0.5
[1] 2.5
```

Normal Distribution $N(\mu, \sigma^2)$: the PDF is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \Big\}, \quad -\infty < x < \infty,$$

where $\mu \in (-\infty, \infty)$ is the 'centre' (or mean) of the distribution, and $\sigma > 0$ is the 'spread' (standard deviation).

Remarks. (i) The most important distribution in statistics: Many phenomena in nature have approximately normal distributions. Furthermore, it provides asymptotic approximations for the distributions of sample means (Central Limit Theorem).

(ii) If $X \sim N(\mu, \sigma^2)$, $EX = \mu$, $\text{Var}(X) = \sigma^2$, and $\psi_X(t) = e^{\mu t + \sigma^2 t^2/2}$.

We compute $\psi_X(t)$ below, the idea is applicable in general.

$$
\begin{aligned}
\psi_X(t) &= \frac{1}{\sqrt{2\pi}\sigma}\int e^{tx}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}dx = \frac{1}{\sqrt{2\pi}\sigma}\int e^{-\frac{1}{2\sigma^2}(x^2-2\mu x-2tx\sigma^2+\mu^2)}dx \\
&= \frac{1}{\sqrt{2\pi}\sigma}\int e^{-\frac{1}{2\sigma^2}[\{x-(\mu+t\sigma^2)\}^2-(\mu+t\sigma^2)^2+\mu^2]}dx \\
&= e^{\frac{1}{2\sigma^2}\{(\mu+t\sigma^2)^2-\mu^2\}}\frac{1}{\sqrt{2\pi}\sigma}\int e^{-\frac{1}{2\sigma^2}\{x-(\mu+t\sigma^2)\}^2}dx \\
&= e^{\frac{1}{2\sigma^2}\{(\mu+t\sigma^2)^2-\mu^2\}} = e^{\mu t+t^2\sigma^2/2}
\end{aligned}
$$

(iii) Standard normal distribution: $N(0,1)$.

If $X \sim N(\mu,\sigma^2)$, $Z \equiv (X-\mu)/\sigma \sim N(0,1)$. Hence

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right),$$

where

$$\Phi(x) = P(Z < x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du$$

is the CDF of $N(0, 1)$. Its values are tabulated in all statistical tables.

Example 3. Let $X \sim N(3, 5)$.

$$P(X > 1) = 1 - P(X < 1) = 1 - P\big(Z < \frac{1-3}{\sqrt{5}}\big) = 1 - \Phi(-0.8944) = 0.81.$$

Now find $x = \Phi^{-1}(0.2)$, i.e. x satisfies the equation

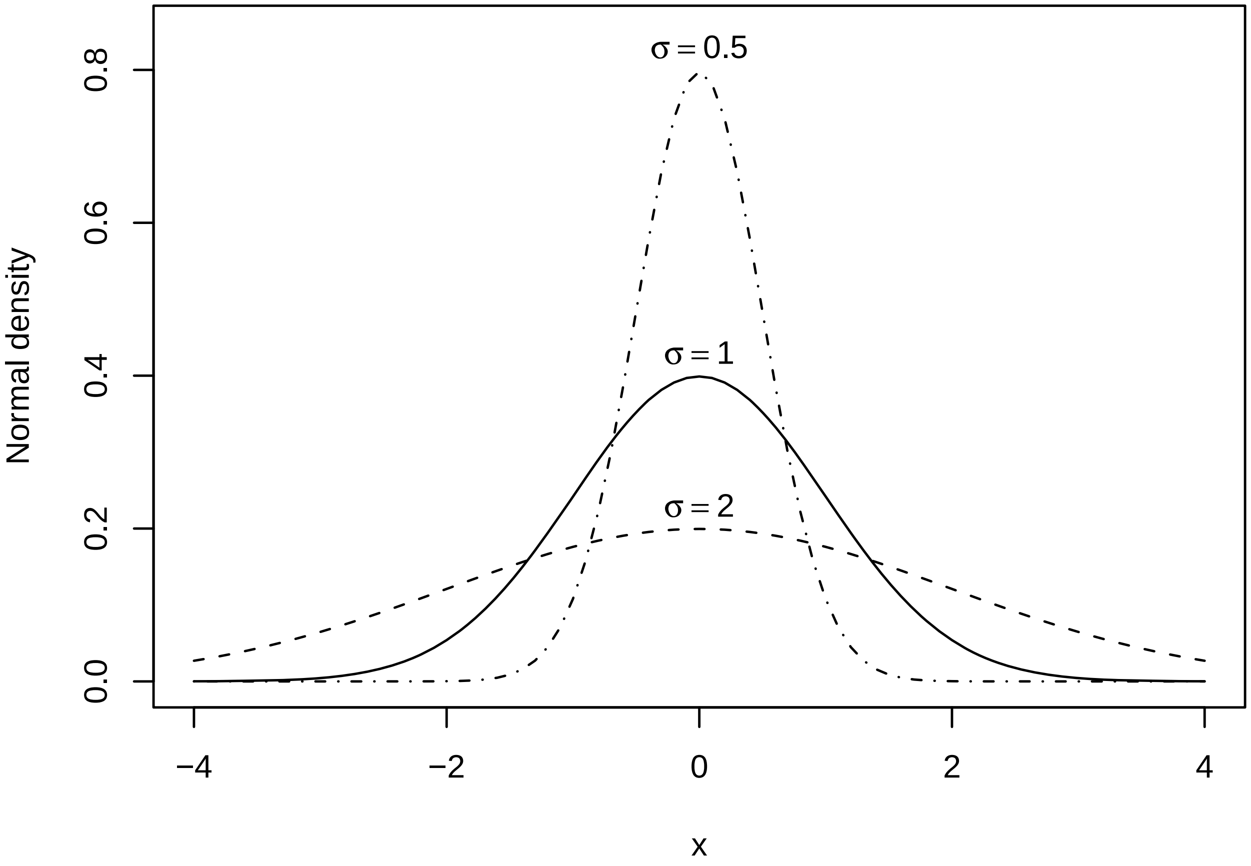
$$0.2 = P(X < x) = P\big(Z < \frac{x-3}{\sqrt{5}}\big).$$

From the normal table, $\Phi(-0.8416) = 0.2$. Therefore $(x-3)/\sqrt{5} = -0.8416$, leading to the solution $x = 1.1181$.

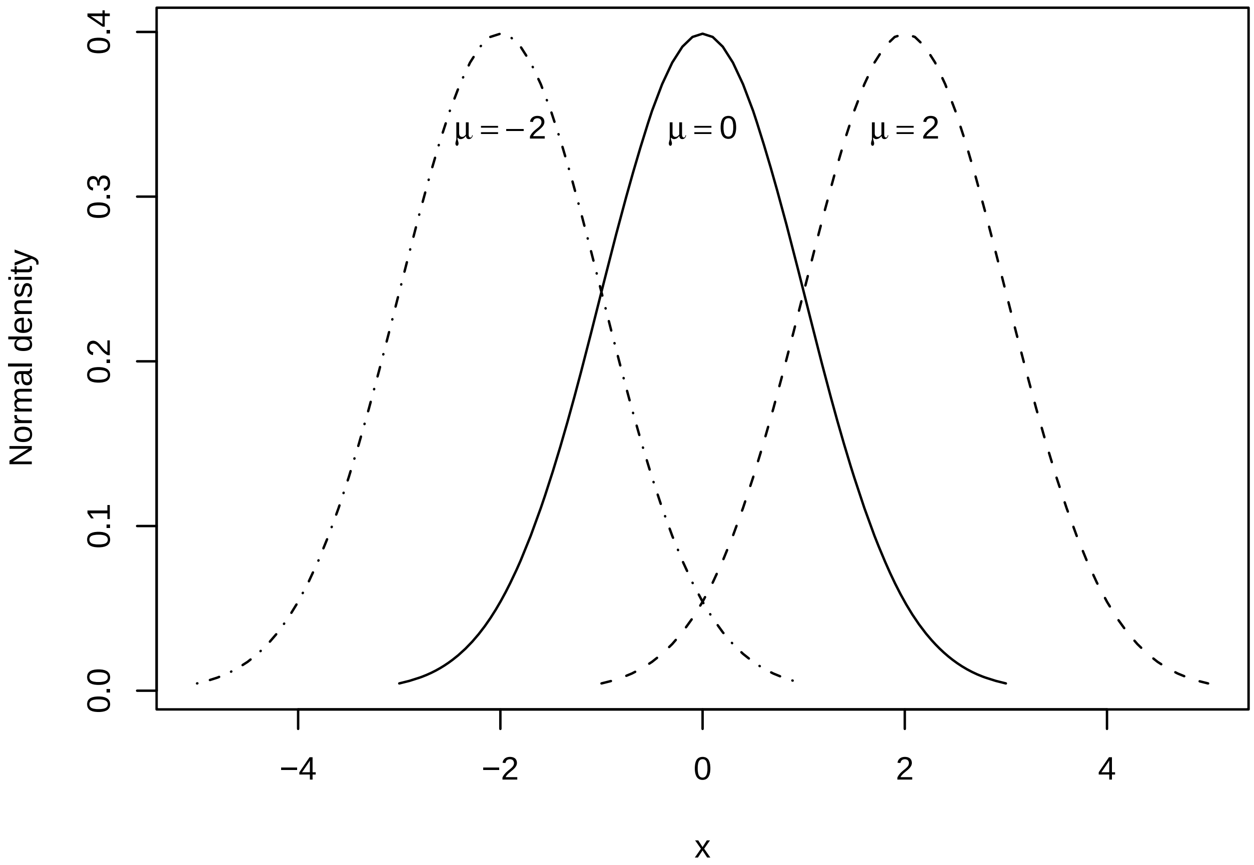
Note. You may check the answers using R:

```
1 - pnorm(1, 3, sqrt(5)),
 qnorm(0.2, 3, sqrt(5))
```

Density functions of $N(0, \sigma^2)$



Density functions of $N(\mu, 1)$



Below are the R codes which produce the two normal density plots.

```r
x <- seq(-4, 4, 0.1)      # x = (-4, -3.9, -3.8, ..., 3.9, 4)
plot(x, dnorm(x, 0, 1), type='l', xlab='x', ylab='Normal density',
                           ylim=c(0, 0.85))
text(0,0.43, expression(sigma==1))
lines(x, dnorm(x, 0, 2), lty=2)
text(0,0.23, expression(sigma==sqrt(2)))
lines(x, dnorm(x, 0, 0.5), lty=4)
text(0,0.83, expression(sigma==sqrt(0.5)))

x <- seq(-3, 3, 0.1)
plot(x, dnorm(x, 0, 1), type='l', xlab='x', ylab='Normal density',
                   xlim=c(-5, 5))
text(0,0.34, expression(mu==0))
lines(x+2, dnorm(x+2, 2, 1), lty=2)
text(2,0.34, expression(mu==2))
lines(x-2, dnorm(x-2, -2, 1), lty=4)
text(-2,0.34, expression(mu==-2))
```

Exponential Distribution $\text{Exp}(\lambda)$: $X \sim \text{Exp}(\lambda)$ if X has the PDF

$$f(x) = \begin{cases} \frac{1}{\lambda} e^{-x/\lambda} & x > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $\lambda > 0$ is a parameter.

$$E(X) = \lambda, \quad \text{Var}(X) = \lambda^2, \quad \psi_X(t) = 1/(1 - t\lambda).$$

Background. $\text{Exp}(\lambda)$ is used to model the lifetime of electronic components and the waiting times between rare events.

Gamma Distribution $\text{Gamma}(\alpha, \beta)$: $X \sim \text{Gamma}(\alpha, \beta)$ if X has the PDF

$$f(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $\alpha, \beta > 0$ are two parameters, and $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$.

$$E(X) = \alpha\beta, \quad \text{Var}(X) = \alpha\beta^2, \quad \psi_X(t) = (1 - t\beta)^{-\alpha}.$$

Note. Gamma$(1, \beta)$ = Exp(β).

Cauchy Distribution: the PDF of the Cauchy distribution is

$$f(x) = \frac{1}{\pi(1 + x^2)}, \qquad x \in (-\infty, \infty).$$

As $E(|X|) = \infty$, the mean and variance of the Cauchy distribution do not exist. Cauchy Distribution is particularly useful to model the data with excessively large, or negatively large outliers.

Chapter 4. Multivariate Distributions

4.1 Bivariate Distributions.

For a pair r.v.s (X, Y), the Joint CDF is defined as

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y).$$

Obviously, the marginal distributions may be obtained easily from the joint distribution:

$$F_X(x) = P(X \leq x) = P(X \leq x, Y < \infty) = F_{X,Y}(x, \infty),$$

and $F_Y(y) = F_{X,Y}(\infty, y)$.

Covariance and correlation of X and Y:

$$\text{Cov}(X, Y) = E\{(X - EX)(Y - EY)\} = E(XY) - (EX)(EY),$$
$$\text{Corr}(X, Y) = \text{Cov}(X, Y)/\sqrt{\text{Var}(X)\text{Var}(Y)}.$$

Discrete bivariate distributions

If X takes discrete values $x_1, \cdots, x_m$ and Y takes discrete values $y_1, \cdots, y_n$, their joint probability function may be presented in a table:

$X \backslash Y$	y_1	y_2	$\cdots$	y_n	
x_1	p_{11}	p_{12}	$\cdots$	p_{1n}	$p_{1\cdot}$
x_2	p_{21}	p_{22}	$\cdots$	p_{2n}	$p_{2\cdot}$
		$\cdots$	$\cdots$		
x_m	p_{m1}	p_{22}	$\cdots$	p_{mn}	$p_{m\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot n}$	

where $p_{ij} = P(X = x_i, Y = y_j)$, and

$$p_{i\cdot} = P(X = x_i) = \sum_{j=1}^{n} P(X = x_i, Y = y_j) = \sum_{j} p_{ij},$$

$$p_{\cdot j} = P(Y = y_j) = \sum_{i=1}^{m} P(X = x_i, Y = y_j) = \sum_{i} p_{ij}.$$

In general, $p_{ij} \neq p_{i\cdot} \times p_{\cdot j}$. However if $p_{ij} = p_{i\cdot} \times p_{\cdot j}$ for all i and j, X and Y are *independent*, i.e.

$$P(X = x_i, Y = y_j) = P(X = x_i) \times P(Y = y_j), \quad \forall i, j.$$

For independent X and Y, $\mathrm{Cov}(X, Y) = 0$.

Example 1. Flip a fair coin two times. Let $X = 1$ if H occurs in the first flip, and 0 if T occurs in the first flip. Let $Y = 1$ if the outcomes in the two flips are the same, and 0 if the two outcomes are different. The joint probability function is

$X\backslash Y$	1	0	
1	1/4	1/4	1/2
0	1/4	1/4	1/2
	1/2	1/2	

It is easy to see that X and Y are independent, which is a bit anti-intuitive.

Continuous bivariate distribution

If the CDF $F_{X,Y}$ can be written as

$$F_{X,Y}(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u,v) du dv \qquad \text{for any } x \text{ and } y,$$

where $f_{X,Y} \geq 0$, (X, Y) has a continuous joint distribution, and $f_{X,Y}(x,y)$ is the joint PDF.

As $F_{X,Y}(\infty, \infty) = 1$, it holds that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(u,v) du dv = 1.$$

In fact any non-negative function satisfying this condition is a PDF. Furthermore for any subset A in R^2,

$$P\{(X,Y) \in A\} = \int_A f_{X,Y}(x,y) dx dy.$$

Also

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \int (x-EX)(y-EY)f_{X,Y}(x,y)dxdy\\ &= \int xyf_{X,Y}(x,y)dxdy - EX\,EY.\end{aligned}$$

Note that

$$F_X(x) = F_{X,Y}(x,\infty) = \int_{-\infty}^{\infty}\int_{-\infty}^{x} f_{X,Y}(u,v)dudv = \int_{-\infty}^{x}\Big\{\int_{-\infty}^{\infty} f_{X,Y}(u,v)dv\Big\}du,$$

hence the *marginal PDF* of X can be derived from the joint PDF as follows

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy.$$

Similarly, $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$.

Note. Different from discrete cases, it is not always easy to work out marginal PDFs from joint PDFs, especially when PDFs are discontinuous.

When $f_{X,Y}(x,y) = f_X(x)f_y(y)$ for any x and y, X and Y are independent, as then

$$P(X \leq x, Y \leq y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u,v)dudv = \int_{-\infty}^{y}\int_{-\infty}^{x} f_X(u)f_Y(v)dudv$$
$$= \int_{-\infty}^{x} f_X(u)du \int_{-\infty}^{y} f_Y(v)dv = P(X \leq x)P(Y \leq y),$$

and also $\mathrm{Cov}(X,Y) = 0$.

Example 2. *Uniform distribution on unit square – $U[0,1]^2$.*

$$f(x,y) = \begin{cases} 1 & 0 \leq x \leq 1,\ 0 \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

This is well-defined PDF, as $f \geq 0$ and $\int\int f(x, y)dxdy = 1$. It is easy to see that X and Y are independent. Let us calculate some probabilities

$$\begin{aligned} P(X < 1/2, Y < 1/2) &= F(1/2, 1/2) = \int_{-\infty}^{1/2}\int_{-\infty}^{1/2} f_{X,Y}(x, y)dxdy \\ &= \int_0^{1/2}\int_0^{1/2} dxdy = 1/4. \end{aligned}$$

$$\begin{aligned} P(X + Y < 1) &= \int_{\{x+y<1\}} f_{X,Y}(x, y)dxdy = \int_{\{x>0,\, y>0,\, x+y<1\}} dxdy \\ &= \int_0^1 dy \int_0^{1-y} dx = \int_0^1 (1 - y)dy = 1/2. \end{aligned}$$

Example 3. Let (X, Y) have the joint PDF

$$f(x, y) = \begin{cases} x^2 + xy/3 & 0 \leq x \leq 1,\ 0 \leq y \leq 2, \\ 0 & \text{otherwsie.} \end{cases}$$

Calculate $P(0 < X < 1/2, 1/4 < Y < 3)$ and $P(X < Y)$. Are X and Y independent with each other?

$$P(0 < X < 1/2, 1/4 < Y < 3) = P(0 < X < 1/2, 1/4 < Y < 2)$$
$$= \int_{1/4}^{2} dy \int_{0}^{1/2} (x^2 + \frac{xy}{3}) dx = \int_{1/4}^{2} \frac{1+y}{24} dy = \frac{1.75}{24} + \frac{y^2}{48}\Big|_{1/4}^{2} = 0.155.$$

$$P(X < Y) = \int_0^1 dx \int_x^2 (x^2 + \frac{xy}{3}) dy = \int_0^1 (\frac{2}{3}x + 2x^2 - \frac{7}{6}x^3) dx = 17/24 = 0.708.$$

$$f_X(x) = \int_0^2 (x^2 + \frac{xy}{3}) dy = 2x^2 + \frac{2x}{3}, \quad f_Y(y) = \int_0^1 (x^2 + \frac{xy}{3}) dx = \frac{1}{3} + \frac{y}{6}.$$

Both $f_X(x)$ and $f_Y(y)$ are well-defined PDFs.

But $f(x, y) \neq f_X(x) f_Y(y)$, hence they are not independent.

4.2 Conditional Distributions

If X and Y are not independent, knowing X should be helpful in determining Y, as X may carry some information on Y. Therefore it makes sense to define the distribution of Y given, say, $X = x$. This is the concept of conditional distributions.

If both X and Y are discrete, the conditional probability function is simply a special case of conditional probabilities:

$$P(Y = y|X = x) = P(Y = y,\ X = x)\big/P(X = x).$$

However this definition does not extend to continuous r.v.s, as then $P(X = x) = 0$.

Definition (Conditional PDF). For continuous r.v.s X and Y, the conditional PDF of Y given $X = x$ is

$$f_{Y|X}(\cdot|x) = f_{X,Y}(x,\cdot)/f_X(x).$$

Remark. (i) As a function of y, $f_{Y|X}(y|x)$ is a PDF:

$$P(Y \in A|X = x) = \int_A f_{Y|X}(y|x)dy,$$

while x is treated as a constant (i.e. not a variable).

(ii) $E(Y|X = x) = \int y f_{Y|X}(y|x)dy$ is a function of x, and

$$\text{Var}(Y|X = x) = \int \{y - E(Y|X = x)\}^2 f_{Y|X}(y|x)dy.$$

(iii) If X and Y are independent, $f_{Y|X}(y|x) = f_Y(y)$.

(iv) $f_{X,Y}(x,y) = f_X(x)f_{Y|X}(y|x) = f_{X|Y}(x|y)f_Y(y)$, which offers alternative ways to determine the joint PDF.

(v) *$E\{E(Y|X)\} = E(Y)$ — This in fact holds for any r.v.s X and Y*. We give a proof here for continuous r.v.s only:

$$E\{E(Y|X)\} = \int \Big\{ \int y f_{Y|X}(y|x) dy \Big\} f_X(x) dx = \int \int y f_{X,Y}(x,y) dx dy$$

$$= \int y \Big\{ \int f_{X,Y}(x,y) dx \Big\} dy = \int y f_Y(y) dy = EY.$$

Example 4. Let $f_{X,Y}(x,y) = e^{-y}$ for $0 < x < y < \infty$, and 0 otherwise. Find $f_{Y|X}(y|x)$, $f_{X|Y}(x|y)$ and $\text{Cov}(X,Y)$.

We need to find $f_X(x)$, $f_Y(y)$ first:

$$f_X(x) = \int f_{X,Y}(x,y)dy = \int_x^\infty e^{-y}dy = e^{-x} \quad x \in (0,\infty),$$

$$f_Y(y) = \int f_{X,Y}(x,y)dx = \int_0^y e^{-y}dx = ye^{-y} \quad y \in (0,\infty).$$

Hence

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = e^{-(y-x)} \qquad y \in (x,\infty),$$

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = 1/y \qquad x \in (0,y).$$

Note that given $Y = y$, $X \sim U(0,\ y)$, i.e. the uniform distribution on $(0,\ y)$.

To find Cov(X, Y), we compute $EX,\ EY$ and $E(XY)$ first.

$$\begin{aligned}
EX &= \int xf_X(x)dx = \int_0^\infty xe^{-x}dx = -e^{-x}(1+x)\Big|_0^\infty = 1, \\
EY &= \int yf_Y(y)dy = \int_0^\infty y^2e^{-y}dy = -y^2e^{-y}\Big|_0^\infty + 2\int_0^\infty ye^{-y}dy = 2, \\
E(XY) &= \int xyf_{X,Y}(x,y)dxdy = \int_0^\infty dy \int_0^y xye^{-y}dx = \frac{1}{2}\int_0^\infty y^3e^{-y}dy \\
&= -\frac{1}{2}y^3e^{-y}\Big|_0^\infty + \frac{3}{2}\int_0^\infty y^2e^{-y}dy = 3.
\end{aligned}$$

Hence Cov$(X, Y) = E(XY) - (EX)(EY) = 3 - 2 = 1$.

4.3 Multivariate Distributions

Let $\mathbf{X} = (X_1, \cdots, X_n)'$ be a random vector (r.v.) consisting of n r.v.s. The joint CDF is defined as

$$F(x_1, \cdots, x_n) \equiv F_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = P(X_1 \leq x_1,\ \cdots,\ X_n \leq x_n).$$

If X is continuous, its PDF f satisfies

$$F(x_1, \cdots, x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(u_1, \cdots, u_n) du_1 \cdots du_n.$$

In general, the PDF admits the factorisation

$$f(x_1, \cdots, x_n) = f(x_1) f(x_2|x_1) f(x_3|x_1, x_2) \cdots f(x_n|x_1, \cdots, x_{n-1}),$$

where $f(x_j|x_1, \cdots, x_{j-1})$ denotes the conditional PDF of X_j given $X_1 = x_1, \cdots, X_{j-1} = x_j$.

However, when $X_1, \cdots, X_n$ are independent,

$$f_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n).$$

IID Samples. If $X_1, \cdots, X_n$ are independent and each has the same CDF F, we say that $X_1, \cdots, X_n$ are IID (independent and identically distributed) and write

$$X_1, \cdots, X_n \sim_{iid} F.$$

We also call $X_1, \cdots, X_n$ *a sample* or *a random sample*.

4.3 Two important multivariate distributions

<ins>Multinomial Distribution</ins> Multinomial$(n, p_1, \cdots, p_k)$ — an extension of Bin(n, p).

Suppose we threw a k-sided die n times, record X_i as the number of times ended with the i-th side, $i = 1, \cdots, k$. Then

$$(X_1, \cdots, X_k) \sim \text{Multinomial}(n, p_1, \cdots, p_k),$$

where p_i is the probability of the event that the i-th side occurs in one threw. Obviously $p_i \geq 0$ and $\sum_i p_i = 1$.

We may immediately make the following observation from the above definition.

(i) $X_1 + \cdots + X_k \equiv n$, therefore $X_1, \cdots, X_n$ are not independent.

(ii) $X_i \sim \text{Bin}(n, p_i)$, hence $EX_i = np_i$ and $\text{Var}(X_i) = np_i(1 - p_i)$.

The joint probability function for Multinomial$(n, p_1, \cdots, p_k)$:

For any $j_1, \cdots, j_k \geq 0$ and $j_1 + \cdots + j_k = n$,

$$P(X_1 = j_1, \cdots, X_k = j_k) = \frac{n!}{j_1! \cdots j_k!} p_1^{j_1} \cdots p_k^{j_k}.$$

Multivariate Normal Distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: a k-variable r.v. $\mathbf{X} = (X_1, \cdots, X_k)'$ is normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ if its PDF is

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\Big\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big\} \qquad \mathbf{x} \in R^k,$$

where $\boldsymbol{\mu}$ is k-vector, and $\boldsymbol{\Sigma}$ is a $k \times k$ positive-definite matrix.

Some properties of $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: Let $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_k)'$ and $\boldsymbol{\Sigma} \equiv (\sigma_{ij})$, then

(i) $E\mathbf{X} = \boldsymbol{\mu}$, and the covariance matrix

$$\text{Cov}(\mathbf{X}) = E\{(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})'\} = \boldsymbol{\Sigma},$$

and

$$\sigma_{ij} = \text{Cov}(X_i, X_j) = E\{(X_i - \mu_i)(X_j - \mu_j)\}.$$

(ii) When $\sigma_{ij} = 0$ for all $i \neq j$, i.e. the components of $\mathbf{X}$ are *uncorrelated*, $\mathbf{\Sigma} = \text{diag}(\sigma_{11}, \cdots, \sigma_{kk})$, $|\mathbf{\Sigma}| = \prod_i \sigma_{ii}$. Hence the PDF admits a simple form

$$f(\mathbf{x}) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma_{ii}}} \exp\{-\frac{1}{2\sigma_{ii}}(x_i - \mu_i)^2\}.$$

Thus $X_1, \cdots, X_n$ are independent when $\sigma_{ij} = 0$ for all $i \neq j$.

(iii) Let $\mathbf{Y} = \mathbf{AX} + \mathbf{b}$, where $\mathbf{A}$ is a constant matrix and $\mathbf{b}$ is a constant vector. Then $\mathbf{Y} \sim N(\mathbf{A}\mu + \mathbf{b}, \mathbf{A\Sigma A'})$.

(iv) $X_i \sim N(\mu_i, \sigma_{ii})$. For any constant k-vector **a**, $\mathbf{a'X}$ is a scale r.v. and $\mathbf{a'X} \sim N(\mathbf{a'}\boldsymbol{\mu}, \mathbf{a'\Sigma a})$.

(v). **Standard Normal Distribution**: $N(0, \mathbf{I_k})$, where $\mathbf{I}_k$ is the $k \times k$ identity matrix.

Example 5. Let X_1, X_2, X_3 be jointly normal with the common mean 0, variance 1 and

$$\mathrm{Corr}(X_i, X_j) = 0.5, \qquad 1 \le i \ne j \le 3.$$

Find the probability $P(|X_1| + |X_2| + |X_3| \le 2)$.

It is difficult to calculate this probability by the integration of the joint PDF. We provide an estimate by simulation. (The justification will be in Chapter 6). We solve a general problem first.

Let $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{X}$ has p component. For any set $A \subset R^p$, we may estimate the probability $P(\mathbf{X} \in A)$ by the relative frequency

$$\#\{1 \le i \le n : \mathbf{X}_i \in A\} \big/ n,$$

where n is a large integer, and $\mathbf{X}_1, \cdots, \mathbf{X}_n$ are n vectors generated independently from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Note

$$\mathbf{X} = \mu + \mathbf{\Sigma}^{1/2}\mathbf{Z},$$

where $\mathbf{Z} \sim N(0, \mathbf{I_p})$ is standard normal, and $\mathbf{\Sigma}^{1/2} \geq 0$ and $\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2} = \mathbf{\Sigma}$. We generate $\mathbf{Z}$ by `rnorm(p)`, and apply the above linear transformation to obtain $\mathbf{X}$.

$\mathbf{\Sigma}^{1/2}$ may be obtained by an eigenanalysis for $\mathbf{\Sigma}$ using *R*-function `eigen`. Since $\mathbf{\Sigma} \geq 0$, it holds that

$$\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}',$$

where $\mathbf{\Gamma}$ is an orthogonal matrix (i.e. $\mathbf{\Gamma}'\mathbf{\Gamma} = \mathbf{I}_p$), $\mathbf{\Lambda} = \text{diag}(\lambda_1, \cdots, \lambda_p)$ is a diagonal matrix. Then

$$\mathbf{\Sigma}^{1/2} = \mathbf{\Gamma}\mathbf{\Lambda}^{1/2}\mathbf{\Gamma}', \quad \text{where } \mathbf{\Lambda}^{1/2} = \text{diag}(\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_p}).$$

The *R* function `rMNorm` below generate random vectors from $N(\boldsymbol{\mu}, \mathbf{\Sigma})$.

```r
rMNorm <- function(n, p, mu, Sigma) {
   # generate n p-vectors from N(mu, Sigma)
   # mu is p-vector of mean, Sigma >=0 is pxp matrix
t <- eigen(Sigma, symmetric=T) # eigenanalysis for Sigma
ev <- sqrt(t$values) # square-roots of the eigenvalues
G <- as.matrix(t$vectors) # line up eigenvectors into a matrix G
D <- G*0; for(i in 1:p) D[i,i] <- ev[i]; # D is diagonal matrix
P <- G%*%D%*%t(G) # P=GDG' is the required transformation matrix
Z <- matrix(rnorm(n*p), byrow=T, ncol=p)
  # Z is nxp matrix with elements drawn from N(0,1)
Z <- Z%*%P # Now each row of Z is N(0, Sigma)
X <- matrix(rep(mu, n), byrow=T, ncol=p) + Z
   # each row of X is N(mu, Sigma)
}
```

This function is saved in the file ‘rMNorm.r’. We may use it to perform the required task:

```r
source("rMNorm.r")
```

```r
mu <- c(0, 0, 0)
Sigma <- matrix(c(1,0.5,0.5,0.5,1,0.5,0.5,0.5,1), byrow=T, ncol=3)
X <- rMNorm(20000, 3, mu, Sigma)
dim(X)  # check the size of X
t <- abs(X[,1]) + abs(X[,2]) + abs(X[,3])
cat("Estimated probability:", length(t[t<=2])/20000, "\n")
```

It returned the value:

```
Estimated probability:  0.446
```

I repeated it a few more times and obtained the estimates 0.439, 0.445, 0.441 etc.

4.4 Transformations of random variables

Let a random vector $\mathbf{X}$ have PDF $f_{\mathbf{X}}$. We are interested in the distribution of a scalar function of $\mathbf{X}$, say, $Y = r(\mathbf{X})$. We introduce a general procedure first.

Three steps to find the PDF of $Y = r(\mathbf{X})$:

(i) For each y, find the set $A_y = \{\mathbf{x} : \ r(\mathbf{x}) \leq y\}$
(ii) Find the CDF

$$F_Y(y) = P(Y \leq y) = P\{r(\mathbf{X}) \leq y\} = \int_{A_y} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}.$$

(iii) $f_Y(y) = \frac{d}{dy} F_Y(y)$.

Example 6. Let $X \sim f_X(x)$ (X is a scalar). Find the PDF of $Y = e^X$.

$A_y = \{x : e^x \le y\} = \{x : x \le \log y\}$. Hence

$$F_Y(y) = P(Y \le y) = P\{e^X \le y\} = P(X \le \log y) = F_X(\log y).$$

Hence

$$f_Y(y) = \frac{d}{dy} F_X(\log y) = f_X(\log y) \frac{d \log y}{dy} = y^{-1} f_X(\log y).$$

Note that $y = e^x$ and $\log y = x$, $\frac{dy}{dx} = e^x = y$. The above result can be written as

$$f_Y(y) = f_X(x) \Big/ \frac{dy}{dx}, \quad \text{or} \quad f_Y(y) dy = f_X(x) dx.$$

For 1-1 transformation $Y = r(X)$ (i.e. the inverse function $X = r^{-1}(Y)$ is uniquely defined), it holds that

$$f_Y(y) = f_X(x)/|r'(x)| = f_X(x)\left|\frac{dx}{dy}\right|.$$

Note. You should replace all x in the above by $x = r^{-1}(y)$.

Example 7. Let $X \sim U(-1, 3)$. Find the PDF of $Y = X^2$. Now this is not a 1-1 transformation. We have to use the general 3-step procedure.

Note that Y takes values in $(0, 9)$. Consider two cases:

(i) For $y \in (0, 1)$, $A_y = (-\sqrt{y}, \sqrt{y})$, $F_y(y) = \int_{A_y} f_X(x)dx = 0.5\sqrt{y}$. Hence $f_Y(y) = F'_Y(y) = 0.25/\sqrt{y}$.

(ii) For $y \in [1, 9)$, $A_y = (-1, \sqrt{y})$, $F_y(y) = \int_{A_y} f_X(x)dx = 0.25(\sqrt{y} + 1)$. Hence $f_Y(y) = F'_Y(y) = 0.125/\sqrt{y}$.

Collectively we have

$$f_Y(y) = \begin{cases} 0.25/\sqrt{y} & 0 < y < 1 \\ 0.125/\sqrt{y} & 1 \leq y < 9 \\ 0 & \text{otherwise.} \end{cases}$$

Chapter 5. Inequalities

Inequalities are useful tools in establishing various properties of statistical inference methods. They may also provide estimates for probabilities with little assumption on probability distributions.

5.1 Probability inequalities

Markov's inequality. Let X be a non-negative r.v. and $EX < \infty$. Then for any $t > 0$, $P(X > t) \leq EX \big/ t$.

An immediate corollary of Markov's inequality: For any r.v. X and any constant $t > 0$,

$$P(|X| > t) \le \frac{E|X|}{t} \quad \text{provided } E|X| < \infty,$$

$$P(|X| > t) \le \frac{E\{|X|^k\}}{t^k} \quad \text{provided } E\{|X|^k\} < \infty. \tag{1}$$

The tail-probability $P(|X| > t)$ is a useful measure in insurance and risk management in finance. (1) implies that the more moments X has, the smaller the tail probabilities are.

Proof of Markov's inequality. Since $X > 0$,

$$\begin{aligned} EX &= \int_0^\infty xf(x)dx = \int_0^t xf(x)dx + \int_t^\infty xf(x)dx \\ &\geq \int_t^\infty xf(x)dx \geq t\int_t^\infty f(x)dx = tP(X > t). \end{aligned}$$

Chebyshev's inequality. Suppose a r.v. X have mean μ and variance $\sigma^2 \in (0, \infty)$. Then $P(|X-\mu| \geq t) \leq \sigma^2/t^2$ for any $t > 0$.

Remarks. (i) Chebyshev's inequality follows from (1) with X replaced by $X - \mu$ and $k = 2$.

(ii) Replacing t by $t\sigma$, we have $P(|Z| > t) \leq 1/t^2$, where $Z = (X-\mu)/\sigma$ is a standardization of X.

(iii) For any r.v. X with mean 0 and variance 1, it holds that

$$P(|X| > 2) \leq 1/4, \qquad P(|X| > 3) \leq 1/9.$$

Example 1. We flipped a coin n times with Head occurred k $(< n)$ times. Therefore a natural estimate for the probability $p = P(\text{Head})$ is k/n. What is the error $k/n - p$ in this estimation?

Let $X_i = 1$ if Head occurred in the i-th flip, and 0 otherwise. Then $k = \sum_{i=1}^n X_i$, and $k/n = n^{-1}\sum_{i=1}^n X_i \equiv \bar{X}_n$. Note k, therefore also $\bar{X}_n$, may take different value if we repeat the experiment. Hence it makes sense to quantify the estimation error in probability such as $P(|\bar{X}_n - p| > \epsilon)$ for some small constant $\epsilon > 0$.

Note $E(\bar{X}_n) = n^{-1}\sum EX_i = p$, $\text{Var}(\bar{X}_n) = n^{-2}\sum \text{Var}(X_i) = n^{-1}\text{Var}(X_1)$ $= n^{-1}p(1-p)$. It follows from Chebyshev's inequality that

$$P(|\bar{X}_n - p| > \epsilon) \le \frac{p(1-p)}{n\epsilon^2} \le \frac{1}{4n\epsilon^2}.$$

Let $\epsilon = 0.1$ and $n = 500$, $P(|\bar{X}_n - p| > 0.1) \le 1/(20) = 0.05$.

5.2 Inequalities for expectations

Cauchy-Schwartz inequality. Let $E(X^2) < \infty$ and $E(Y^2) < \infty$. Then $E|XY| \leq \{E(X^2)E(Y^2)\}^{1/2}$.

A function g is *convex* if for any x, y and any $\alpha \in [0, 1]$,

$$g(\alpha x + (1 - \alpha)y) \leq \alpha g(x) + (1 - \alpha)g(y).$$

If $g''(x) \geq 0$ for all x, g is convex. Examples of convex functions include $g(x) = x^2$ and $g(x) = e^x$.

A function g is *concave* if $-g$ is convex. Examples of concave functions are $g(x) = -x^2$ and $g(x) = \log(x)$.

Jensen's inequality. If g is convex, $E\{g(X)\} \geq g(EX)$.

From Jensen's inequality, we have $E(X^2) \geq (EX)^2$. If $X \geq 0$, $E(1/X) \geq 1/EX$ and $E(\log X) \leq \log(EX)$.

Chapter 6. Convergence of Random Variables and Monte Carlo Methods

6.1 Type of convergence

The two main types of convergence are defined as follows.

Let $X_1, X_n, \cdots$ be a sequence of r.v.s, and X be another r.v.

1. X_n **converges to** X **in probability**, denoted by $X_n \xrightarrow{P} X$, if for any constant $\epsilon > 0$, $P(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$.
2. X_n **converges to** X **in distribution**, denoted by $X_n \xrightarrow{D} X$, if $\lim_n F_{X_n}(x) = F_X(x)$ for any x (at which F_X is continuous).

Remarks. (i) X may be a constant (as a constant is a r.v. with probability mass concentrated on a single point.)

(ii) If $X_n \xrightarrow{P} X$, it also holds that $X_n \xrightarrow{D} X$, but not visa versa.

Example 1. Let $X \sim N(0,1)$ and $X_n = -X$ for all $n \geq 1$. Then $F_{X_n} \equiv F_X$. Hence $X_n \xrightarrow{D} X$. But $X_n \not\xrightarrow{P} X$, as for any $\epsilon > 0$

$$P(|X_n - X| > \epsilon) = P(2|X| > \epsilon) = P(|X| > \epsilon/2) > 0.$$

However if $X_n \xrightarrow{D} c$ and c is a constant, it holds that $X_n \xrightarrow{P} c$.

Note. We need the two types of convergence.

For example, let $\widehat{\theta}_n = h(X_1, \cdots, X_n)$ be an estimator for θ.

Naturally we require $\widehat{\theta}_n \xrightarrow{P} \theta$.

But $\widehat{\theta}_n$ is a random variable, it takes different values with different samples. To consider how good it is as an estimator for θ, we hope that the distribution of $(\widehat{\theta}_n - \theta)$ becomes more concentrated around 0 when n increases.

(iii) It is sometimes more convenient to consider the mean square convergence:

$$E\{(X_n - X)^2\} \to 0 \qquad \text{as } n \to \infty,$$

denoted by $X_n \xrightarrow{m.s.} X$. It follows from Markov's inequality that

$$P(|X_n - X| > \epsilon) = P(|X_n - X|^2 > \epsilon^2) \leq \frac{E\{|X_n - X|^2\}}{\epsilon^2}.$$

Hence if $X_n \xrightarrow{m.s.} X$, it also holds that $X_n \xrightarrow{P} X$, but not visa versa.

Example 2. Let $U \sim U(0,1)$ and $X_n = nI_{\{U<1/n\}}$. Then $P(|X_n| > \epsilon) \leq P(U < 1/n) = 1/n \to 0$, hence $X_n \xrightarrow{P} 0$. However

$$E(X_n^2) = n^2 P(U < 1/n) = n \to \infty.$$

Hence $X_n \overset{m.s.}{\not\longrightarrow} 0$.

(iv) $X_n \xrightarrow{P} X$ does not imply $EX_n \to EX$.

Example 3. Let $X_n = n$ with probability $1/n$ and 0 with probability $1 - 1/n$. Then $X_n \xrightarrow{P} 0$. However $EX_n = 1 \not\to 0$.

(v) When $X_n \xrightarrow{D} X$, we also write $X_n \xrightarrow{D} F_X$, where F_X is the CDF of X.

However the notation $X_n \xrightarrow{P} F_X$ does not make sense!

Slutsky's Theorem. Let X_n, Y_n, X, Y be r.v.s, g be a continuous function, and c is a constant.

(i) If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n + Y_n \xrightarrow{P} X + Y$, $X_n Y_n \xrightarrow{P} XY$, and $g(X_n) \xrightarrow{P} g(X)$.

(ii) If $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{D} c$, then $X_n + Y_n \xrightarrow{D} X + c$, $X_n Y_n \xrightarrow{D} cX$, and $g(X_n) \xrightarrow{D} g(X)$.

Note. $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{D} Y$ does not in general imply $X_n + Y_n \xrightarrow{D} X + Y$.

Slutzky's theorem is very useful, as it implies, e.g., $\bar{X}_n^2 \xrightarrow{P} \mu^2$, and $\bar{X}_n / S_n \xrightarrow{P} \mu/\sigma$ (see Exercise 4.3).

Recall the limits of sequences of real numbers $x_1, x_2, \cdots$: if $\lim_{n\to\infty} x_n = x$ (or, simply, $x_n \to x$), we mean $|x_n - x| \to 0$ as $n \to \infty$.

For a sequence of r.v.s $X_1, X_2, \cdots$, we say X is the limit of $\{X_n\}$ if $|X_n - X| \to 0$. Now there are some subtle issues here:

(i) $|X_n - X|$ is a r.v., it takes different values in the sample space Ω. Therefore $|X_n - X| \to 0$ should hold (almost) on the entirely sample space. This calls for some probability statement.

(ii) Since r.v.s have distributions, we may also consider $F_{X_n}(x) \to F_X(x)$ for all x.

Recall two simple facts: for any r.v.s $Y_1, \cdots, Y_n$ and constants $a_1, \cdots, a_n$,

$$E\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i E Y_i, \tag{2}$$

and if $Y_1, \cdots, Y_n$ are uncorrelated (i.e. $\mathrm{Cov}(Y_i, Y_j) = 0 \ \forall\ i \neq j$)

$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(Y_i). \tag{3}$$

Proof for (3). First note that for any r.v. U, $\text{Var}(U) = \text{Var}(U - EU)$. Because of (2), we may assume $EY_i = 0$ for all i. Thus

$$\text{Var}\big(\sum_{i=1}^{n} a_i Y_i\big) = E\big(\sum_{i=1}^{n} a_i Y_i\big)^2 = E\big(\sum_{i=1}^{n} a_i^2 Y_i^2 + \sum_{i \neq j} a_i a_j Y_i Y_j\big)$$

$$= \sum_{i=1}^{n} a_i^2 E(Y_i^2) + \sum_{i \neq j} a_i a_j E(Y_i Y_j) = \sum_{i=1}^{n} a_i^2 \text{Var}(Y_i) + \sum_{i \neq j} a_i a_j (EY_i)(EY_j)$$

$$= \sum_{i=1}^{n} a_i^2 \text{Var}(Y_i).$$

6.2 Two important limit theorems: LLN and CLT

Let $X_1, X_2, \cdots$ be IID with mean μ and variance $\sigma^2 \in (0, \infty)$. Let $\bar{X}_n$ denote the sample mean:

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n), \qquad n = 1, 2, \cdots .$$

We recall two simple facts:

$$E\bar{X}_n = \mu, \qquad \text{Var}(\bar{X}_n) = \sigma^2/n.$$

The (weak) Law of Large Numbers (LLN):

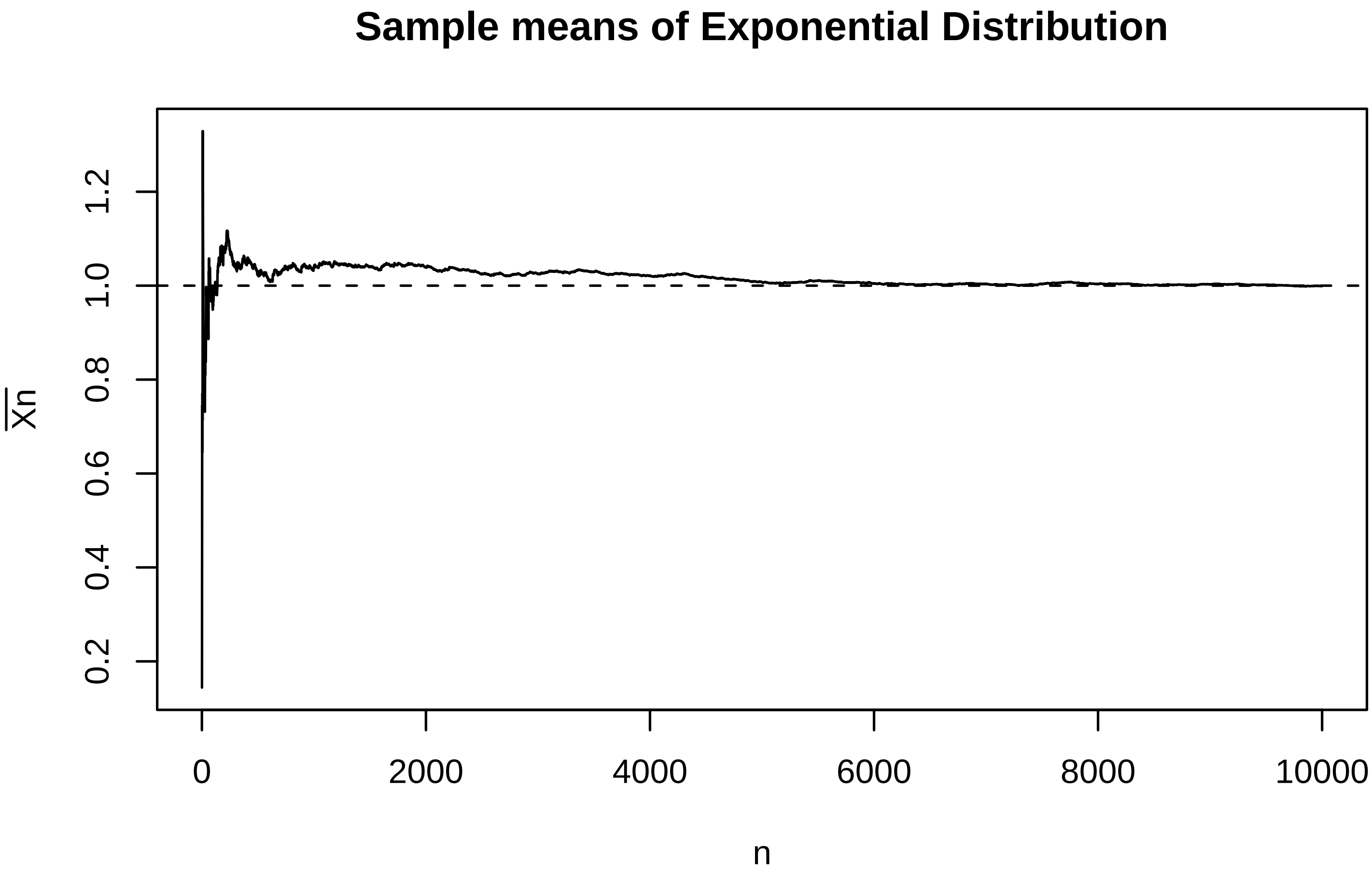
$$\text{As } n \to \infty, \ \bar{X}_n \xrightarrow{P} \mu.$$

The LLN is very natural: When the sample size increases, the sample mean becomes more and more close to the population mean. Furthermore, the distribution of $\bar{X}_n$ degenerates to a single point distribution at μ.

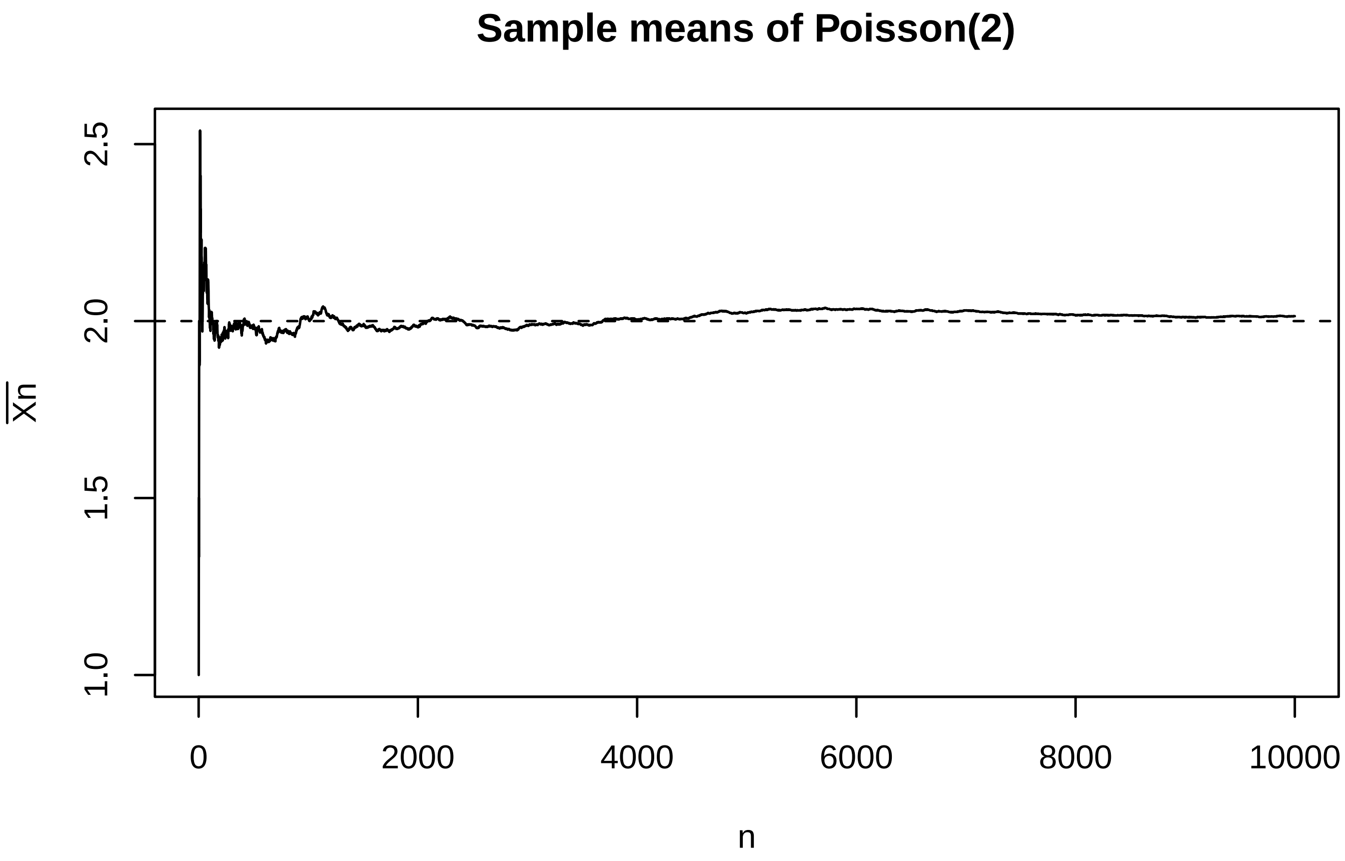
Proof. It follows from Chebyshev's inequality directly.

To visualize the LLN, we simulate the sample paths for

```
> x <- rexp(10000)   # generate 10000 random numbers from Exp(1)
> summary(x)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0001666 0.2861000 0.7098000 1.0220000 1.4230000 8.6990000
> n <- 1:10000
> ms <- n
> for(i in 1:10000) ms[i] <- mean(x[1:i])
> plot(n, ms, type='l', ylab=expression(bar(Xn)),
     main='Sample means of Exponential Distribution')
> abline(1,0,lty=2)  # draw a horizontal line at y=1
```


Sample means of Exponential Distribution
$\overline{Xn}$
0.2
0.4
0.6
0.8
1.0
1.2
0
2000
4000
6000
8000
10000
n

We repeat this exercise for Poisson(2):



The Central Limit Theorem (CLT):

$$\text{As } n \to \infty, \sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{D} N(0, 1).$$

Note the CLT can be expressed as

$$P\left\{ \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} \leq x \right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du = \Phi(x)$$

for any x, as $n \to \infty$, i.e. the *standardized* sample mean is approximately standard normal when the sample size is large. Hence in addition to $\sqrt{n}(\bar{X}_n - \mu)/\sigma \approx N(0, 1)$, we also see the expressions such as

$$\bar{X}_n \approx N(\mu, \sigma^2/n), \quad \bar{X}_n - \mu \approx N(0, \sigma^2/n), \quad \sqrt{n}(\bar{X}_n - \mu) \approx N(0, \sigma^2).$$

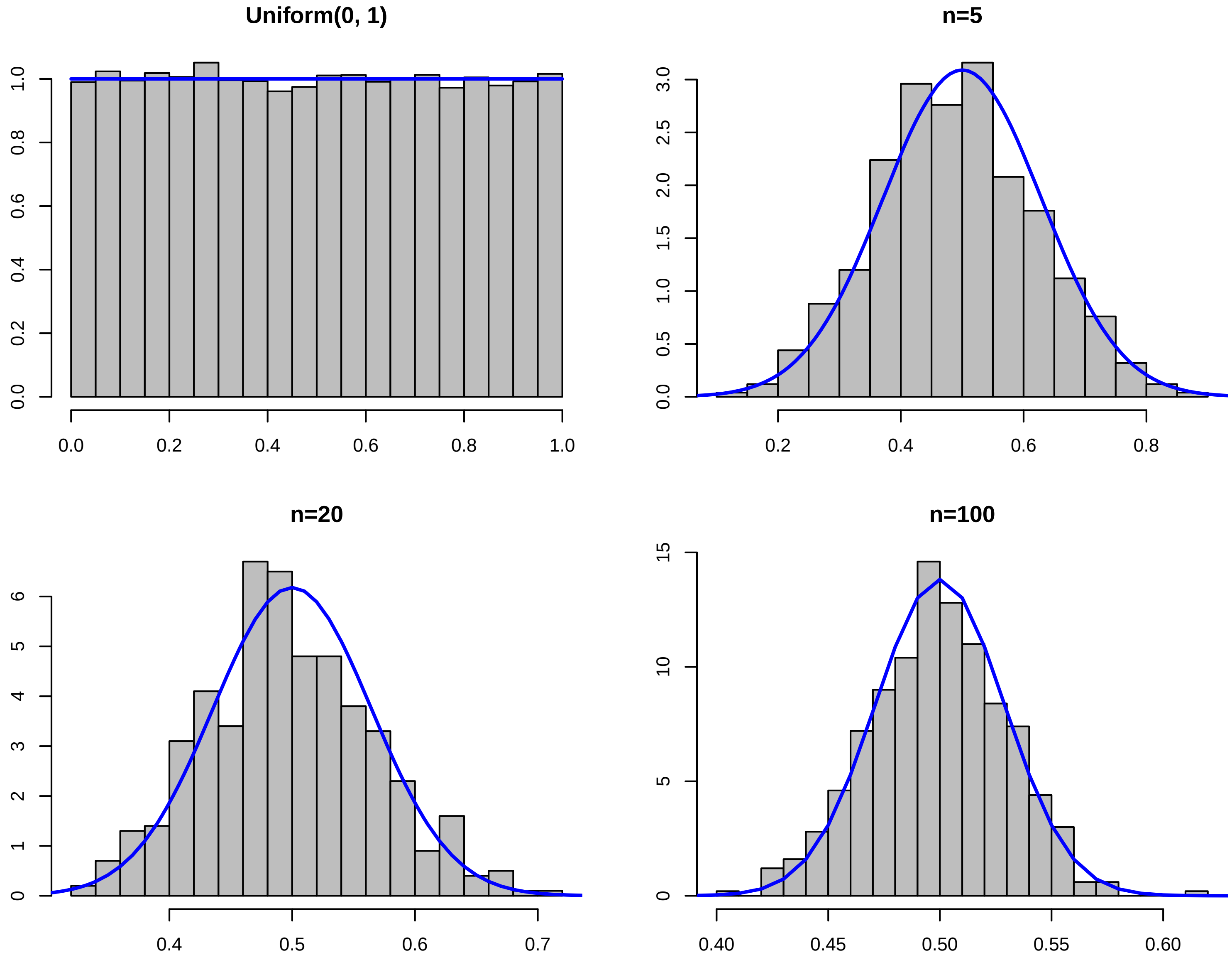
Note. The CLT is one of the reasons why normal distribution is the most useful and important distribution in statistics.

Example 4. If we take a sample $X_1, \cdots, X_n$ from $U(0,1)$, the standardized histogram will resemble the density function $f(x) = I_{(0,1)}(x)$, and the sample mean $\bar{X}_n = n^{-1} \sum_i X_i$ will be close to $\mu = EX_i = 0.5$, provided n is sufficiently large.

However the CLT implies $\bar{X}_n \approx N(0.5, 1/(12n))$ as $\text{Var}(X_i) = 1/12$. What does this mean?

If we take many samples of size n and compute the sample mean for each sample, we then obtain many sample means. The standardized histogram of those samples means resembles the PDF of $N(0.5, 1/(12n))$ provided n is sufficiently large.

```
> x <- runif(50000)  # generate 50,000 random numbers from U(0,1)
> hist(x, probability=T) # plot histogram of the 50,000 data
> z <- seq(0,1,0.01)
> lines(z,dunif(z)) # superimpose the PDF of U(0,1)
> x <- matrix(x, ncol=500)   # line up x into a 100x500 matrix
           # each column represents a sample of size 100
> par(mar=c(3,3,3,2),mfrow=c(2,2)) # plot 4 figures together
> meanx <- 1:500
> for(i in 1:500) meanx[i] <- mean(x[1:5,i])
        # compute the mean of the first 5 data in each column
> hist(meanx, probability=T, nclass=20, main='n=5')
> lines(z,dnorm(z,1/2,sqrt(1/(12*5))))
        # superimpose the PDF of N(.5, 1/(12*5))
> for(i in 1:500) meanx[i] <- mean(x[1:20,i])
> hist(meanx, probability=T, nclass=20, main='n=20')
> lines(z,dnorm(z,1/2,sqrt(1/(12*20))))
> for(i in 1:500) meanx[i] <- mean(x[1:60,i])
> hist(meanx, probability=T, nclass=20, main='n=60')
> lines(z,dnorm(z,1/2,sqrt(1/(12*60))))
> for(i in 1:500) meanx[i] <- mean(x[,i])
> hist(meanx, probability=T, nclass=20, main='n=100')
> lines(z,dnorm(z,1/2,sqrt(1/(12*100))))
```


Uniform(0, 1)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
n=5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.2
0.4
0.6
0.8
n=20
0
1
2
3
4
5
6
0.4
0.5
0.6
0.7
n=100
0
5
10
15
0.40
0.45
0.50
0.55
0.60

Example 5. Suppose a large box contains 10,000 poker chips distributed as follows

Values of chips	$5	$10	$15	$30
No. of chips	5000	3000	1000	1000

Take one chip randomly from the box, let X be its nomination. Then its probability function is

X	5	10	15	30
probability	0.5	0.3	0.1	0.1

Furthermore $\mu = EX = 10$ and $\sigma^2 = \mathrm{Var}(X) = 55$.

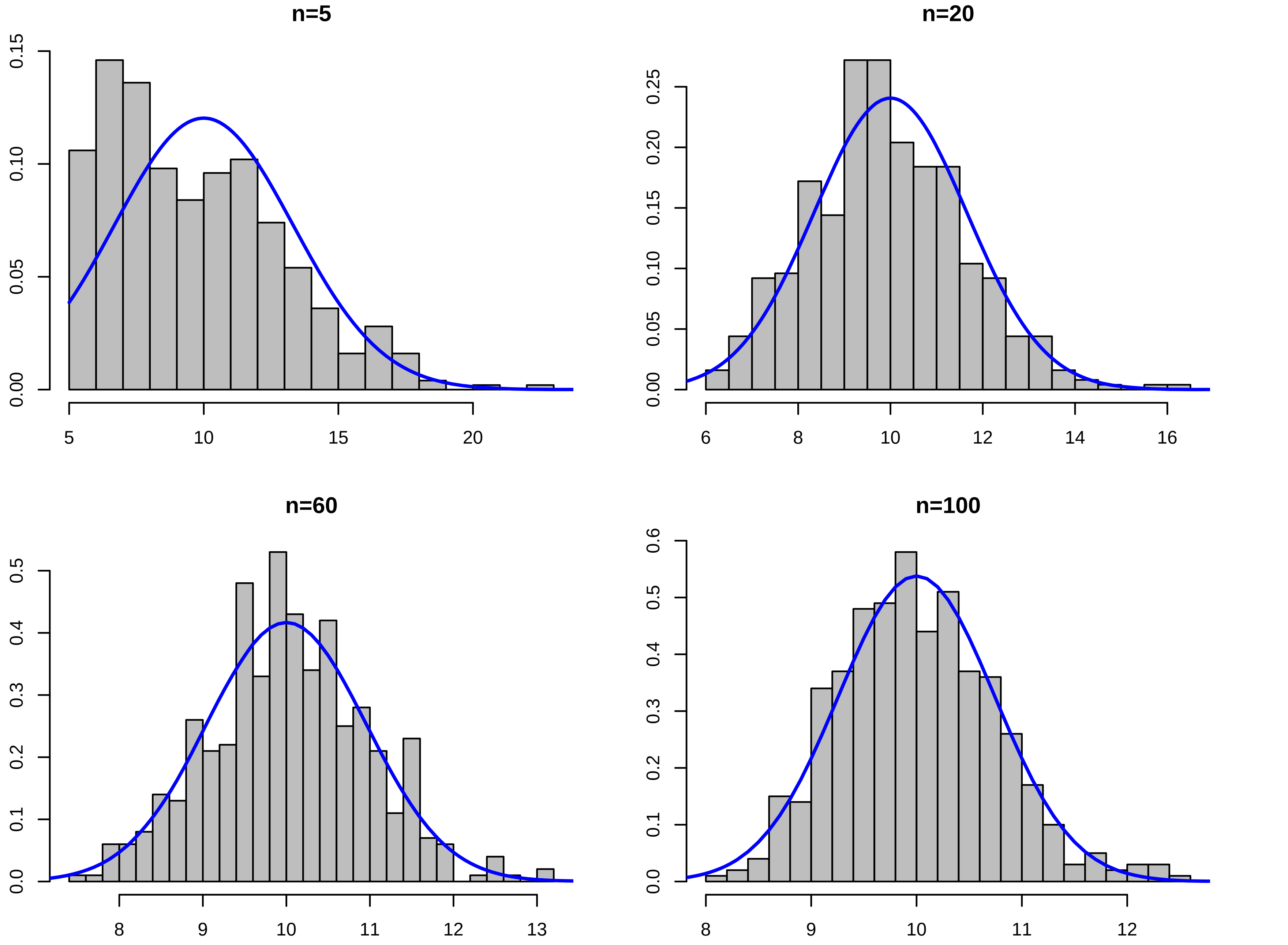
We draw 500 samples from this distribution, compute the sample means $\bar{X}_n$. When n is sufficiently large, we expect $\bar{X}_n \approx N(10, 55/n)$.

We create a plain text file ‘porkerChip.r’ as below, which illustrate the central limiting phenomenon for the samples from this simple distribution.

```
y<- runif(50000) # generate 50,000 U(0,1) random numbers
x<- y
for(i in 1:50000)
    if(y[i]<0.5) x[i]<-5 else {
        if(y[i]<0.8) x[i]<-10 else {
            ifelse(y[i]<0.9, x[i]<-15, x[i]<-30)
        }
    }   # By now x are random numbers from the required distribution
        # of the poker chips
cat("mean", mean(x), "\n")
cat("variance", var(x), "\n")

x <- matrix(x, ncol=500)  # line up x into a 100x500 matrix
                          # each column represents a sample of size 100
par(mar=c(3,3,3,2),mfrow=c(2,2)) # plot 4 figures together

meanx <- 1:500
```

```
z<-seq(5,25,0.1)

for(i in 1:500) meanx[i] <- mean(x[1:5,i])
        # compute the mean of the first 5 data in each column
hist(meanx, probability=T, main='n=5')
lines(z,dnorm(z,10,sqrt(55/5)))
       # draw N(10, 55/n) together with the histogram

for(i in 1:500) meanx[i] <- mean(x[1:20,i])
        # compute the mean of the first 20 data in each column
hist(meanx, probability=T, main='n=20')
lines(z,dnorm(z,10,sqrt(55/20)))

for(i in 1:500) meanx[i] <- mean(x[1:60,i])
        # compute the mean of the first 60 data in each column
hist(meanx, probability=T, main='n=60')
lines(z,dnorm(z,10,sqrt(55/60)))

for(i in 1:500) meanx[i] <- mean(x[,i])
        # compute the mean of the whole 100 data in each column
hist(meanx, probability=T, main='n=100')
```

```
lines(z,dnorm(z,10,sqrt(55/100)))
```


n=5
0.00
0.05
0.10
0.15
5
10
15
20
n=20
0.00
0.05
0.10
0.15
0.20
0.25
6
8
10
12
14
16
n=60
0.0
0.1
0.2
0.3
0.4
0.5
8
9
10
11
12
13
n=100
0.0
0.1
0.2
0.3
0.4
0.5
0.6
8
9
10
11
12

Example 6. Suppose $X_1, \cdots, X_n$ is an IID sample. A natural estimator for the population mean $\mu = EX_i$ is the sample mean $\bar{X}_n$. By the CLT, we can easily gauge the error of this estimation as follows:

$$\begin{aligned} P(|\bar{X}_n - \mu| > \epsilon) &= P\big(\sqrt{n}|\bar{X}_n - \mu|/\sigma > \sqrt{n}\epsilon/\sigma\big) \approx P\{|N(0,1)| > \sqrt{n}\epsilon/\sigma\} \\ &= 2P\{N(0,1) > \sqrt{n}\epsilon/\sigma\} = 2\{1 - \Phi(\sqrt{n}\epsilon/\sigma)\}. \end{aligned}$$

With ϵ, n given, we can find the value $\Phi(\sqrt{n}\epsilon/\sigma)$ from the table for standard normal distribution, *if we know* σ.

Remarks. (i) Let $\epsilon = 2\sigma/\sqrt{n} = 2 \times \mathrm{STD}(\bar{X}_n)$ (as $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$), $P(|\bar{X}_n - \mu| < 2\sigma/\sqrt{n}) \approx 2\Phi(2) - 1 = 0.954$. Hence

> *If one estimates μ by $\bar{X}_n$ and repeats it a large number times, about the 95% of times μ is within $2 \times \mathrm{STD}(\bar{X}_n)$-distance from $\bar{X}_n$.*

(ii) Typically $\sigma^2 = \mathrm{Var}(X_i)$ is unknown in practice. We estimate it using the sample variance

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

In fact it still holds that

$$\sqrt{n}(\bar{X}_n - \mu)/S_n \xrightarrow{D} N(0,1), \qquad \text{as } n \to \infty.$$

Similar to Example 6 above, we have now

$$P(|\bar{X}_n - \mu| > \epsilon) \approx 2\{1 - \Phi(\sqrt{n}\epsilon/S_n)\}$$

Let $\epsilon = S_n/\sqrt{n}$, $P(|\bar{X}_n - \mu| > \epsilon) \approx 2\{1 - \Phi(1)\} = 0.317$, or
$P(|\bar{X}_n - \mu| < S_n/\sqrt{n}) \approx 1 - 0.317 = 0.683$.

Let $\epsilon = 2S_n/\sqrt{n}$, we obtain:

$$P(|\bar{X}_n - \mu| < 2S_n/\sqrt{n}) \approx 0.954.$$

Hence

> *If one estimates μ by $\bar{X}_n$ and repeats it a large number times, about the 95% of times the true value is within $(2S_n/\sqrt{n})$-distance from $\bar{X}_n$.*

Standard Error: $\mathrm{SE}(\bar{X}_n) \equiv S_n/\sqrt{n}$ is called the standard error of the sample mean. Note

$$\mathrm{SE}(\bar{X}_n) = \left\{\frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right\}^{1/2}.$$

The Delta Method. If $\sqrt{n}(Y_n - \mu)/\sigma \xrightarrow{D} N(0,1)$ and g is a differentiable function and $g'(\mu) \neq 0$. Then

$$\frac{\sqrt{n}\{g(Y_n) - g(\mu)\}}{|g'(\mu)|\sigma} \xrightarrow{D} N(0,1).$$

Hence if $Y_n \approx N(\mu, \sigma^2/n)$, then $g(Y_n) \approx N(g(\mu), (g'(\mu))^2\sigma^2/n)$.

Example 7. Suppose $\sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{D} N(0,1)$ and $W_n = e^{\bar{X}_n} = g(\bar{X}_n)$ with $g(x) = e^x$. Since $g'(x) = e^x$, the Delta method implies $W_n \approx N(e^\mu, e^{2\mu}\sigma^2/n)$.

6.3 Monte Carlo methods

6.3.1 Basic Monte Carlo integration

The LLN may be interpreted as

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \int xf(x)dx$$

if $\{X_1, \cdots, X_n\}$ is a sample from the distribution with PDF f.

In general, for any function h, we apply the LLN to the sample $H_i \equiv h(X_i)$ $(i = 1, \cdots, n)$, leading to

$$\bar{H}_n \equiv \frac{1}{n}\sum_{i=1}^{n} h(X_i) \xrightarrow{P} E\{h(X_1)\} = \int h(x)f(x)dx. \qquad (4)$$

Monte Carlo integration method: generate a sample $\{X_1, \cdots, X_n\}$ from PDF f, then the integral on the RHS of (4) may be approximated by the mean $\bar{H}_n$.

To measure the accuracy of this Monte Carlo approximation, we may use the standard deviation $\sigma/\sqrt{n}$ (if we know $\sigma^2 = \text{Var}(H_1)$), or the standard error:

$$(\frac{1}{n(n-1)}\sum_{i=1}^{n}\{h(X_i) - \bar{H}_n\}^2)^{1/2}.$$

Example 8. (*Area of the quarter circle*) The area of a quarter of the unit circle is $\pi/4 = 0.7854$.

Suppose we do not know the answer. It can be written as

$$J \equiv \int_0^1 \sqrt{1 - x^2} dx.$$

However it is not obvious how to solve this integral. We provide a Monte Carlo solution. Let

$$h(x) = \sqrt{1 - x^2}, \quad f(x) = I_{(0,1)}(x).$$

Then f is the PDF of $U(0, 1)$ and

$$J = \int h(x) f(x) dx = E\{h(X)\},$$

where $X \sim U(0,1)$. Hence we generate a sample from $U(0,1)$ and estimate J by

$$\widehat{J} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{1-X_i^2}, \quad \mathrm{SE} = \Big\{\frac{1}{n(n-1)}\sum_{i=1}^{n}(\sqrt{1-X_i^2}-\widehat{J})^2\Big\}^{1/2}.$$

The STD of $\widehat{J}$ is $\sigma/\sqrt{n}$, where

$$\sigma^2 = \mathrm{Var}(\sqrt{1-X_1^2}) = E(1-X_1^2) - (\frac{\pi}{4})^2 = \frac{2}{3} - (\frac{\pi}{4})^2 = 0.0498.$$

The R-function ‘quartercircle.r’ below perform this Monte Carlo calculation. It is used to produce the table

n	1000	2000	4000	8000
$\widehat{J}$	.7950	.7834	.7841	.7858
STD	.0071	.0050	.0035	.0025
SE	.0072	.0050	.0036	.0025

R-function ‘quartercircle.r’:

```
quartercircle<-function(n)
        # This function calculates the area of the quarter circle
        # using Monte Carlo method
        # The true value is \pi/4 = 0.7854
        # n is the sample size
{
        x <- runif(n)
        h <- sqrt(1-x*x)
        list(quarterarea=mean(h), STD=sqrt(.0498/n),
             SE=sqrt(var(h)/n), SampleSize=n)
        # use 'list' to keep more than one outputs
}
```

You may call the function to perform the simulation:

```
> source("quartercircle.r")
> t=quartercircle(2000)
> summary(t)
            Length Class  Mode
quarterarea 1      -none- numeric
STD         1      -none- numeric
SE          1      -none- numeric
SampleSize  1      -none- numeric
> t
$quarterarea
[1] 0.7913048
$STD
[1] 0.00498999
$SE
[1] 0.004946009
$SampleSize
[1] 2000
> t$quarterarea
[1] 0.7913048
```

6.3.2 Composition (Sequential sampling)

Let $X \sim f_X(\cdot)$, $Y|X \sim f_{Y|X}(\cdot|X)$. To obtain

$$Y_1, \cdots, Y_n \sim_{iid} f_Y(\cdot) \equiv \int f_{Y|X}(\cdot|x) f_X(x) dx,$$

we may repeat the composition below n times:

Step 1. Draw X_i from $f_X(\cdot)$,
Step 2. Draw Y_i from $f_{Y|X}(\cdot|X_i)$.

Then $\{(X_i, Y_i), 1 \le i \le n\}$ are i.i.d. from the joint density

$$f_{X,Y}(x, y) = f_{Y|X}(y|x) f_X(x).$$

Hence $Y_1, \cdots, Y_n$ are i.i.d. from its marginal density $f_Y(\cdot)$.

Remarks.

(a) This method is applied when it is difficult to sample directly from $f_Y(\cdot)$.

(b) With $Y_1, \cdots, Y_n \sim_{iid} f_Y(y)$, we may estimate $E(Y)$ by $n^{-1}\sum_i \mathbf{Y}_i$. In general we estimate $E\{\psi(Y)\}$, for a known $\psi(\cdot)$, by

$$\bar{\psi} \equiv \frac{1}{n}\sum_{i=1}^{n} \psi(Y_i),$$

with the standard error

$$\frac{1}{\sqrt{n(n-1)}}\Big[\sum_{i=1}^{n}\{\psi(Y_i) - \bar{\psi}\}^2\Big]^{1/2}.$$

(c) The density function $f_Y(\cdot)$ may be estimated by

$$\widehat{f}_Y(y) = \frac{1}{n}\sum_{i=1}^{n} f_{Y|X}(y|X_i).$$

It also provides an estimate for EY: $\int y\widehat{f}_Y(y)dy$.

Example 9. Let $Y = X_1 + \cdots + X_T$, where $X_1, X_2, \cdots$ are IID Bernoulli(p), $T \sim$ Poisson(λ), and T and X_i's are independent. Then a sample from the distribution of Y can be drawn as follows:

(i) Draw $T_1, \cdots, T_n$ independently from Poisson(λ),

(ii) Draw $Y_i \sim \text{Bin}(T_i, p)$, $i = 1, \cdots, n$, independently.

Example 10. Mixture of Normal distributions:

$$p\, N(\mu_1, \sigma_1^2) + (1-p)\, N(\mu_0, \sigma_0^2), \quad p \in (0, 1),$$

(i.e. with PDF $\frac{p}{\sigma_1}\varphi(\frac{x-\mu_1}{\sigma_1}) + \frac{1-p}{\sigma_0}\varphi(\frac{x-\mu_0}{\sigma_0})$.)

A sample $X_1, \cdots, X_n$ can be drawn as follows:

(i) $I_1, \cdots, I_n \sim$ Bernoulli(p) independently,

(ii) $X_i \sim N(\mu_{I_i}, \sigma_{I_i}^2)$, $i = 1, \cdots, n$, independently.

Example 11. The lifetime X of a product follows the exponential distribution with mean $e^{1+U/4}$, where U is a quality index of the raw materials used in producing the product and $U \sim N(\mu, \sigma^2)$. Find the mean, variance and the PDF of X when $\mu = 1$ and $\sigma^2 = 2$.

As $X|U \sim Exp(e^{1+U/4})$ and $U \sim N(\mu, \sigma^2)$, we have

$$f_X(x) = \int f_{X|U}(x|u) f_U(u) du,$$
$$f_{X|U}(x|u) = e^{-(1+u/4)} \exp\{-xe^{-(1+u/4)}\} \quad \text{for } x > 0.$$

We use Monte Carlo simulation as follows:

1. Draw $U_1, \cdots, U_n$ from $N(\mu, \sigma^2)$
2. Draw X_i from $Exp(e^{1+U_i/4})$, $i = 1, \cdots, n$.

Then the estimated mean for X is $\bar{X}_n = n^{-1} \sum_i X_i$ with the standard error $\widehat{\sigma}/\sqrt{n}$, where

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

is an estimator for the variance of X. The estimated PDF is

$$\widehat{f}_X(x) = \frac{1}{n} \sum_{i=1}^{n} f_{X|U}(x|U_i) = \frac{1}{n} \sum_{i=1}^{n} e^{-(1+U_i/4)} \exp\{-x e^{-(1+U_i/4)}\}$$

We write R-function `lifetimeMeanVar` to simulate EX and $\mathrm{Var}(X)$, and `lifetimePDF` to produce the PDF f_X and also EX.

```
lifetimeMeanVar <- function(n, mu, sigma2) {
       u <- rnorm(n, mu, sqrt(sigma2))
             # generate n random numbers from N(mu, sigma2)
       x <- u
       for(i in 1:n) x[i] <- rexp(1, 1/exp(1+u[i]/4))
         # x[i] is a random number from Exponential
         # distribution with mean e^{1+u[i]/4}
       vx <- var(x)
       list(Mean=mean(x), Min=min(x), Max=max(x),
                   StandardError=sqrt(vx/n), Var=vx)
}
```

The function is saved in the file ‘lifetimeMeanVar.r’, we source it into R and produce the required results:

```
> source("lifetimeMeanVar.r")
> outcome <- lifetimeMeanVar(500,1,2)
> outcome$Mean
[1] 3.763913
> outcome$Min
[1] 0.02139847
> outcome$Max
[1] 50.12281
> outcome$StandardError
[1] 0.1906219
> outcome$Var
[1] 18.16836
```

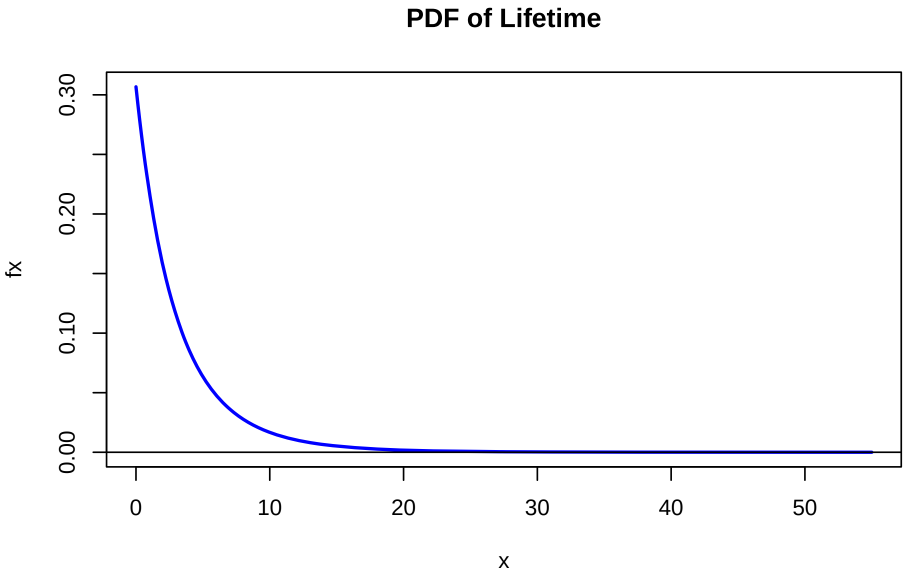
You may also try `summary(outcome)`.

The function `lifetimePDF` produces the PDF curve of X in the given range `(xmin, xmax)`. It also computes EX according to the estimated PDF.

```
lifetimePDF <- function(n,xmin,xmax,mu,sigma2) {
    u <- rnorm(n, mu, sqrt(sigma2))
    eu <- exp(-1-u/4)
    h <- (xmax-xmin)/400
    x <- seq(xmin, xmax, h)
    fx <- x
    for(i in 1:401) fx[i] <- mean(eu*exp(-x[i]*eu))
    m <- sum(x*fx*h)  # calculate the mean
    plot(x, fx, type='l', main="PDF of Lifetime")
    abline(0,0)   # abline(a,b) draw the straight line y=a+bx
    cat("Mean", m, "\n") # print out the mean
}  # Definition of function lifetimePDF' ends here
```

Source it into R to produce the required results:

```
> source("lifetimePDF.r")
> lifetimePDF(500,0,55,1,2)
> Mean 3.779971
```



6.3.3 Importance sampling

Let us consider the composition method discussed in section 6.3.2: To obtain an estimate for

$$f_Y(\cdot) = \int f_{Y|X}(\cdot|x) f_X(x) dx$$

or to obtain a sample from $f_Y(\cdot)$, we need to draw a sample $\{X_1, \cdots, X_n\}$ from $f_X(\cdot)$.

However sometimes we cannot directly sample from $f_X(\cdot)$. Importance sampling offers an indirect way to achieve this goal via an appropriately selected PDF $p(\cdot)$.

Let $p(\cdot)$ be a density satisfying:

(a) the support of p contains the support of f_X,
i.e. $p(\mathbf{x}) = 0$ implies $f_X(\mathbf{x}) = 0$, and
(b) it is easy to sample from $p(\cdot)$.

Importance sampling method for approximating

$$J \equiv E\{h(X)\} = \int h(x) f_X(x) dx$$

(i) Draw $X_1, \cdots, X_n \sim_{i.i.d.} p(\cdot)$
(ii) Compute the estimator

$$\widehat{J} = \sum_{i=1}^{n} w_i h(X_i) \Big/ \sum_{i=1}^{n} w_i,$$

where $w_i = f_X(X_i)/p(X_i)$.

Importance sampling places weights greater than 1 on the regions where $f_X(x) > p(x)$, and downweights the regions where $f_X(x) < p(x)$.

Choice of $p(\cdot)$: as close to $f_X(\cdot)$ as possible among all PDF satisfying (a) and (b) in the previous page.

The standard error of $\widehat{J}$ is

$$\Big[\sum_{i=1}^{n} \{h(X_i) - \widehat{J}\}^2 w_i^2\Big]^{1/2} \Big/ \sum_{i=1}^{n} w_i.$$

which is inflated when $p(\cdot)$ poorly approximates $f_X(\cdot)$.

Note. $\sum_{i=1}^{n} w_i$ can be viewed as a version of the effective sample size in the importance sampling. When $p(\cdot)$ differs substantially from $f_X(\cdot)$, all w_i are small. Hence the sampling is inefficient.

Remark. In the above calculation, we may *replace the PDF $f_X(\cdot)$ by $g(\cdot) \equiv C_0 f_X(\cdot)$*, where $C_0 > 0$ is an unknown constant. The algorithm stays the same but with the weights

$$w_i = g(X_i)/p(X_i).$$

For example, $f_X(x) = C_0^{-1} e^{-x^2/(|x|+2)}$, where the normalised constant $C_0 = \int e^{-x^2/(|x|+2)} dx$ is not easy to compute. In this case we may use $g(x) = e^{-x^2/(|x|+2)}$ instead of $f_X(x)$ in importance sampling.

Proof of Remark. By the LLN, as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} w_i \xrightarrow{P} \int \frac{g(x)}{p(x)} p(x)dx = \int g(x)dx = C_0 \int f_X(x)dx = C_0,$$

and

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} w_i h(X_i) &\xrightarrow{P} \int \frac{g(x)}{p(x)} h(x)p(x)dx \\ = & \int g(x)h(x)dx = C_0 \int f_X(x)h(x)dx = C_0 E\{h(X)\}.\end{aligned}$$

Hence, by Slutzky's theorem,

$$\sum_{i=1}^{n} w_i h(X_i) \Big/ \sum_{i=1}^{n} w_i \xrightarrow{P} E\{h(X)\}.$$

Application to sequential sampling: $f_Y(\cdot) = \int f_{Y|X}(\cdot|x) f_X(x) dx$

(i) Draw $X_1, \cdots, X_N \sim_{i.i.d.} p(\cdot)$,

(ii) Draw $Y_i \sim f_{Y|X}(\cdot|X_i)$, $i = 1, \cdots, n$, independently.

Let $w_i = g(X_i)/p(X_i)$ and $\mu_y = E(Y)$, then

$$\widehat{f}_Y(y) = \sum_{i=1}^{n} w_i f_{Y|X}(y|X_i) \Big/ \sum_{i=1}^{n} w_i,$$

$$\widehat{\mu}_y = \sum_{i=1}^{n} w_i Y_i \Big/ \sum_{i=1}^{n} w_i,$$

which is guaranteed by the fact $(X_i, Y_i) \sim_{i.i.d.} p(x) f_{Y|X}(y|x)$.

Note. Importance sampling does not yield correct samples, as

$$X_i \nsim f_X(\cdot), \qquad Y_i \nsim f_Y(\cdot)$$

Example 11 (Continue). Suppose now the quality index of the raw materials U follows a generalised normal distribution with PDF

$$f_U(u) \propto \exp\left\{-\frac{1}{2}\left|\frac{u-\mu}{\sigma}\right|^{\nu}\right\} \equiv g(u)$$

where $\nu > 0$ is a constant. Recall

$$f_{X|U}(x|u) = e^{-(1+u/4)}\exp\{-xe^{-(1+u/4)}\} \quad \text{for } x > 0.$$

We adopt an importance sampling scheme as follows:

1. Draw $U_1, \cdots, U_n$ from $N(\mu, \sigma^2)$, compute the weight $w_i = g(U_i)/\phi(\frac{U_i-\mu}{\sigma})$, where ϕ denotes the PDF of $N(0,1)$.
2. Draw X_i from $Exp(e^{1+U_i/4})$, $i = 1, \cdots, n$.

Then the estimated mean for X is

$$\bar{X}_n = \sum_{i=1}^{n} w_i X_i \Big/ \sum_{i=1}^{n} w_i .$$

The estimated PDF is

$$\widehat{f}_X(x) = \frac{\sum_{i=1}^{n} w_i f_{X|U}(x|U_i)}{\sum_{i=1}^{n} w_i} = \frac{\sum_{i=1}^{n} w_i e^{-(1+U_i/4)} \exp\{-x e^{-(1+U_i/4)}\}}{\sum_{i=1}^{n} w_i}.$$

The R-function `lifetimeMeanIS` implements the above scheme for calculating EX:

```r
lifetimeMeanIS <- function(n, mu, sigma2, nu) {
u=rnorm(n, mu, sqrt(sigma2)) #generate n numbers from N(mu, sigam2)
w=exp(-0.5*abs((u-mu)/sqrt(sigma2))^nu)/dnorm((u-mu)/sqrt(sigma2))
            # compute the weights w_i
x<-u
for(i in 1:n) x[i]<-rexp(1, 1/exp(1+u[i]/4))
list(Mean=sum(x*w)/sum(w), Min=min(x), Max=max(x))
}
```

The results for $\mu = 1$, $\sigma^2 = 2$ and $\nu = 0.5$ or 3 are as follows:

```
> source("lifetimeMeanIS.r")
> lifetimeMeanIS(5000,1,2,0.5)
$Mean
[1] 0.8827147
$Min
[1] 0.0003652474
$Max
[1] 57.21467
> lifetimeMeanIS(10000,1,2,3)
$Mean
[1] 1.616474
$Min
[1] 0.00125402
$Max
[1] 56.77547
```

The R-function `lifetimePDF.IS` implements the above scheme for estimating PDF f_X and $E(X)$:

```
lifetimePDF.IS <- function(n,xmin,xmax,mu,sigma2,nu) {
  u <- rnorm(n, mu, sqrt(sigma2))
  Eu <- exp(-(1+u/4))   # Eu=e^{-(1+u/4)}
  w=exp(-0.5*abs((u-mu)/sqrt(sigma2))^nu)/dnorm((u-mu)/sqrt(sigma2))
          # compute the weights w_i
  sumw <- sum(w)
  h <- (xmax-xmin)/400
  x <- seq(xmin, xmax, h)
  fx <- x
  t <- 1:n
  m <- 0
  for(i in 1:401) {
      t <- Eu*exp(-x[i]*Eu)
    # t = PDF of  Exp(1/e^(1+u/4)) at x=x[i] --- THIS IS MORE
      fx[i] <- sum(t*w)/sumw
      m <- m+x[i]*fx[i]*h  # calculate the mean
  }
  plot(x, fx, type='l', main="PDF of Lifetime")
  abline(0,0)   # abline(a,b) draw the straight line y=a+bx
  cat("Mean", m, "\n")  # print out the mean
}
```

You may source it in, and try `lifetimePDF.IS(5000,0,60,1,2,0.5)` etc.

Importance of using appropriate sampling distributions

An alternative measure for the effective sample size (ESS) is defined as $n/\{1 + cv(w)\}$, where $cv(w)$ is the sample coefficient of variation of the weights

$$cv(w) = \Big\{\frac{1}{n-1}\sum_{i=1}^{n}(w_i - \bar{w})^2\Big\}^{1/2}\Big/\bar{w}, \qquad \bar{w} = \frac{1}{n}\sum_{i=1}^{n} w_i.$$

We illustrate the importance of choosing 'correct' $p(\cdot)$ in the example below.

Example 12. Estimate μ for $N(\mu, 1)$ based on the importance sampling method using $N(0, 1)$ as the sampling distribution $p(\cdot)$. The table below is produced by R-function `effectN` with $n = 1000$.

μ	0	1	2	3	4	5
Estimated μ	-0.022	1.026	1.756	2.806	2.873	3.325
ESS	1000	448.9	246.1	113.4	65.7	33.8

```r
effectN=function(n, mu) {
x=rnorm(n)
w=dnorm(x,mu,1)/dnorm(x)  # sampling weights
muhat=mean(w*x)/mean(w)  # estimate for mu by importance sampling
ess=n/(1+sqrt(var(w))/mean(w)) # effective sample size
list(SampleSize=n, Mean=mu, EstimatedMean=muhat, ESS=ess)
}
```

Chapter 7. Introduction to Statistical Inference

7.1. What is Statistics: a scientific subject on collecting and analyzing data.

collecting: designing experiments/questionnaires, designing sampling schemes, administration of data collection

analyzing: estimation, testing and forecasting

Statistics is an application-oriented subject, is particularly useful or helpful in answering questions such as:

Those questions are difficult to study in laboratory, and admit no self-evident axioms.

What to learn in Statistics: **basic ideas**, methods (including computation) and theory.

Some guidelines for learning/applying statistics:

- Understand what data say in each specific context. All the methods are just tools to help to understand data
- Concentrate on what to do and why, rather than concrete calculation and graphing
- It may take a while to catch the basic idea of statistics – Keep thinking!!!

7.2 Population, Sample and Parametric Models

Two practical situations:

- A new type of tyre was designed to increase the lifetime. The manufacturer tested 120 new tyres and obtained the average lifetime (over those 120 tyres) 35,391 miles. So it claims that the mean lifetime of the new tyres is 35,391 miles.

- A newspaper sampled 1000 potential voters, and 350 of them were Democratic party supporters. It claims that the proportion of the Democratic voters in the whole Country is 350/1000=35%.

In both cases, the conclusion is drawn on a **population** (i.e. all the objects concerned) based on the information from a **sample** (i.e. a subset of population).

In the first case, it is impossible to measure the whole population. For the second case, it is not economic to measure the whole population. Therefore, **errors are inevitable**!

Population is an entire set of the objects concerned, and those objects are typically represented by some numbers. We do not know the entire population in practice.

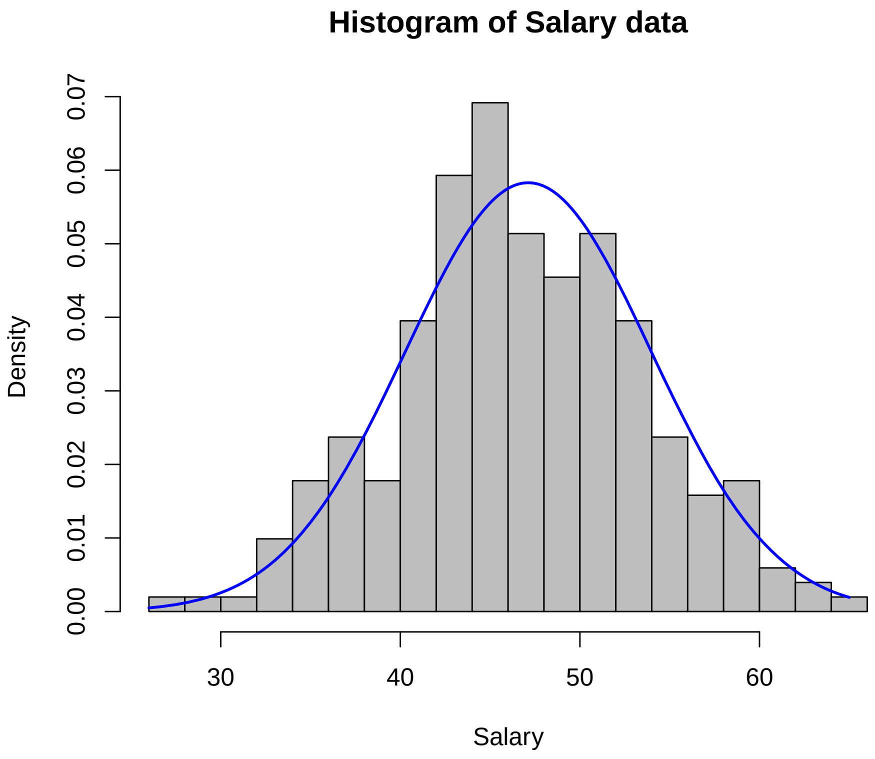
For the tyre example, the population consists of the lifetimes of all the tyres, including those to be produced in the future.

For the opinion poll examples, the population consists of many ‘1’ and ‘0’, where each ‘1’ represents a voter for Democratic party, and each ‘0’ represents a voter for other parties.

A **Sample** is a (randomly) selected subset of a population, and is a set of known data in practice.

Population is unknown. We represent a population by a probability distribution.

```
> jobs <- read.table("Jobs.txt", header=T, row.name=1)
> dim(jobs)
[1] 253   5
> mean(jobs[,4]); var(jobs[,4])
[1] 47.12648
[1] 46.83315
> hist(jobs[,4], probability=T, nclass=15, xlab="Salary",
          main="Histogram of Salary data")
> range(jobs[,4])
[1] 26 65
> x <- seq(26, 65, 0.1)
> lines(x, dnorm(x, 47.12648, sqrt(46.83315)))
      # superimpose the PDF of N(47.12648, 46.83315)
```



The blue curve is the PDF of $N(\bar{X}_n, S_n^2)$.

$n = 253$, $\bar{X}_n = 47.126$, and $S_n^2 = 46.833$, $S_n = 6.843$.

$\bar{X}_n \pm S_n = (40.283,\ 53.969)$

171 points are in this interval:

$171/253 = 67.58\%.$

$\bar{X}_n \pm 1.96 S_n = (33.714,\ 60.538)$

242 points are in this interval:

$242/253 = 95.65\%$

Suggesting $N(\bar{X}_n, S_n^2)$!!!

Parametric Models. For a given problem, we typically assume a population to be a probability distribution $F(\cdot;\theta)$, where the form of distribution F is known (such as normal, Poisson etc), and θ denotes some unknown characteristics (such as mean, variance etc) and is called a parameter. Such an assumed distribution is often called a parametric model.

For the tyre lifetime example, the population may be assumed to be $N(\mu,\sigma^2)$ with $\theta = (\mu,\sigma^2)$, where μ is the 'true' lifetime. Let

$$X = \text{the lifetime of a tyre.}$$

Then $X \sim N(\mu,\sigma^2)$.

For the opinion poll example, the population is a Bernoulli distribution:

$$P(X = 1) = P(\text{ a Democratic voter }) = \pi,$$

$$P(X = 0) = P(\text{ a Republican voter }) = 1 - \pi,$$

where

π = the proportion of Democratic supporters in the USA
= the probability of a voter to be a Democratic supporter

A Sample: a set of data or random variables? – A Duality

A sample of size n: $\{X_1, \cdots, X_n\}$, is also called a random sample. It consists of n concrete numbers in a practical problem.

The word 'random' captures the character that samples (of the same size) taken by different people or at different times may be different, as they are different subsets of a population.

Furthermore, a sample is also viewed as n independent and identically distributed (i.i.d.) random variables, when we assess the performance of a statistical method.

For the tyre lifetime example, the sample (of size $n = 120$) used gives the sample mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 35,391$$

A different sample may give a different sample mean, say, 36,721.

Question: Is the sample mean $\bar{X}$ a good estimator for the unknown ‘true’ lifetime μ?

Obviously we cannot use the concrete number 35,391 to assess how good this estimator is, as a different sample may give a different average value, say, 36,721.

Key idea: By treating $X_1, \cdots, X_n$ as random variables, $\bar{X}$ is also a random variable. If the distribution of $\bar{X}$ concentrates closely around (unknown) μ, $\bar{X}$ is a good estimator for μ.

Statistic. Any known function of a random sample is called a statistic.

Statistic is used for statistical inference such as estimation, testing etc.

Example. Let $X_1, \cdots, X_n$ be a sample from population $N(\mu, \sigma^2)$. Then

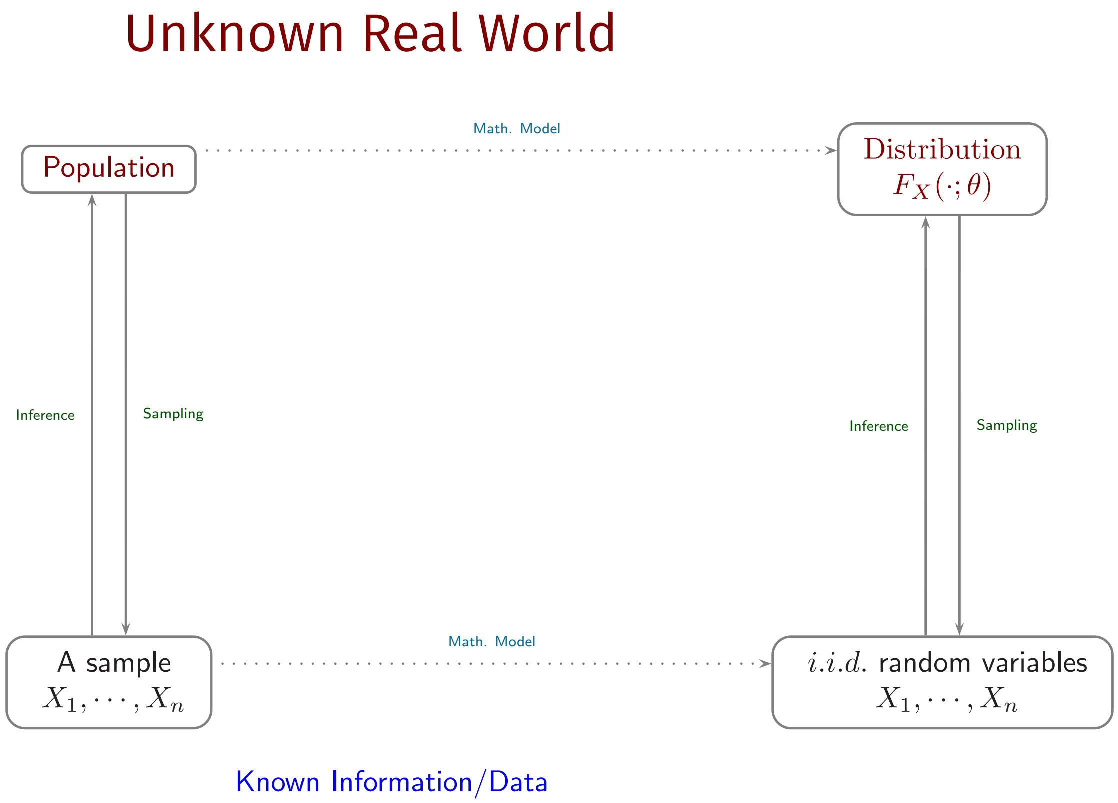
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad X_1 + X_n^2, \quad \sin(X_3) + 6$$

are all statistics. But

$$(X_1 - \mu)/\sigma$$

is not a statistic, as it depends on unknown quantities μ and σ^2.

Note. It is often to denote a random sample as $x_1, \cdots, x_n$, indicating they are n concrete numbers. They are seen as a realization or an instance of n i.i.d. random variables $X_1, \cdots, X_n$. But we do not make this difference, as it makes statements laboursome from time to time.



θ is called a **parameter**.

A known function of $X_1, \cdots, X_n$ is called a **statistic**.

Difference between Probability and Statistics

Probability: a mathematical subject

Statistics: an application oriented subject (which uses Probability heavily)

Example. Let

$$X = \text{No. of StatsI lectures attended by a student}$$

Then $X \sim \text{Bin}(17, p)$, i.e.

$$P(X = k) = \frac{17!}{k!(17-k)!} p^k (1-p)^{17-k}, \qquad k = 0, 1, \cdots, 17.$$

Probability questions: treating p as known

- what is E(X)? (the average lectures attended)
- what is $P(X \geq 14)$? (the proportion of the students attending at least 14 lectures)
- what is $P(X \leq 8)$? (the proportion of the students attending fewer than the half of lectures)

Statistics questions:

- what is p? (the average attendance rate)
- Is p not smaller than 0.9?
- Is p smaller than 0.5?

7.3 Fundamental concepts in statistical inference

Let $X_1, \cdots, X_n$ be a sample from a population $F(\cdot, \theta)$. Most inference problems can be identified as one of the three types: *(point) estimation, confidence sets* and *hypothesis testing* for parameter θ.

7.3.1 Point estimation: Provide a single "best guess" of θ, based on observations $X_1, \cdots, X_n$. Formally we may write

$$\widehat{\theta} \equiv \widehat{\theta}_n = g(X_1, \cdots, X_n)$$

as a point estimator for θ, where $g(X_1, \cdots, X_n)$ is a statistic.

For example, a natural point estimator for the mean $\mu = EX_1$ is the sample mean $\widehat{\mu} = \bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$.

Remark. Parameters to be estimated are unknown constants. Their estimators are viewed as r.v.s, although in practice $\widehat{\theta}$, $\widehat{\mu}$ admit some concrete values.

A good estimator should make $|\widehat{\theta} - \theta|$ as small as possible. However

(i) θ is unknown,
(ii) the value of $\widehat{\theta}$ changes with the sample observed.

Hence we seek for an estimator $\widehat{\theta}$ which makes the MSE as small as possible **for all possible values of** θ.

The **mean square error** of the estimator $\widehat{\theta}$ is defined as

$$\text{MSE}_\theta(\widehat{\theta}) = E_\theta\{(\widehat{\theta} - \theta)^2\} = \{\text{Bias}_\theta(\widehat{\theta})\}^2 + \text{Var}_\theta(\widehat{\theta}), \tag{5}$$

where $\text{Bias}_\theta(\widehat{\theta}) = E_\theta(\widehat{\theta}) - \theta$ is called the bias.

When $\text{Bias}_\theta(\widehat{\theta}) = 0$ for all possible values of θ, $\widehat{\theta}$ is called an *unbiased estimator*.

Note. The subscript 'θ' in E_θ etc indicates that the expectation etc are taken with θ being the true value.

The standard error of the estimator $\widehat{\theta}$: $\mathrm{SE}(\widehat{\theta}) = \{\mathrm{Var}_{\widehat{\theta}}(\widehat{\theta})\}^{1/2}$

Note. The standard deviation of $\widehat{\theta}$ $\{\mathrm{Var}_{\theta}(\widehat{\theta})\}^{1/2}$ may depend on the unknown θ. The standard error $\mathrm{SE}(\widehat{\theta})$ is known, and is an estimator for the standard deviation of $\widehat{\theta}$.

Example 1. Let $Y_1, \cdots, Y_n$ be a sample from Bernoulli(p). Let $\widehat{p} \equiv \widehat{p}_n = \bar{Y}_n = \sum_i Y_i / n$. Then

$$E(\widehat{p}) = \frac{1}{n}\sum_{i=1}^{n} EY_i = p, \quad \mathrm{Var}(\widehat{p}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(Y_i) = \frac{p(1-p)}{n}.$$

Therefore $\widehat{p}_n$ is an unbiased estimator for p with the standard deviation $\sqrt{p(1-p)/n}$, and SE($\widehat{p}$) = $\sqrt{\widehat{p}(1-\widehat{p})/n}$.

For example, if $n = 10$ and $\bar{Y}_n = 0.3$, we have $\widehat{p} = 0.3$ and SE($\widehat{p}$) = 0.1449 while the standard deviation of $\widehat{p}$ is $\sqrt{p(1-p)/10}$ unknown.

Proof of (5).

$$\begin{aligned}\mathrm{MSE}(\widehat{\theta}) &= E[\{(\widehat{\theta} - E\widehat{\theta}) + (E\widehat{\theta} - \theta)\}^2] \\ &= E\{(\widehat{\theta} - E\widehat{\theta})^2\} + (E\widehat{\theta} - \theta)^2 + 2(E\widehat{\theta} - \theta)E(\widehat{\theta} - E\widehat{\theta}) \\ &= \mathrm{Var}(\widehat{\theta}) + \mathrm{Bias}(\widehat{\theta})^2 + 0.\end{aligned}$$

Consistency. $\widehat{\theta}_n$ is a consistent estimator for θ if $\widehat{\theta}_n \xrightarrow{P} \theta$ as $n \to \infty$.

The consistency is a natural condition for a reasonable estimator, as $\widehat{\theta}_n$ should converge to θ if we have infinity amount of information. Therefore a non-consistent estimator should not be used in practice!

If $\text{MSE}(\widehat{\theta}_n) \to 0$, $\widehat{\theta} \xrightarrow{m.s.} \theta$. Therefore $\widehat{\theta} \xrightarrow{P} \theta$, i.e. $\widehat{\theta}$ is a consistent estimator for θ.

In Example 1 above, $\text{MSE}(\widehat{p}) = \text{Var}(\widehat{p}) = p(1-p)/n \to 0$. Hence $\widehat{p}$ is consistent.

Asymptotic Normality. An estimator $\widehat{\theta}_n$ is asymptotically normal if

$$(\widehat{\theta} - \theta)/\mathrm{SE}(\widehat{\theta}) \xrightarrow{D} N(0, 1).$$

Remark. Many good estimators such as MLE, LSE and MME are asymptotically normal under some mild conditions, such as finite moments and smooth likelihood function (as function of parameters)

7.3.2 Confidence sets

A point estimator is simple to construct and to use. But it is not very informative. For example, it does not reflect the uncertainty in the estimation.

Confidence Interval is the most commonly used confidence set, is more informative than a point estimator.

Example 2. Let us start with a simple example. A random sample $X_1, \cdots, X_n$ are drawn from $N(\mu, 1)$. Then $\sqrt{n}(\bar{X} - \mu) \sim N(0, 1)$. Hence

$$P(-1.96 \le \sqrt{n}(\bar{X} - \mu) \le 1.96) = 0.95,$$

or

$$P(\bar{X} - 1.96/\sqrt{n} < \mu < \bar{X} + 1.96/\sqrt{n}) = 0.95.$$

So a 95% confidence interval for μ is

$$(\bar{X} - 1.96/\sqrt{n},\ \bar{X} + 1.96/\sqrt{n}).$$

Suppose $n = 4$, $\bar{X} = 2.25$. Then a 95% C.I. is $(2.25 - 0.98, 2.25 + 0.98) = (1.27, 3.23)$.

Question: what is $P(1.27 < \mu < 3.23)$? — Note μ is a unknown constant!

Answer: $(1.27, 3.23)$ is one instance of the **random interval** $(\bar{X} - 0.98,\ \bar{X} + 0.98)$ which covers μ with probability 0.95.

If one draw 10,000 samples, with size $n = 4$ each, to construct 10,000 intervals of the form $(\bar{X} - 0.98,\ \bar{X} + 0.98)$, about 9,500 intervals cover the true value of μ.

Definition. If $L \equiv L(X_1, \cdots, X_n)$ and $U \equiv U(X_1, \cdots, X_n)$ are two statistics for which

$$P\{L(X_1, \cdots, X_n) < \theta < U(X_1, \cdots, X_n)\} = 1 - \alpha,$$

$(L,\ U)$ is called a $100(1-\alpha)\%$ *confidence interval* for θ.

Remark. $1 - \alpha$ is called the confidence level, which is usually set at 0.90, 0.95 or 0.99. Naturally for given α, we shall search for the interval with the shortest length $U - L$, which gives the most accurate estimation.

Approximate confidence interval based on an asymptotically normal estimator: If $(\widehat{\theta} - \theta)/\text{SE}(\widehat{\theta}) \xrightarrow{D} N(0, 1)$. Then $\widehat{\theta} \pm Z_{\alpha/2}\text{SE}(\widehat{\theta})$ is an approximate $1 - \alpha$ confidence interval for θ, where Z_α is the top-α point of $N(0, 1)$, i.e. $P\{N(0, 1) > Z_\alpha\} = \alpha$.

For $\alpha = 0.05,\ Z_{\alpha/2} = 1.96 \approx 2$, one of the most used 95% confidence interval is

$$\widehat{\theta} \pm 2 \times \text{SE}(\widehat{\theta}) = \left(\widehat{\theta} - 2 \times \text{SE}(\widehat{\theta}),\ \widehat{\theta} + 2 \times \text{SE}(\widehat{\theta})\right).$$

Example 1 (continue). Let $Y_1, \cdots, Y_n$ be a sample from Bernoulli(p). Let $\widehat{p} \equiv \widehat{p}_n = \bar{Y}_n = \sum_i Y_i / n$. By the CLT, $(\widehat{p}_n - p)/\mathrm{SE}(\widehat{p}_n) \sim N(0,1)$ asymptotically. Hence an approximate $1 - \alpha$ confidence interval for p is

$$\widehat{p}_n \pm Z_{\alpha/2}\mathrm{SE}(\widehat{p}_n) = \widehat{p}_n \pm Z_{\alpha/2}\sqrt{\widehat{p}_n(1 - \widehat{p}_n)/n}.$$

For example, if $n = 10$, $\widehat{p}_n = 0.3$, an approximate 95% confidence interval for p is

$$0.3 \pm 2\sqrt{0.3(1 - 0.3)/10} = 0.3 \pm 0.145 = (0.155,\ 0.445).$$

However if now $n = 100$ and $\widehat{p}_n = 0.3$, an approximate 95% confidence interval for p is

$$0.3 \pm 2\sqrt{0.3(1 - 0.3)/100} = 0.3 \pm 0.046 = (0.254,\ 0.346).$$

Remark. The point estimator $\widehat{p}_n$ unchanged with $n = 10$ or 100. However the confidence interval is much shorter when $n = 100$, giving much more accurate estimation.

7.3.3 Hypothesis testing: We start with some default statement – called a null hypothesis denoted by H_0. We ask if the date provide significant evidence to reject the null hypothesis.

For example we may test if a coin is fair by using the hypothesis $H_0 : p = 0.5$.

Remark. (i) Estimation and testing address different needs in practice.

(ii) A statistical test often takes binary decision: '*reject* H_0' or '*not reject* H_0'. However technically a testing problem is more complex that an estimation problem.

7.4 Nonparametric models and Empirical distribution functions

The outcome of a statistical inference depends on two factors: *data* and *assumption*.

The data are objective, while the assumption is more subjective. We would like to let *data speak* as much as possible.

The classical statistical inference is typically based on an assumption of a parametric model: $X_1, \cdots, X_n$ is a sample from the distribution $F(\cdot,\ \theta)$, where the form of the distribution F is known (such as Normal, Exponential etc), and the parameter θ is unknown. The inference is on either estimation or testing of parameter θ.

Nonparametric model: let $X_1, \cdots, X_n$ is a sample from a distribution F belong to a class of distributions $\mathcal{F}$. For example, $\mathcal{F}$ may consist of all continuous distributions on $(-\infty, \infty)$. The statistical inference is either to estimate or to test some characteristics of F, or F itself.

Parametric models limit the tasks of statistical inference. It facilitates more efficient inference **if** the assume parametric form is correct. Nonparametric models impose less model-bias, however its statistical inference is more challenging.

Empirical Distribution Functions. Let $X_1, \cdots, X_n$ be a sample from CDF F. A natural estimator for F is defined as

$$\widehat{F}(x) = \frac{\text{No. of } X_i's \text{ not greater than } x}{n} = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x),$$

where $I(A) = 1$ if A occurs, and 0 otherwise. $\widehat{F}(\cdot)$ is a well-defined CDF, and is called the empirical distribution function.

Note $I(X_i \le x)$ is a sequence of Bernoulli r.v.s with $p = F(x)$. Hence,

$$E\{\widehat{F}(x)\} = F(x), \quad \text{Var}\{\widehat{F}(x)\} = F(x)\{1 - F(x)\}/n,$$

and $\widehat{F}(x) \xrightarrow{m.s.} F(x)$. In fact it also holds that

$$\sup_x |\widehat{F}(x) - F(x)| \xrightarrow{P} 0, \quad P\{\sup_x |\widehat{F}(x) - F(x)| > \epsilon\} \le 2e^{-2n\epsilon^2}.$$

Chapter 8. Nonparametric bootstrap

Bootstrap is a computational method for estimating standard errors and confidence intervals.

Let $X_1, \cdots, X_n \sim_{iid} F$. We use statistic

$$T = g(X_1, \cdots X_n)$$

for inference (i.e. estimation or testing). It is important to know, e.g. the standard deviation or the standard error of T.

Bootstrap idea: Let $\widehat{F}_n(x) = n^{-1} \sum_i I(X_i \leq x)$.

$$\begin{array}{rl} \text{Real world: } F & \longrightarrow \; X_1, \cdots, X_n \; \longrightarrow \; T = g(X_1, \cdots X_n) \\ \text{Bootstrap world: } \widehat{F}_n & \longrightarrow \; X_1^*, \cdots, X_n^* \; \longrightarrow \; T^* = g(X_1^*, \cdots X_n^*) \end{array}$$

Although we do not know F, $\widehat{F}_n$ is known. Therefore we know the distribution of T^* (in principle), which is taken as an approximation for the distribution of T. We compute the distribution of T^* by simulation.

8.1 Bootstrap variance estimation

Suppose we need to know variance $v = \mathrm{Var}(T) = \mathrm{Var}\{g(X_1, \cdots X_n)\}$. The bootstrap scheme below provides an estimator v^* for v.

---

1. Draw $X_1^*, \cdots X_n^*$ independently from $\widehat{F}_n$.
2. Compute $T^* = g(X_1^*, \cdots X_n^*)$.
3. Repeat Steps 1 & 2 B times, to obtain $T_1^*, \cdots, T_B^*$.
4. Compute the sample variance $v^* = (B-1)^{-1} \sum_{1 \le i \le B} (T_i^* - \bar{T}^*)^2$,
where $\bar{T}^* = B^{-1} \sum_{1 \le i \le B} T_i^*$.

Remark. Step 1 can be easily implemented in R. Let `X` be n-vector $(X_1, \cdots, X_n)$, then a bootstrap sample is obtained using `sample` as follows:

```
> Xstar <- sample(X, n, replace=T)
```

Bootstrap MSE estimation. Let $T = g(X_1, \cdots X_n)$ be an estimator for $\theta = \theta(F)$. Let

$$m = \text{MSE}(T) = E\{(T - \theta)^2\} = \text{Var}(T) + (ET - \theta)^2.$$

The bootstrap scheme below provides an estimator m^* for m.

1. Draw $X_1^*, \cdots X_n^*$ independently from $\widehat{F}_n$.
2. Compute $T^* = g(X_1^*, \cdots X_n^*)$.
3. Repeat Steps 1 & 2 B times, to obtain $T_1^*, \cdots, T_B^*$.
4. Compute the sample MSE

$$m^* = \frac{1}{B} \sum_{i=1}^{B} \{T_i^* - \theta(\widehat{F}_n)\}^2,$$

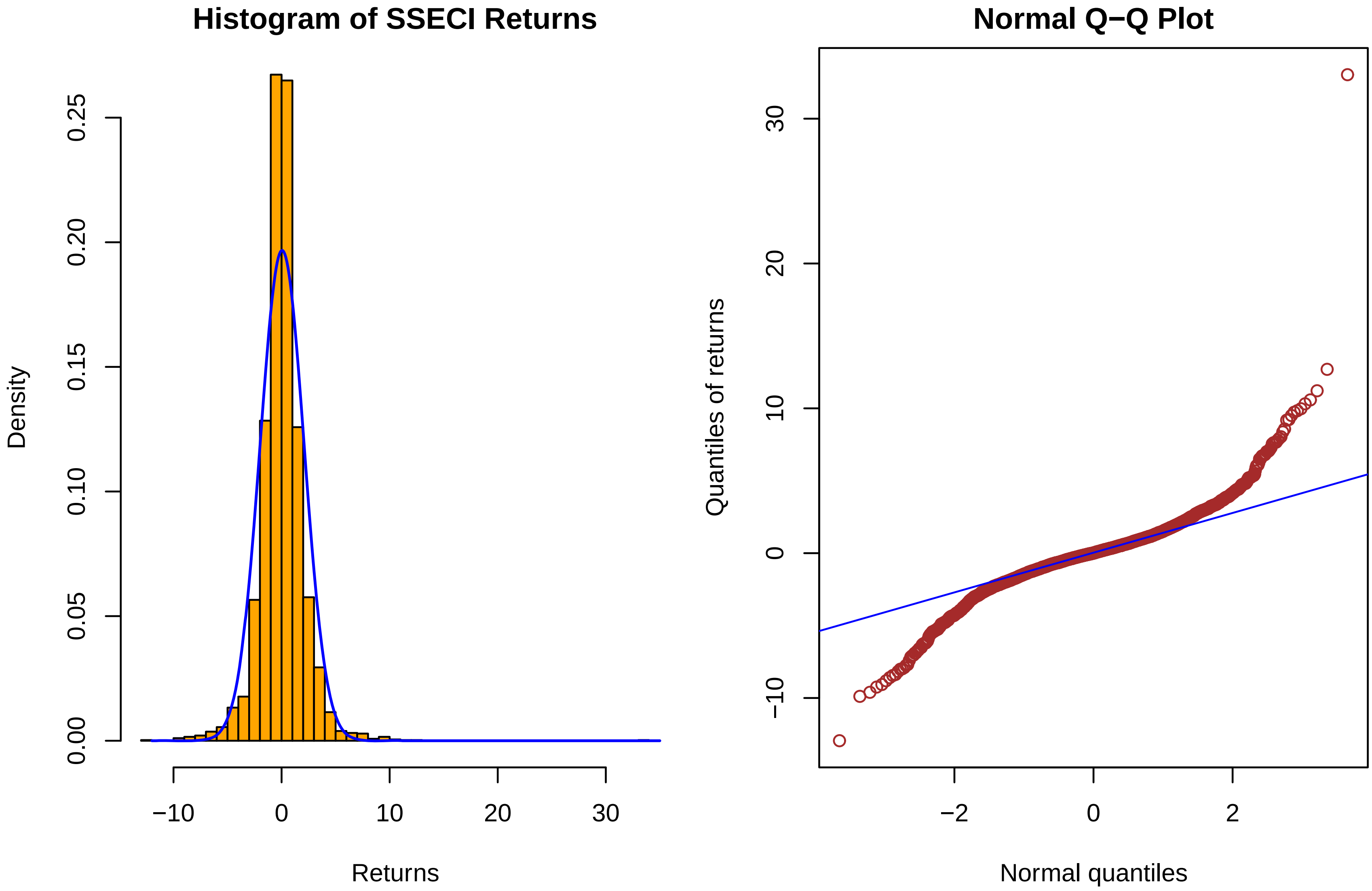
where $\widehat{F}_n(x) = n^{-1} \sum_i I(X_i \le x)$ is the empirical distribution.

Example 1. Consider the daily returns of the Shanghai Stock Exchange Composite Index in December 1994 – September 2010

The data are saved in the file `shanghaiSECI.txt`. The sample size is $n = 3839$.

```
> x <- read.table("shanghaiSECI.txt", skip=3, header=T)
> x[1:4,]   # print out the first 4 rows
  idxcd           idxnmabbr             date            idxdret
1     8           SSE-Composite-Index   1994-12-08     -0.0165
2     8           SSE-Composite-Index   1994-12-09     -0.0014
3     8           SSE-Composite-Index   1994-12-12     -0.0085
4     8           SSE-Composite-Index   1994-12-13      0.0000
> dim(x)
[1] 3839    4
> y <- x[,4]*100   # daily return in percentages
> summary(y)
     Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
```

```
-12.95000  -0.89000   0.01000   0.04994   0.96000  33.04000
> var(y)
[1] 4.111112
> hist(y, nclass=40, prob=T, main='Histogram of SSECI Returns')
> x <- seq(-12, 33, 0.1)
> lines(x, dnorm(x, mean(y), sqrt(var(y))), col='blue', lwd=2)
    # superimpose a normal PDF with the same mean/var onto the histogram
> qqnorm(y, xlab="Normal quantiles", ylab="Quantiles of returns")
                         # quantiles of the empirical distribution
                         # vs quantiles of Normal distribution
> qqline(y, col="blue")  # add a line passing through 1st and 3rd
                         # quartiles
```


Histogram of SSECI Returns
Density
0.00
0.05
0.10
0.15
0.20
0.25
−10
0
10
20
30
Returns
Normal Q−Q Plot
Quantiles of returns
−10
0
10
20
30
−2
0
2
Normal quantiles

The histogram shows that the returns do not follow a normal distribution, as the peak around 0 is much higher.

The Q-Q plot shows that both the tails of the return distribution are much heavier than the tails of normal distributions.

Recall: For any univariate CDF $F(\cdot)$, the quantile of F is defined as $F^{-1}(\alpha) = \inf\{x : F(x) \geq \alpha\}, \ \alpha \in [0, 1]$.

Q-Q plot of two distributions F **and** G: the curves $\{(G^{-1}(\alpha), F^{-1}(\alpha)), \ \alpha \in [0, 1]\}$.

Lemma 1. Let F, G are two univariate CDFs, $b > 0$ and a are two constants. Then $G(x) = F(\frac{x-a}{b})$ for any x iff $G^{-1}(\alpha) = a + bF^{-1}(\alpha)$ for any $\alpha \in [0, 1]$.

Hence, a Q-Q plot is a straight line iff the two distributions are of the same form (i.e. one is a scale-location transformation of the other).

R-functions: `qqnorm, qqline, qqplot`

We introduce two measures related to the 3rd and 4th moments, which are often used as the measures for non-Gaussianality. Let $X \sim F$ and $E(X^4) < \infty$. Write $\mu = EX$ and $\sigma^2 = \mathrm{Var}(X)$.

Skewness of F: $\gamma = E\{(X-\mu)^3\}/\sigma^3$.
Kurtosis of F: $\kappa = E\{(X-\mu)^4\}/\sigma^4$.

Remark. (i) The skewness is a measure for symmetry of distributions. If F is symmetric w.r.t the mean μ (such as $N(\mu, \sigma^2)$), $\gamma = 0$.

(ii) The kurtosis is a measure for tail-heaviness (i.e. fat-tails). For $N(\mu, \sigma^2)$, $\kappa = 3$. When $\kappa > (<)3$, we say that the tails of F are heavier (lighter) than normal distributions.

(iii) Estimators for Skewness and Kurtosis: Let $\bar{X}$ and S^2 be the sample mean and the sample variance. Then

$$\widehat{\gamma} = \frac{1}{nS^3}\sum_{i=1}^{n}(X_i - \bar{X})^3, \qquad \widehat{\kappa} = \frac{1}{nS^4}\sum_{i=1}^{n}(X_i - \bar{X})^4.$$

Example 1 (Continue). We compute the estimates for skewness and kurtosis for the Shanghai SECI returns:

```
> mean((y-mean(y))^3) /var(y)^(1.5)
[1] 1.204415         # estimated skewness
> mean((y-mean(y))^4) /var(y)^2
[1] 25.05686        # estimated kurtosis
```

Since $\widehat{\gamma} = 1.204415 > 0$, the distribution is skewed to the right. The distribution is also heavy-tailed, since $\widehat{\kappa} = 25.05686$.

How accurate are those estimates? — use bootstrap to find the standard errors of the estimators.

```
> skew <- 1:1000
> kurt<- 1:1000
> for(i in 1:1000) {
+ ystar <- sample(y, 3839, replace=T)
+ skew[i] <- mean((ystar-mean(ystar))^3) /var(ystar)^(1.5)
+ kurt[i] <- mean((ystar-mean(ystar))^4) /var(ystar)^2
+ }
> sqrt(var(skew)); sqrt(var(kurt))
[1] 0.9514143  # bootstrap estimate for SE(estimated skewness)
[1] 13.96478   # bootstrap estimate for SE(estimated kurtosis)
```

Hence, the estimated skewness is 1.2044 with the standard error 0.9514, the estimated kurtosis is 25.06 with the standard error 13.97.

In the above we draw $B = 1000$ bootstrap samples. For this example, the results are insensitive for $B \geq 100$.

The analysis indicates that the returns are skewed to its right (unusual!) and heavy-tailed. Certainly their distribution is not normal.

8.2 Bootstrap confidence intervals

8.2.1 Approximate normal intervals

If $(\widehat{\theta}-\theta)/\{\mathrm{Var}(\widehat{\theta})\}^{1/2} \xrightarrow{D} N(0,1)$, an approximate $(1-\alpha)$ confidence interval for θ is

$$\widehat{\theta} \pm Z_{\alpha/2}\{\mathrm{Var}(\widehat{\theta})\}^{1/2},$$

where $Z_{\alpha/2}$ is the top-$\alpha/2$ point of $N(0,1)$.

However $\mathrm{Var}(\widehat{\theta})$ is often unknown. Replacing it by its bootstrap estimate (see section 8.1 above), we obtain a bootstrap interval:

$$\widehat{\theta} \pm Z_{\alpha/2}\{\mathrm{Var}(\theta^*)\}^{1/2}.$$

In practice, we repeat bootstrap sampling B times, obtaining bootstrap estimates $\theta_1^*, \cdots, \theta_B^*$. We take the sample variance of $\{\theta_1^*, \cdots, \theta_B^*\}$ as $\mathrm{Var}(\theta^*)$.

8.2.2 Pivotal intervals

Let $X_1, \cdots, X_n$ be a sample from distribution F. We are interested in estimating a characteristics $\theta = \theta(F)$ (such as mean, skewness etc). Let $\widehat{\theta} = g(X_1, \cdots, X_n) = \theta(\widehat{F_n})$ be the estimator for θ. Let r_α be the α-th percentile of the pivotal $\widehat{\theta} - \theta$, i.e.

$$\alpha = P(\widehat{\theta} - \theta \le r_\alpha).$$

Then

$$P(r_{\alpha/2} < \widehat{\theta} - \theta \le r_{1-\alpha/2}) = 1 - \alpha.$$

This gives a $(1-\alpha)$-th confidence interval of θ:

$$(\widehat{\theta} - r_{1-\alpha/2}, \ \ \widehat{\theta} - r_{\alpha/2}).$$

This is a valid interval estimation if r_α does not depend on θ, i.e. the distribution of the pivotal $\widehat{\theta} - \theta$ does not depend on θ. However this requirement is not necessary if we adopt a bootstrap approach.

Under some standard conditions,

$$P(\widehat{\theta} - \theta < r) \approx P(\theta^* - \widehat{\theta} < r \,\Big|\, X_1, \cdots, X_n)$$

when n is large, where $\theta^* = g(X_1^*, \cdots, X_n^*)$. Thus we may replace $r_{\alpha/2}$ and $r_{1-\alpha/2}$ by their bootstrap counterparts as follows:

> Repeat bootstrap sampling B times to form estimates $\theta_1^*, \cdots, \theta_B^*$. Let θ_α^* be the $[B\alpha]$-th smallest value among $\theta_1^*, \cdots, \theta_B^*$, where $[B\alpha]$ denotes the integer part of $B\alpha$ (i.e. $[a]$ is the largest integer smaller than a). Then
>
> $$r^*_{\alpha/2} = \theta^*_{\alpha/2} - \widehat{\theta}, \qquad r^*_{1-\alpha/2} = \theta^*_{1-\alpha/2} - \widehat{\theta}.$$

The $(1-\alpha)$ bootstrap pivotal interval for θ is:

$$(2\widehat{\theta} - \theta^*_{1-\alpha/2}, \;\; 2\widehat{\theta} - \theta^*_{\alpha/2})$$

8.2.3 Percentile intervals

The $(1-\alpha)$ bootstrap percentile interval for θ is:

$$(\theta^*_{\alpha/2}, \quad \theta^*_{1-\alpha/2})$$

Example 2. We calculate the three bootstrap intervals for the median of the salary for the graduates in a business school based on data in `Jobs.txt` using the following R function:

```
jobsMedianCIs <- function(alpha, B) {
jobs <- read.table("Jobs.txt", header=T, row.names=1)
y <- jobs[,4]  # salary data
cat("Point estimate for median of salaries:", median(y), "\n\n")
my <- 1:B
for(i in 1:B) {
```

```
ystar <- sample(y, 253, replace=T)  # draw bootstrap sample
my[i] <- median(ystar) # bootstrap estimate for median
}
my <- sort(my)   # sort bootstrap estimates in ascending order
i <- as.integer(alpha*B/2)  # i=[B x alpha/2]
cat(1-alpha, "Bootstrap confidence intervals for median of salaries", "\n")
cat("Normal interval:", median(y)-qnorm(1-alpha/2)*sqrt(var(my)),
           median(y)+qnorm(1-alpha/2)*sqrt(var(my)), "\n")
cat("Pivotal interval:", 2*median(y)-my[B-i], 2*median(y)-my[i], "\n")
cat("Percentile interval:", my[i], my[B-i], "\n")
}
```

Calling `jobsMedianCIs(0.05, 5000)`, we obtain the results below. Note that the three intervals for this example are very similar.

```
Point estimate for median of salaries: 47

0.95 Bootstrap confidence intervals for median of salaries
Normal interval: 45.72511 48.27489
Pivotal interval: 46 48
Percentile interval: 46 48
```

Example 1 (Continue). We calculate the three bootstrap intervals for the skewness using the following *R*-function:

```
SSECIbootstrapCIs <- function(B) {
x <- read.table("shanghaiSECI.txt", skip=3, header=T)
y <- x[,4]*100
skew0 <- mean((y-mean(y))^3) /var(y)^(1.5)
cat("Point estimate for skewness:", skew0, "\n\n")
skew <- 1:B
for(i in 1:B) {
ystar <- sample(y, 3839, replace=T)   # draw bootstrap sample
skew[i] <- mean((ystar-mean(ystar))^3) /var(ystar)^(1.5)
}
skew <- sort(skew)   # sort the data in ascending order
i <- as.integer(0.025*B)  # i =[0.025B]
cat("95% Bootstrap confidence intervals for skewness", "\n")
cat("Normal interval:", skew0-2*sqrt(var(skew)),
                  skew0+2*sqrt(var(skew)), "\n")
cat("Pivotal interval:", 2*skew0-skew[B-i], 2*skew0-skew[i], "\n")
cat("Percentile interval:", skew[i], skew[B-i], "\n")
}
```

Call `SSECIbootstrapCIs(1000)`, yielding the following output:

```
Point estimate for skewness: 1.204415

95% Bootstrap confidence intervals for skewness
Normal interval: -0.6486737 3.057503
Pivotal interval: -0.6067615 2.577141
Percentile interval: -0.1683116 3.015591
```

Final Remark. All the bootstrap intervals work well when $\widehat{\theta}$ is asymptotically normal.

Chapter 9. Point Estimation

Let $\{X_1, \cdots, X_n\}$ be a random sample from a population $F(\cdot, \boldsymbol{\theta})$, where the form of F is known but the parameter $\boldsymbol{\theta}$ is unknown, and it has p components. Often we may specify $\boldsymbol{\theta} \in \Theta$, where Θ is called the parameter space.

For $N(\mu, \sigma^2)$, $p = 2$ and $\Theta = (-\infty, \ \infty) \times (0, \ \infty)$. For $Poisson(\lambda)$, $p = 1$ and $\Theta = (0, \ \infty)$.

Goal: to find a (point) estimator for $\boldsymbol{\theta}$.

9.1 Method of Moments Estimation

Let $\mu_k \equiv \mu_k(\boldsymbol{\theta}) = E(X_1^k)$ denote the k-th moment of the population, $k = 1, 2, \cdots$. Then μ_k depends on unknown parameter $\boldsymbol{\theta}$, as everything else on the distribution $F(\cdot, \boldsymbol{\theta})$ are known. Denote the k-th sample moment by

$$M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k = \frac{1}{n}(X_1^k + \cdots + X_n^k).$$

The **MM estimator** $\widehat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ is the solution of the p equations

$$\mu_1(\widehat{\boldsymbol{\theta}}) = M_1,\ \mu_2(\widehat{\boldsymbol{\theta}}) = M_2,\ \cdots,\ \mu_p(\widehat{\boldsymbol{\theta}}) = M_p.$$

Example 1. Let $\{X_1, \cdots, X_n\}$ be a sample from a population with mean μ and variance $\sigma^2 < \infty$. Find the MM estimator for (μ, σ^2).

There are two unknown parameters. Let

$$\mu = \mu_1 = M_1, \qquad \mu_2 = M_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

This gives us $\widehat{\mu} = M_1 = \bar{X}$. Since $\sigma^2 = \mu_2 - \mu_1^2$,

$$\widehat{\sigma}^2 = M_2 - M_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Note. $E(\widehat{\sigma}^2) = E(X_1^2) - E(\bar{X}^2) = \sigma^2 + \mu^2 - (\sigma^2/n + \mu^2) = \frac{n-1}{n}\sigma^2$. We call $E(\widehat{\sigma}^2) - \sigma^2 = -\sigma^2/n$ the estimation bias. The **sample variance**

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

is a more frequently used estimator for σ^2, and it has zero-bias.

Theorem 1. Under some mild regularity conditions, the MME $\widehat{\boldsymbol{\theta}}$ is a consistent estimator in the sense that as $n \to \infty$, $\widehat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}$, i.e.

$$P\{||\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}|| > \epsilon\} \to 0 \qquad \text{for any } \epsilon > 0.$$

Further it is asymptotically normal, i.e. $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges in distribution to a p-dimensional normal distribution.

9.2. Maximum likelihood estimation

9.2.1 Likelihood

Likelihood is one of the most fundamental concepts in all types of statistical inference.

Definition 1 Suppose that $\mathbf{X}$ has density function or probability function $f(\mathbf{x};\boldsymbol{\theta})$. We have observed $\mathbf{X}=\mathbf{x}$. Then the likelihood function with observation $\mathbf{x}$ is defined as

$$L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta};\mathbf{x}) = f(\mathbf{x};\boldsymbol{\theta}).$$

Density/probability function: a function of $\mathbf{x}$, specifying the distribution of random variable $\mathbf{X}$

Likelihood: a function of $\boldsymbol{\theta}$, reflecting information on $\boldsymbol{\theta}$ contained in observation $\mathbf{x}$

Note. A likelihood function represents the uncertainty on a unknown non-random constant $\boldsymbol{\theta}$, and it is **not a density or probability function**! It provides

a rational degree of belief, or
an order of preferences

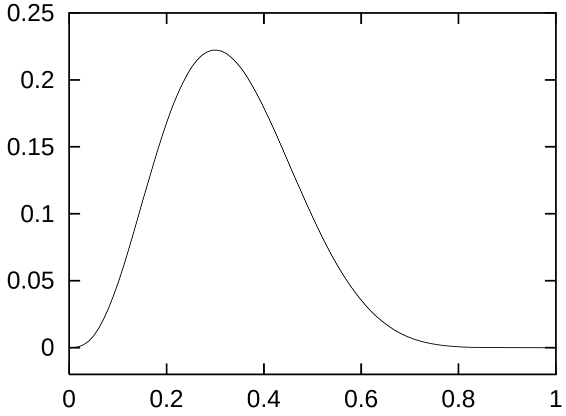
on possible values of the parameter $\boldsymbol{\theta}$. This can be seen more clearly in the simple example on next slide.

In fact, a likelihood function is often defined up to a positive constant — the constant here refers to a quantity independent of $\boldsymbol{\theta}$. But it may depending on $\mathbf{x}$. (Note $\mathbf{x}$ is a given constant.)

Example 2. Suppose that x is the number of successes from a known number n of independent trials with unknown probability of success π. The probability function, and so the likelihood function is

$$L(\pi) = f(x;\pi) = \binom{n}{x}\pi^x(1-\pi)^{n-x}.$$

The likelihood function $L(\pi; x)$ can be graphed as a function of π. It changes shape for different values of x. A likelihood function for a $x = 3$ when $n = 10$ is shown in the Figure below.


0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1

Notice that the likelihood function shown above is *not* a density function. It does not have an area of 1 below it.

We use the likelihood function to compare the plausibility of different possible parameter values. For instance, the likelihood is much larger for $\pi = 0.3$ than for $\pi = 0.8$, that is the data $x = 3$ have a greater probability of being observed if $\pi = 0.3$ than if $\pi = 0.8$. This makes $\pi = 0.3$ much more **likely** as the true value for π than 0.8.

Note. In the above argument, we do not need to calculate exact probabilities under different values of $\boldsymbol{\theta}$. Only the order of those quantities matters!

Let $X_1, \cdots, X_n$ be i.i.d. with PDF $f(\cdot, \boldsymbol{\theta})$. Write $\mathbf{X} = (X_1, \cdots, X_n)'$. Then the likelihood function is

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \mathbf{X}) = \prod_{i=1}^{n} f(X_i, \boldsymbol{\theta}),$$

which is a product of n terms. Then the **log-likelihood function** is

$$l(\boldsymbol{\theta}) = l(\boldsymbol{\theta}; \mathbf{X}) \equiv \log\{L(\boldsymbol{\theta}; \mathbf{X})\} = \sum_{i=1}^{n} \log\{f(X_i, \boldsymbol{\theta})\},$$

which is a sum of n terms.

This explains why log-likelihood functions are often used with independent observations.

9.2.3 Maximum likelihood estimator (MLE)

The MLE is by far the most popular estimator.

Definition 2 — MLE

A *Maximum Likelihood Estimator* (MLE), $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{X}) \in \Theta$, of parameter $\boldsymbol{\theta}$ is an estimator satisfying

> $L(\widehat{\boldsymbol{\theta}}; \mathbf{X}) \geq L(\boldsymbol{\theta}; \mathbf{X})$ for all $\boldsymbol{\theta} \in \Theta$, or equivalently $l(\widehat{\boldsymbol{\theta}}; \mathbf{X}) \geq l(\boldsymbol{\theta}; \mathbf{X})$ for all $\boldsymbol{\theta} \in \Theta$.

Obviously, a maximum likelihood estimator is the most plausible value for $\boldsymbol{\theta}$ as judged by the likelihood function. In many cases where Θ is continuous and the maximum does not occur at a boundary of Θ, **$\widehat{\boldsymbol{\theta}}$ is often the solution** of the equation

$$s(\boldsymbol{\theta}; \mathbf{X}) = \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}; \mathbf{X}) = 0.$$

We call $s(\boldsymbol{\theta}) \equiv s(\boldsymbol{\theta}; \mathbf{X})$ **a score function**.

Example 3. Suppose that $Y_1, Y_2, \ldots, Y_n$ is a random sample from $N(\mu, \sigma^2)$ where neither μ or σ^2 is known. Then we can find the maximum likelihood estimator from the log-likelihood

$$\begin{aligned} l(\mu, \sigma^2) &= -n \log \sqrt{2\pi} - n/2 \log \sigma^2 - \sum_1^n (Y_i - \mu)^2/(2\sigma^2) \\ &= -n \log \sqrt{2\pi} - n/2 \log \sigma^2 - \sum_1^n (Y_i - \bar{Y})^2/(2\sigma^2) - n(\bar{Y} - \mu)^2/(2\sigma^2). \end{aligned}$$

This is maximised by choosing $\mu = \bar{Y}$, so $\widehat{\mu} = \bar{Y}$ is the MLE for μ. It is easy to see

$$E(\widehat{\mu}) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \mu.$$

Such a estimator is called **unbiased**.

The **profile log-likelihood** remaining is

$$l(\widehat{\mu}, \sigma^2) = -n \log \sqrt{2\pi} + (n/2)(\log \sigma^{-2} - \widehat{\sigma}^2 \sigma^{-2}),$$

where $\widehat{\sigma}^2 = \sum_1^n (Y_i - \bar{Y})^2/n$. By the lemma below, the MLE for σ^2 is $\widehat{\sigma}^2$. Note that the MLE of σ^2 is *biased* since

$$E(\widehat{\sigma}^2) = (1 - 1/n)\sigma^2 \neq \sigma^2.$$

Lemma. Define $L(x) = \log(x^{-1}) - b/x$, where $b > 0$ are constants. Then $L(b) \geq L(x)$ for all $x > 0$.

Example 4. Let $X_1, \cdots, X_n$ be i.i.d. Bernoulli(π). Then

$$L(\pi) = \prod_{i=1}^{n} \pi^{X_i}(1 - \pi)^{1-X_i} = \pi^{n\bar{X}}(1 - \pi)^{n(1-\bar{X})}.$$

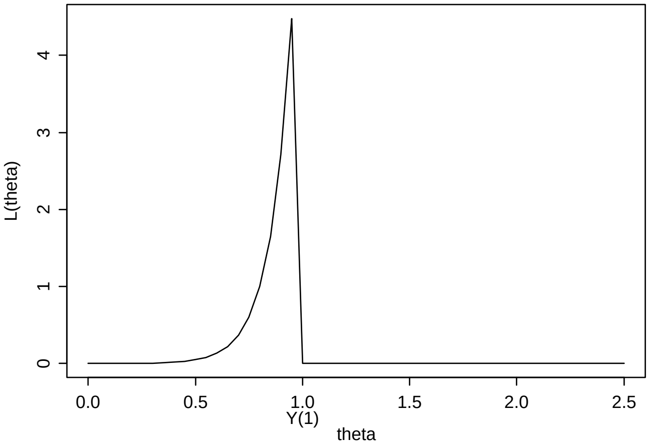
$$l(\pi) = n\bar{X} \log \pi + n(1 - \bar{X}) \log(1 - \pi).$$

Let $s(\pi) = \frac{\partial}{\partial \pi} l(\pi) = 0$, leading to $\widehat{\pi} = \bar{X}$.

Example 5. Suppose that $Y_1, Y_2, \ldots, Y_n$ is a random sample from an exponential distribution with density function $e^{-(y-\theta)}$ for $y \geq \theta$. This is the usual exponential distribution shifted to start at θ. The Likelihood is

$$L(\theta; \mathbf{Y}) = e^{-n(\bar{Y}-\theta)} I_{\{(\theta,\infty)\}}(Y_{(1)}),$$

where $Y_{(1)}$ is the smallest observation. This likelihood is zero for $\theta > Y_{(1)}$ and increases in θ for $\theta \leq Y_{(1)}$. So the MLE $\widehat{\theta} = Y_{(1)}$, which is a boundary maximum.



Invariance property of MLEs

Suppose $\mathbf{X} \sim f(\mathbf{x}, \boldsymbol{\theta})$, and $\boldsymbol{\psi} = g(\boldsymbol{\theta})$. Let $\widehat{\boldsymbol{\theta}}$ be the MLE for $\boldsymbol{\theta}$, i.e.

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} f(\mathbf{X}, \boldsymbol{\theta}).$$

It is obvious to see that the MLE for $\boldsymbol{\psi}$ is $\widehat{\boldsymbol{\psi}} = g(\widehat{\boldsymbol{\theta}})$.

If $\boldsymbol{\psi} = g(\boldsymbol{\theta})$ is a 1-1 transform and $\widehat{\boldsymbol{\psi}}$ is the MLE for $\boldsymbol{\psi}$, $\widehat{\boldsymbol{\theta}} \equiv g^{-1}(\widehat{\boldsymbol{\psi}})$ is the MLE for $\boldsymbol{\theta}$.

9.2.4 Numerical computation of MLEs

In modern statistical applications, it is typically difficult to find explicit analytic forms for the maximum likelihood estimators. These estimators are found more often by iterative procedures built into computer software. An iterative scheme starts with some guess at the MLE and then steadily improves it with each iteration. The estimator is considered found when it has become numerically stable. Sometimes the iterative procedures become trapped at a local maximum which is not a global maximum. There may be a very large number of parameters in a model, which makes such local entrapment much more common.

Newton-Raphson Scheme

Suppose that the log-likelihood function function $l(\boldsymbol{\theta})$ is sufficiently smooth. Then

$$s(\widehat{\boldsymbol{\theta}}) = 0,$$

where $\widehat{\boldsymbol{\theta}}$ is the MLE and $s(\boldsymbol{\theta}) = \frac{\partial}{\partial\theta} l(\boldsymbol{\theta})$ is the score function. Let

$$\dot{s}(\boldsymbol{\theta}) = \ddot{l}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial\theta\partial\theta'} l(\boldsymbol{\theta}) = \left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right).$$

Suppose $\widehat{\boldsymbol{\theta}}$ is close to the true value $\boldsymbol{\theta}^0$. By a simple Taylor expansion,

$$s(\boldsymbol{\theta}^0) = \dot{s}(\boldsymbol{\theta}^0)(\boldsymbol{\theta}^0 - \widehat{\boldsymbol{\theta}}) + o_p(||\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0||).$$

This leads to the approximation

$$\widehat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}^0 - \{\dot{s}(\boldsymbol{\theta}^0)\}^{-1} s(\boldsymbol{\theta}^0).$$

Since θ^0 is unknown, we use iterative estimators

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \{\dot{s}(\boldsymbol{\theta}_k)\}^{-1} s(\boldsymbol{\theta}_k) \tag{6}$$

for $k = 1, 2, \cdots$, where $\boldsymbol{\theta}_0$ is a prescribed initial value. We define $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_j$ if $\boldsymbol{\theta}_j$ and $\boldsymbol{\theta}_{j-1}$ differ by a small amount.

Example 6. Let $X_1, \cdots, X_n$ be a sample from Cauchy distribution with PDF

$$f(x, \theta) = \frac{1}{\pi\{1 + (x - \theta)^2\}},$$

where θ is the location parameter. The log-likelihood is

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

The MLE is the solution of $s(\widehat{\theta}) = 0$, where

$$s(\theta) = 2\sum_{i=1}^{n} \frac{X_i - \theta}{1 + (X_i - \theta)^2}.$$

Since $s(\theta) = 0$ does not admit an explicit solution, we adopt a Newton-Raphson scheme: $\theta_{k+1} = \theta_k - s(\theta_k)/\dot{s}(\theta_k)$, where

$$\dot{s}(\theta) = 2\sum_{i=1}^{n} \frac{(X_i - \theta)^2 - 1}{\{1 + (X_i - \theta)^2\}^2}.$$

The *R*-function below implements the above scheme.

```
cauchyMLE <- function(n, theta, init, Tiny) {
x <- rcauchy(n, theta) # x is a sample
i <- 0   # No. of iterations
theta0 <- init +10*Tiny
theta1 <- init
while(abs(theta1-theta0)>Tiny) {
    theta0 <- theta1
    x2 <- x-theta0
    x22 <- x2*x2
    t1 <- mean(x2/(x22+1))         #  s(theta0)/(2n)
    t2 <- mean((x22-1)/(x22+1)^2)  #  derivation of s(theta0)/(2n)
    theta1 <- theta0 - t1/t2
    i <- i+1
    cat(i, "iteration:", theta1, "\n")  # print out iteration values
}
cat("MLE:", theta1, "No. of iterations:", i, "\n")
}
```

By calling `cauchyMLE(100,10,11.12,0.01)`, we excute the above iterative algorithm as follows:

```
> source("cauchyMLE.r")
> cauchyMLE(100, 10, 11.12, 0.01)
1 iteration: 9.594835
2 iteration: 10.08787
3 iteration: 10.0752
4 iteration: 10.07521
```

```
MLE: 10.07521 No. of iterations: 4
```

Note the initial value is important: the iterations will not converge if $\theta_0 < 8.75$ or $\theta_0 > 11.2$ on my PC.

Choosing a good initial value is always important. For this example, the PDF is symmetric around θ, it makes sense to consider either the sample mean or sample median as an initial estimate. However $E(X_1)$ is not well-defined, so the sample mean may not be a good estimator θ. Thus we may use the sample median as the initial value for our algorithm.

The Fisher Scoring method: replace $\dot{s}(\widehat{\boldsymbol{\theta}}_k)$ in (6) by $E_{\boldsymbol{\theta}}\{\dot{s}(\boldsymbol{\theta})\}$ under $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_k$. So the algorithm is now

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - [E\{\dot{s}(\boldsymbol{\theta}_k)\}]^{-1} s(\boldsymbol{\theta}_k) = \boldsymbol{\theta}_k + \{\mathcal{I}(\boldsymbol{\theta}_k)\}^{-1} s(\boldsymbol{\theta}_k),$$

where

$$\mathcal{I}(\boldsymbol{\theta}) = \mathrm{Var}\{s(\boldsymbol{\theta})\} = -E\{\dot{s}(\boldsymbol{\theta})\}$$

is the Fisher information.

Example 6 (continue). It can be shown that

$$E\{\dot{s}(\theta)\} = \frac{2n}{\pi} \int_{-\infty}^{\infty} \frac{(x-\theta)^2 - 1}{\{1 + (x-\theta)^2\}^3} dx = -n/2.$$

Hence the Fisher scoring method is

$$\theta_{k+1} = \theta_k + \frac{4}{n} \sum_{i=1}^{n} \frac{X_i - \theta_k}{1 + (X_i - \theta_k)^2}.$$

```
cauchyMLEscoring <- function(n, theta, init, Tiny) {
   # Fisher scoring MLE for Cauchy(theta)
   # n is sample size, Tiny is the tolerance limit controls
   # iteration estimates, init is the initial value used iteration
x <- rcauchy(n, theta)
i <- 0   # No. of iterations
theta0 <- init +10*Tiny
theta1 <- init
while(abs(theta1-theta0)>Tiny) {
    theta0 <- theta1
    x2 <- x-theta0
    t1 <- mean(x2/(x2*x2+1))         # s(theta0)/(2n)
    theta1 <- theta0 + 4*t1
    i <- i+1
    cat(i, "iteration:", theta1, "\n")  # print out iteration values
}
cat("\n", "MLE:", theta1, "No. of iterations:", i, "\n")
}
```

Calling `cauchyMLEscoring(100, 10, 15, 0.1)` yields

```
MLE: 10.14528 No.  of iterations:  7
```

Note now that the range of valid initial values is much bigger.

Like most iterative algorithms, the choice of appropriate **initial values is important** to ensure the convergence to right limits. In practice multiple initial values are often used.

The differences between the Newton-Raphson and Fisher scoring methods are subtle. We make observations below

- The convergence of the Newton-Raphson algorithm is often faster when both algorithms converge

- The radius of convergence for the Fisher scoring method is often larger, making the choice of initial values less important for the scoring method.

9.2.5 EM algorithms

Goal: to find the MLE $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{Y})$ for $\boldsymbol{\theta}$ from the likelihood based on data $\mathbf{Y}$:

$$L(\boldsymbol{\theta}; \mathbf{Y}) = f_{\mathbf{Y}}(\mathbf{Y}, \boldsymbol{\theta}),$$

while the '*complete*' data $\mathbf{X}' = (\mathbf{Y}', \mathbf{Z}')$ contain a '*missing*' component $\mathbf{Z}$. The likelihood based on the complete data is

$$L(\boldsymbol{\theta}; \mathbf{X}) = f_{\mathbf{X}}(\mathbf{X}, \boldsymbol{\theta}).$$

EM (*E*xpectation and *M*aximisation) **algorithms**

- *E-step*: compute the conditional expectation

$$Q(\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\theta}_0) \equiv E\{\log L(\boldsymbol{\theta}; \mathbf{X})|\mathbf{Y}, \boldsymbol{\theta}_0\}$$

- *M-step*: maximise $Q(\boldsymbol{\theta})$ to give an updated value $\boldsymbol{\theta}_1$

then go to the E-step using $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_1$, and keep iterating until convergence. The limit of $\boldsymbol{\theta}_0$ is taken as $\widehat{\boldsymbol{\theta}}(\mathbf{Y})$.

Example 7. The genetic example from (Rao 1973, p.396) assumes that the phenotype data

$$\begin{aligned}\mathbf{Y} &= (Y_1, Y_2, Y_3, Y_4)' \sim \\ &\text{Multinomial}(4; \frac{1}{2}+\frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}),\end{aligned}$$

where $\theta \in (0,1)$. Then log-likelihood is

$$l(\theta, \mathbf{Y}) = Y_1 \log(2+\theta) + (Y_2+Y_3)\log(1-\theta) + Y_4 \log\theta + C,$$

which **does not yield a closed form $\widehat{\theta}$**.

Now we treat **Y** as incomplete data from $\mathbf{X} = (X_1, \ldots, X_5)'$ with multinomial probabilities

$$(\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}).$$

Then

$$Y_1 = X_1 + X_2, \quad Y_i = X_{i+1} \quad \text{for } i = 2, 3, 4.$$

The log-likelihood of based on **X** is

$$l(\theta, \mathbf{X}) = (X_2 + X_5) \log \theta + (X_3 + X_4) \log(1 - \theta) + C,$$

which **readily yields**

$$\widehat{\theta}(\mathbf{X}) = \frac{X_2 + X_5}{X_2 + X_3 + X_4 + X_5}.$$

Now the E-step is to find

$$\begin{aligned} Q(\theta) &= E\{l(\theta, \mathbf{X})|\mathbf{Y}, \theta_0\} \\ &= \log \theta E(X_2 + X_5|\mathbf{Y}, \theta_0) + \log(1 - \theta) E(X_3 + X_4|\mathbf{Y}, \theta_0) \\ &= (\widehat{X}_2 + Y_4) \log \theta + (Y_2 + Y_3) \log(1 - \theta), \end{aligned}$$

where $\widehat{X}_2 = E(X_2|\mathbf{Y}, \theta_0)$. Since the conditional distribution of X_2 given $Y_1 (= X_1 + X_2)$ is a binomial distribution with $n = Y_1$ and

$$p = \frac{\theta_0/4}{1/2 + \theta_0/4} = \frac{\theta_0}{2 + \theta_0}.$$

Hence

$$\widehat{X}_2 = np = \frac{Y_1\theta_0}{2+\theta_0}. \tag{7}$$

The M-step leads to

$$\theta_1 = \frac{\widehat{X}_2 + Y_4}{\widehat{X}_2 + Y_4 + Y_2 + Y_3}. \tag{8}$$

For $Y = (125, 18, 20, 34)$ and the initial value

$$\theta_0 = 4 \times 34/(125 + 18 + 20 + 34)$$

which is a relative frequency estimate, the first 5 iterations between (7) and (8) are 0.690, 0.635, 0.628, 0.627 and 0.627, giving the MLE $\widehat{\theta} = 0.627$.

What can be said about *convergence properties* of the EM algorithm?

Let $\boldsymbol{\theta}_0$ be an arbitrary initial value and $\boldsymbol{\theta}_1$ be the updated value obtained from applying the iteration once. Then it can be shown that

$$L(\boldsymbol{\theta}_1; \mathbf{Y}) \geq L(\boldsymbol{\theta}_0; \mathbf{Y}).$$

Unfortunately, it does not imply that the iterations will always lead to the MLE eventually.

It is important to choose appropriate initial values to ensure the algorithm converges to the MLE. (This is also true for both Gaussian-Raphson and score methods!) In practice, it is a good idea to use **a variety of initial values**.

Further discussion on the convergence of EM algorithms, see Wu (1983) *Annals of Statistics, Vol.11, pp.95-103* and §12.4 of Pawitan (2001).

General comments on EM algorithms:

- The EM algorithm is a general procedure for computing MLEs. It is not really a numerical algorithm. The calculation of M-Step typically involves other numerical algorithms such as Newton-Raphson and the score methods.

- It can be applied when some data are *genuinely* missing. It can also be applied when the missing information is merely a concept based on which we transform a difficult optimisation problem into a sequence of easier problems; see Example 7 above. This is particularly relevant when $\widehat{\boldsymbol{\theta}}(\mathbf{Y})$ is difficult to calculate while $\widehat{\boldsymbol{\theta}}(\mathbf{X})$ is easier to obtain.

- The convergence of EM algorithm may be very slow, depending on the amount of missing information.

Example 8. Mixture distributions

Let $Y_1, \cdots, Y_n$ be i.i.d. with a mixture PDF

$$f(y) = \sum_{j=1}^{k} \alpha_j f_j(y, \boldsymbol{\lambda}_j),$$

where f_j are PDFs, $\alpha_j \geq 0$ and $\sum_j \alpha_j = 1$. The parameter $\boldsymbol{\theta}$ contains all α_j and $\boldsymbol{\lambda}_j$. This model becomes important because

1. it represents heterogeneous data well since each f_j represents one heterogeneous component, and
2. it provides very good approximations to a large class of distributions.

The likelihood based on $\mathbf{Y} = (Y_1, \cdots, Y_n)'$ is

$$L(\boldsymbol{\theta}; \mathbf{Y}) = \prod_{i=1}^{n} \Big\{ \sum_{j=1}^{k} \alpha_j f_j(Y_i, \boldsymbol{\lambda}_j) \Big\}.$$

Maximising this likelihood is difficult, due to the presence of the summations, which reflect the fact that we are typically lacking in knowledge of which component any particular sample value comes from. This is the missing information!

Let $\mathbf{X}_i' = (Y_i, \mathbf{Z}_i')$, where $\mathbf{Z}_i = (Z_{i1}, \cdots, Z_{ik})'$ is a $k \times 1$ vector with a 1 in the position corresponding to the component of the mixture that Y_i comes from, and 0 elsewhere.

Let $\mathbf{e}_j$ be the $k \times 1$ vector with the j-th component 1 and all the other components 0. The joint probability-density function for $(\mathbf{Z}_1, Y_1)$ is

$$\begin{aligned} & P\{Z_1 = \mathbf{e}_\ell,\ Y_1 \in [y_1, y_1 + dy_1)\}/dy_1 \\ = \;& P(Z_1 = \mathbf{e}_\ell)\, P\{Y_1 \in [y_1, y_1 + dy_1)|Z_1 = \mathbf{e}_\ell\}/dy \\ = \;& \alpha_\ell f_\ell(y_1, \boldsymbol{\lambda}_\ell) \;=\; \prod_{j=1}^{k} \alpha_j^{e_{\ell j}} f_j(y_1, \boldsymbol{\lambda}_j)^{e_{\ell j}}, \end{aligned}$$

where $\mathbf{e}_\ell = (e_{\ell 1}, \cdots, e_{\ell k})'$. Let $\mathbf{X}' = (\mathbf{X}_1', \cdots, \mathbf{X}_n')$.

$$L(\boldsymbol{\theta}; \mathbf{X}) = \prod_{i=1}^{n} \prod_{j=1}^{k} \alpha_j^{Z_{ij}} f_j(Y_i, \boldsymbol{\lambda}_j)^{Z_{ij}},$$

$$l(\boldsymbol{\theta}; \mathbf{X}) = \sum_{i=1}^{n} \mathbf{Z}_i' \left\{ \begin{pmatrix} \log \alpha_1 \\ \vdots \\ \log \alpha_k \end{pmatrix} + \begin{pmatrix} \log f_1(Y_i, \boldsymbol{\lambda}_1) \\ \vdots \\ \log f_k(Y_i, \boldsymbol{\lambda}_k) \end{pmatrix} \right\}.$$

Note that

$$\begin{aligned} E(\mathbf{Z}_i | \mathbf{Y}, \boldsymbol{\theta}_0) &= \left(\frac{\alpha_1^0 f_1(Y_i, \boldsymbol{\lambda}_1^0)}{\sum_{\ell=1}^{k} \alpha_\ell^0 f_\ell(Y_i, \boldsymbol{\lambda}_\ell^0)}, \cdots, \frac{\alpha_k^0 f_k(Y_i, \boldsymbol{\lambda}_k^0)}{\sum_{\ell=1}^{k} \alpha_\ell^0 f_\ell(Y_i, \boldsymbol{\lambda}_\ell^0)} \right)' \\ &\equiv \left(b_1(Y_i, \boldsymbol{\theta}_0), \cdots, b_k(Y_i, \boldsymbol{\theta}_0) \right)', \end{aligned}$$

which are constants as far as the M-step is concerned.

Now the E-step implies that

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{n} E(\mathbf{Z}_i'|\mathbf{Y}, \boldsymbol{\theta}_0) \left\{ \begin{pmatrix} \log \alpha_1 \\ \vdots \\ \log \alpha_k \end{pmatrix} + \begin{pmatrix} \log f_1(Y_i, \boldsymbol{\lambda}_1) \\ \vdots \\ \log f_k(Y_i, \boldsymbol{\lambda}_k) \end{pmatrix} \right\},$$

The M-step requires to maximise $Q(\boldsymbol{\theta})$, i.e.

$$\boldsymbol{\theta}_1 = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}).$$

Note that $\{\boldsymbol{\lambda}_j^1\}$ and $\{\alpha_j^1\}$ can be evaluated separately, which is much easier than minimizing $l(\boldsymbol{\theta}, \mathbf{Y})$ directly. For example,

$$\alpha_j^1 = \frac{\sum_{i=1}^k b_j(Y_i, \boldsymbol{\theta}_0)}{\sum_{\ell=1}^k \sum_{i=1}^k b_\ell(Y_i, \boldsymbol{\theta}_0)}, \qquad j = 1, \cdots, k.$$

When f_j is a normal PDF, $\boldsymbol{\lambda}_j^1$ admits an explicit formula.

Let $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_1$, keep iterating between E-step and M-step until two successive values of $\boldsymbol{\theta}_1$ differ by a small amount.

9.3 Evaluating estimation

To measure the accuracy of an MLE or, more general, any estimation procedure, we need to define some measures for the goodness (or badness) of an estimator.

Let $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{X})$ be an estimator of $\boldsymbol{\theta}$, and $\boldsymbol{\theta}_o$ be the (unknown) *true value* of $\boldsymbol{\theta}$. Note that

(i) exact estimation error $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o$ is unknown, and
(ii) $\widehat{\boldsymbol{\theta}}$ is a random variable

we have to gauge the error

(i) in terms of a probability average, and
(ii) for all possible values of $\boldsymbol{\theta}_o \in \Theta$.

Let $P_{\boldsymbol{\theta}}$, $E_{\boldsymbol{\theta}}$ and $\text{Var}_{\boldsymbol{\theta}}$ denote the probability distribution, expectation and variance under $\boldsymbol{\theta}_o = \boldsymbol{\theta}$.

Bias: $\text{Bias}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) = E_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\theta}$

Variance: $\text{Var}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})$

Standard deviation: $\{\text{Var}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})\}^{1/2}$

Standard error: $\{\text{Var}_{\widehat{\boldsymbol{\theta}}}(\widehat{\boldsymbol{\theta}})\}^{1/2}$

Mean square error (MSE): $E_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2$

Mean absolute error (MAE): $E_{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}|$

Note that

- standard error is a meaningful measure of accuracy for (approximately) unbiased estimators only, and

- MSE (or its squared-root) should be used in general as

$$\mathrm{MSE}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) = \{\mathrm{Bias}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})\}^2 + \mathrm{Var}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}).$$

Ideally we would seek for the estimator which minimises MSE or MAE **for all $\boldsymbol{\theta} \in \Theta$** over all possible candidate estimators. Unfortunately such a global optimum rarely exists. However if we confine to some subclass of estimators, the MLE is often optimal or asymptotically optimal.

The MSE is most frequently used largely due to is technical tractability while the MAE leads to estimators which is more robust against outliers in observations.

Fisher Information

Suppose $\mathbf{X} \sim f(\mathbf{x}, \boldsymbol{\theta})$. The score function is

$$s(\boldsymbol{\theta}) = \dot{l}(\boldsymbol{\theta}; \mathbf{X}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log\{f(\mathbf{X}, \boldsymbol{\theta})\}.$$

We assume certain regularity conditions so that we can take derivatives under the integral sign.

Mean of $s(\boldsymbol{\theta})$:

$$\begin{aligned} E_{\boldsymbol{\theta}}\{s(\boldsymbol{\theta})\} &= \int s(\boldsymbol{\theta}) f(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} = \int \frac{\partial}{\partial \boldsymbol{\theta}} \log\{f(\mathbf{x}, \boldsymbol{\theta})\} f(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} \\ &= \int \frac{\partial}{\partial \boldsymbol{\theta}} f(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} = \frac{\partial}{\partial \boldsymbol{\theta}} \int f(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} = 0. \end{aligned}$$

Variance of $s(\boldsymbol{\theta})$ — **Fisher information** matrix:

$$\mathcal{I}(\boldsymbol{\theta}) = \mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{\theta}}\{s(\boldsymbol{\theta})\} = E_{\boldsymbol{\theta}}\{s(\boldsymbol{\theta})s(\boldsymbol{\theta})'\} = -E_{\boldsymbol{\theta}}\Big[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log\{f(\mathbf{X},\boldsymbol{\theta})\}\Big],$$

because

$$E_{\theta}\{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}l(\boldsymbol{\theta})\} = \int \frac{\ddot{L}(\boldsymbol{\theta})L(\boldsymbol{\theta}) - \dot{L}(\boldsymbol{\theta})\dot{L}(\boldsymbol{\theta})'}{L(\theta)}d\mathbf{x} = \int \ddot{L}(\boldsymbol{\theta})d\mathbf{x} - \int \frac{\dot{L}(\boldsymbol{\theta})\dot{L}(\boldsymbol{\theta})'}{L(\theta)}d\mathbf{x}$$
$$= -\int \frac{\dot{L}(\boldsymbol{\theta})\dot{L}(\boldsymbol{\theta})'}{L(\theta)}d\mathbf{x} = -\int s(\boldsymbol{\theta})s(\boldsymbol{\theta})'f(\mathbf{x},\boldsymbol{\theta})d\mathbf{x} = -E_{\boldsymbol{\theta}}\{s(\boldsymbol{\theta})s(\boldsymbol{\theta})'\}.$$

Fisher information $\mathcal{I}(\boldsymbol{\theta})$ measures **the information on θ contained in data $\mathbf{X}$**. Further if $\mathbf{X} = (X_1, \cdots, X_n)'$, and $X_1, \cdots, X_n$ are IID,

$$\mathcal{I}(\boldsymbol{\theta}) = \mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}) = \sum_{j=1}^{n} \mathcal{I}_{X_j}(\boldsymbol{\theta}) = n\mathcal{I}_{X_1},$$

i.e. the information is additive.

For $\boldsymbol{\theta} = \theta$ is a scalar, the Fisher information is

$$\mathcal{I}(\theta) = E_\theta\{s(\theta)^2\} = -E_\theta\{\ddot{l}(\theta)\}.$$

Theorem 2. *(Cramér-Rao inequality)*
Let $\mathbf{X} \sim f(\cdot, \theta)$ which satisfying some regularity conditions. Let $T = T(\mathbf{X})$ be a statistic with $g(\theta) = E_\theta(T)$. Then for any $\theta \in \Theta$,

$$\mathrm{Var}_\theta(T) \geq \{\dot{g}(\theta)\}^2/\mathcal{I}(\theta).$$

The Cramér-Rao inequality specifies **a lower bound** for any *unbiased estimator* for the parameter $g(\theta)$. When the equality holds, T is the **minimum variance unbiased estimator (MVUE)** of $g(\theta)$.

Important case: For any unbiased estimator $\widehat{\theta} = \widehat{\theta}(\mathbf{X})$,

$$\mathrm{Var}(\widehat{\theta}) \geq 1/\mathcal{I}(\theta).$$

Multivariate case: For any unbiased estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{X})$, $\mathrm{Var}(\widehat{\boldsymbol{\theta}}) - \{\mathcal{I}(\boldsymbol{\theta})\}^{-1}$ is a non-negative definite matrix. Hence $\mathrm{Var}(\widehat{\theta}_j) \geq I^{jj}(\boldsymbol{\theta})$, where $\widehat{\theta}_j$ is the j-th component of $\widehat{\boldsymbol{\theta}}$, and $I^{jj}(\boldsymbol{\theta})$ is the (j, j)-th element of $\{\mathcal{I}(\boldsymbol{\theta})\}^{-1}$.

Example 9. Let $X_1, \cdots, X_n$ be a sample from $N(\mu, \sigma^2)$. We consider estimators for μ, treating σ^2 as known. The score function (for one observation) is

$$\begin{aligned} s(\mu; X_1) &= \frac{\partial}{\partial \mu} \log[e^{-\frac{1}{2\sigma^2}(X_1-\mu)^2}/\sqrt{2\pi\sigma^2}] \\ &= \frac{\partial}{\partial \mu}[-\frac{1}{2\sigma^2}(X_1-\mu)^2] = (X_1-\mu)/\sigma^2. \end{aligned}$$

Note $\ddot{l}(\mu) = \dot{s}(\mu) = -\sigma^{-2}$. Hence the Fisher information based on a single observation is $\mathcal{I}_{X_1}(\mu) = \sigma^{-2}$. Therefore

$$\mathcal{I}(\mu) = \mathcal{I}_{X_1,\cdots,X_n}(\mu) = n/\sigma^2.$$

For any unbiased estimator $\widehat{\mu}$ for μ, it holds that

$$\mathrm{Var}_\mu(\widehat{\mu}) \geq \sigma^2/n,$$

which is the variance of $\bar{X}$. Hence $\bar{X}$ is the MVUE for μ.

Asymptotic properties of MLEs

Let $X_1, \cdots, X_n$ be i.i.d. with PDF $f(\cdot, \boldsymbol{\theta})$. Write

$$l(\boldsymbol{\theta}) = l(\boldsymbol{\theta}; \mathbf{X}) = \sum_{j=1}^{n} \log f(X_j, \boldsymbol{\theta}).$$

Let $\widehat{\boldsymbol{\theta}}$ be the MLE which maximises $l(\boldsymbol{\theta})$. Suppose f fulfils certain regularity conditions.

(a) Consistency.
The MLE is consistent in the sense that as $n \to \infty$,

$$P_{\boldsymbol{\theta}}\{||\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}|| > \varepsilon\} \longrightarrow 0$$

for any $\varepsilon > 0$.

Consistency requires that an estimator converges to the parameter to be estimated. It is a very mild and modest condition that any reasonable estimator should fulfil. The consistency condition is often used to *rule out bad estimators.*

(b) Asymptotic normality

As $n \longrightarrow \infty$,

$$n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N\left(0,\ \{\mathcal{I}_{X_1}(\boldsymbol{\theta})\}^{-1}\right).$$

For large n, it holds **approximately** that

$$\widehat{\boldsymbol{\theta}} \sim N\left(\boldsymbol{\theta},\ \{\mathcal{I}_{X_1}(\boldsymbol{\theta})\}^{-1}/n\right).$$

Therefore *asymptotically* the MLE is unbiased and attains the Cramér-Rao lower bound. Any estimator fulfilling this condition is called **efficient**.

An approximate standard error of the j-th component of $\widehat{\boldsymbol{\theta}}$ is the square-root of the (j,j)-th element of $\{\mathcal{I}_{X_1}(\widehat{\boldsymbol{\theta}})\}^{-1}$ divided by $n^{1/2}$.

Bootstrapping MSEs — Parametric bootstrap

An MSE provides a measure for the accuracy of the estimator. But it is not always feasible to derive an explicit expression for MSE. Furthermore it depends on the unknown parameter. Alternatively we may estimate the MSE by Bootstrapping.

Let $X_1, \cdots, X_n$ be i.i.d. with PDF $f(\cdot, \theta)$, where θ is a scalar. Let

$$\widehat{\theta} = T(X_1, \cdots, X_n)$$

be an estimator. The *goal* here is to estimate

$$\nu \equiv \{\text{MSE}_{\theta_o}(\widehat{\theta})\}^{1/2},$$

where θ_o is the true value.

If we *knew* $f(\cdot, \theta_o)$ completely, ν is known in principle, and may be estimated easily via a repeated sampling as follows. We draw B independent samples of size n from $f(\cdot, \theta_o)$. For each sample, we calculate $\widehat{\theta}$, obtaining $\widehat{\theta}_1, \cdots, \widehat{\theta}_B$. Then the sample root-MSE

$$\{\frac{1}{B}\sum_{b=1}^{B}(\widehat{\theta}_b - \theta_o)^2\}^{1/2}$$

is a reasonable estimator for ν. By the LLN, this estimator converges to ν as $B \longrightarrow \infty$.

The **basic idea of parametric bootstrap** is to adopt the above sampling procedure in the so-called *bootstrap world*: now the population is $f(\cdot, \widehat{\theta})$ which is known. We draw a sample denoted as $(X_1^*, \cdots, X_n^*)$ from this distribution. Define the *bootstrap version* of the estimator

$$\widehat{\theta}^* = T(X_1^*, \cdots, X_n^*).$$

Then the quantity

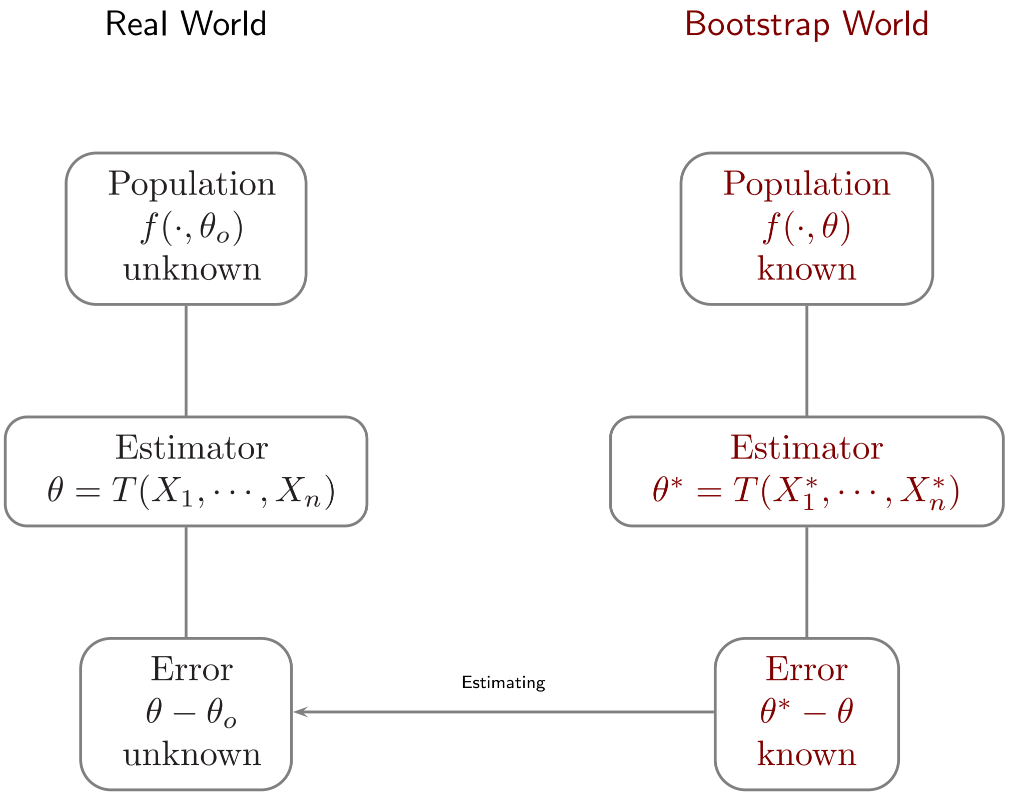
$$\nu^* = \{\mathrm{MSE}_{\widehat{\theta}}(\widehat{\theta}^*)\}^{1/2}$$

is known in principle since the distribution $f(\cdot, \widehat{\theta})$ is completely known. Define ν^* as a **bootstrap estimator** for ν.

In practice, we draw B sets samples from $f(\cdot, \widehat{\theta})$, forming B bootstrap versions of estimator $\widehat{\theta}_1^*, \cdots, \widehat{\theta}_B^*$. The ν^* is calculated as

$$\nu^* = \{\frac{1}{B}\sum_{j=1}^{B}(\widehat{\theta}_j^* - \widehat{\theta})^2\}^{1/2}.$$

Remark. The bootstrap methods introduced in Chapter 8 are in the category of *Nonparametric Bootstrap*. If we know the form of the underlying distribution, parametric bootstrap methods are typically more efficient.



Real World

Bootstrap World

Population
$f(\cdot, \theta_o)$
unknown

Population
$f(\cdot, \theta)$
known

Estimator
$\theta = T(X_1, \cdots, X_n)$

Estimator
$\theta^* = T(X_1^*, \cdots, X_n^*)$

Error
$\theta - \theta_o$
unknown

Estimating

Error
$\theta^* - \theta$
known

Bootstrap is a powerful tool for statistical inference. It has different forms for different applications. The diagram above indicates that the basic idea of a *parametric bootstrap* method.

Chapter 10. Hypothesis Testing (I)

Hypothesis Testing, together with statistical estimation, are the two most frequently used statistical inference methods. It addresses a different type of practical problems from statistical estimation.

10.1 Basic idea, *p*-values

Based on the data, a (statistical) test is to make a binary decision on a well-defined hypothesis, denoted as H_0:

$$\text{Reject } H_0 \quad \text{or} \quad \text{Not reject } H_0$$

Consider a simple experiment: toss a coin n times.

Let $X_1, \cdots, X_n$ be the outcomes: Head – $X_i = 1$, Tail – $X_i = 0$

Probability distribution: $P(X_i = 1) = \pi = 1 - P(X_i = 0)$, $\pi \in (0,\ 1)$

Estimation: $\widehat{\pi} = \bar{X} = (X_1 + \cdots + X_n)/n$.

Test: to assess if a hypothesis such as "*a fair coin*" is true or not, which may be formally represented as

$$H_0: \ \pi = 0.5.$$

$\boxed{\text{The answer cannot be resulted from the estimator } \widehat{\pi}}$

If $\widehat{\pi} = 0.9$, H_0 is $\underline{\text{unlikely}}$ to be true

If $\widehat{\pi} = 0.45$, H_0 $\underline{\text{may}}$ be true (and also may be untrue)

If $\widehat{\pi} = 0.7$, what to do then?

A customer complaint: the amount of coffee in a Hilltop coffee bottle is less than the advertised weight 3 pounds.

Sample 20 bottles, yielding the average 2.897

Is this sufficient to substantiate the complaint?

Again statistical estimation cannot provide a satisfactory answer, due to random fluctuation among different samples

We cast the problem into a hypothesis testing problem:

Let the weight of coffee be a normal random variable $X \sim N(\mu, \sigma^2)$. We need to test the hypothesis $\mu < 3$. In fact, we use the data to test the hypothesis

$$H_0: \ \mu = 3 \quad (\text{or } H_0: \ \mu \geq 3)$$

If we could reject H_0, the customer complaint will be vindicated.

Suppose one is interested in estimating the mean income of a community. Suppose the income population is normal $N(\mu,\ 25)$ and a random sample of $n = 25$ observations is taken, yielding the sample mean $\bar{X} = 17$.

Three expert economists give their own opinions as follows:

- Mr A claims the mean income $\mu = 16$
- Mr B claims the mean income $\mu = 15$
- Mr C claims the mean income $\mu = 14$

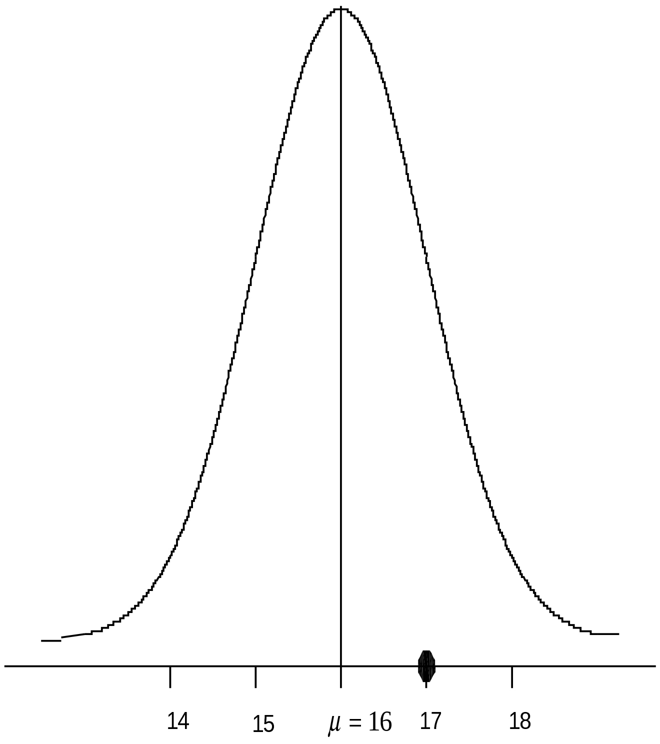
How would you assess those experts' statements?

Note. $\bar{X} \sim N(\mu, \sigma^2/n) = N(\mu, 1)$ — we assess the statements based on this distribution.

If Mr A's claim were correct, $\bar{X} \sim N(16,\ 1)$.

The observed value $\bar{X} = 17$ is *one standard deviation* away from μ, and may be regarded as a *typical observation* from the distribution.

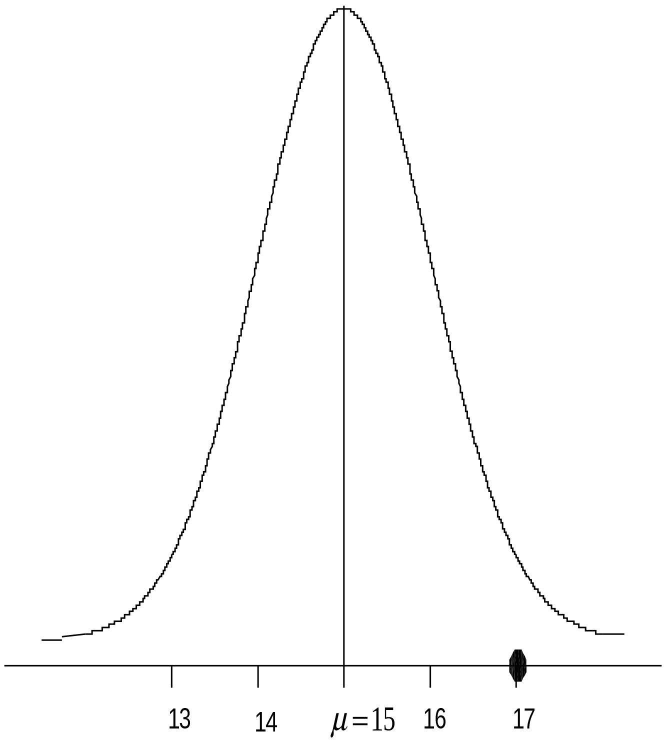
Little inconsistency between the claim and the data evidence.



If Mr B's claim were correct, $\bar{X} \sim N(15,\ 1)$.

The observed value $\bar{X} = 17$ begins to look *a bit extreme*, as it is *two standard deviation* away from μ.

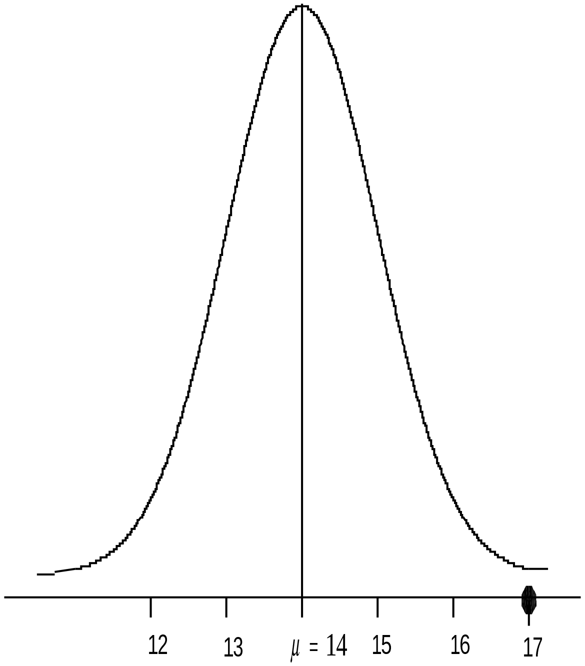
Inconsistency between the claim and the data evidence.



If Mr C’s claim were correct, $\bar{X} \sim N(14,\ 1)$.

The observed value $\bar{X} = 17$ is *extreme* indeed, as it is *three standard deviation* away from μ.

Strong inconsistency between the claim and the data evidence.



A measure of the discrepancy between the hypothesised (claimed) value for μ and the observed value $\bar{X} = x$ is the probability of observing $\bar{X} = x$ or more extreme values. This probability is called the p-value. That is

- under $H_0: \ \mu = 16$,
$$P(\bar{X} \geq 17) + P(\bar{X} \leq 15) = P(|\bar{X} - 16| \geq 1) = 0.317$$
- under $H_0: \ \mu = 15$,
$$P(\bar{X} \geq 17) + P(\bar{X} \leq 13) = P(|\bar{X} - 15| \geq 2) = 0.046$$
- under $H_0: \ \mu = 14$,
$$P(\bar{X} \geq 17) + P(\bar{X} \leq 11) = P(|\bar{X} - 14| \geq 3) = 0.003$$

In summary, we reject the hypothesis $\mu = 15$ or $\mu = 14$, as, for example, if the hypothesis $\mu = 14$ is true, the probability of observing $\bar{X} = 17$ or more extreme values is merely 0.003. We are comfortable with this decision, as *a small probability event would not occur in a single experiment*.

On the other hand, we cannot reject the hypothesis $\mu = 16$.

But this does not imply that this hypothesis is necessarily true, as, for example, $\mu = 17$ or 18 are at least as likely as $\mu = 16$.

Not Reject $\neq$ Accept

A statistical test is incapable to accept a hypothesis.

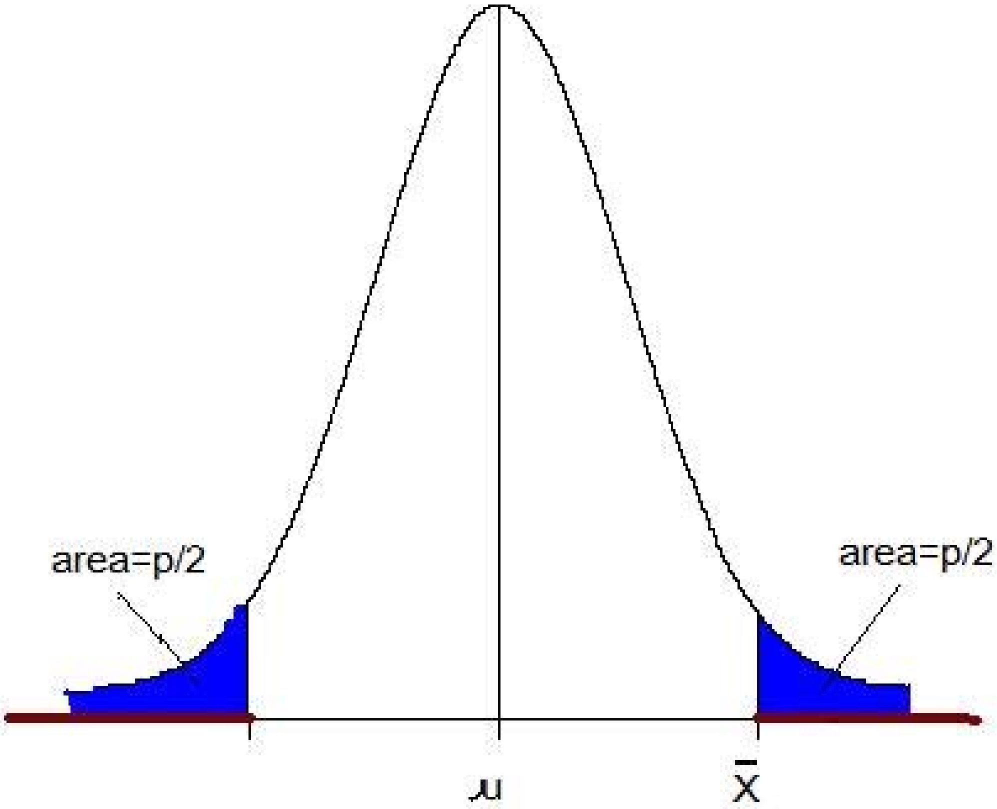
p-value: the probability of the event that a test statistic takes the observed value or more extreme (i.e. more unlikely) values under H_0

It is a measure of the discrepancy between a hypothesis and data.

p-value small: hypothesis is not supported by data

p-value large: hypothesis is not inconsistent with data

p-value may be seen as a risk measure of rejecting hypothesis H_0



General setting of hypothesis test

Let $\{X_1, \cdots, X_n\}$ be a random sample from a distribution $F(\cdot,\ \theta)$. We are interested in testing the hypotheses

$$H_0 :\ \theta = \theta_0 \qquad \text{vs} \qquad H_1 :\ \theta \in \Theta_1,$$

where θ_0 is a fixed value, Θ_1 is a set, and $\theta_0 \notin \Theta_1$.

- H_0 is called a null hypothesis
- H_1 is called an alternative hypothesis

Significance level α: a small number between 0 and 1 selected subjectively.

Often we choose $\alpha = 0.1,\ 0.05$ or 0.01, i.e. tests are often conducted as the significance levels 10%, 5% or 1%.

Decision: **Reject H_0 if p-value $\leq\ \alpha$**

Statistical testing procedure:

Step 1. Find a test statistic $T = T(X_1, \cdots, X_n)$. Denote T_0 the value of T with the given sample of observations.

Step 2. Compute the p-value, i.e.

$$p = P_{\theta_0}(\, T = T_0 \text{ or more extreme values}\,),$$

where P_{θ_0} denotes the probability distribution with $\theta = \theta_0$.

Step 3. If $p \leq \alpha$, reject H_0. Otherwise, H_0 is not rejected.

Remarks. 1. The alternative hypothesis H_1 is helpful to identify powerful test statistic T.

2. The significance level α controls how small is small for p-values.

3. “More extreme values" refers to those more unlikely values (than T_0) under H_0 in favour of H_1.

Example 1. Let $X_1, \cdots, X_{20}$, taking values either 1 or 0, be the outcomes of an experiment of tossing a coin 20 times, i.e.

$$P(X_i = 1) = \pi = 1 - P(X_i = 0), \quad \pi \in (0, 1).$$

We are interested in testing

$$H_0 : \pi = 0.5 \qquad \text{against} \qquad H_1 : \pi \neq 0.5.$$

Suppose there are 17 X_i's taking value 1, and 3 taking value 0. Will you reject the null hypothesis at the significance level 5%?

Let $Y = X_1 + \cdots + X_{20}$. Then $Y \sim Bin(20, \pi)$. We use Y as the test statistic.

With the given sample, we observe $Y = 17$. What are the more extreme values for Y if H_0 is true?

Under H_0, $EY = n\pi_0 = 10$. Hence 3 is as extreme as 17, and the more extreme values are

$$18, \ 19, \ 20, \ \text{ and } \ 0, \ 1, \ 2.$$

Thus the p-value is

$$\begin{aligned}
& \Big(\sum_{i=0}^{3} + \sum_{i=17}^{20}\Big) P_{H_0}(Y = i) \\
= \;& \Big(\sum_{i=0}^{3} + \sum_{i=17}^{20}\Big) \frac{20!}{i!(20-i)!}(0.5)^i(1-0.5)^{20-i} \\
= \;& 2 \times (0.5)^{20} \sum_{i=0}^{3} \frac{20!}{i!(20-i)!} \\
= \;& 2 \times (0.5)^{20} \times \{1 + 20 + 20 \times 19/2 + 20 \times 19 \times 18/(2 \times 3)\} \\
= \;& 0.0026.
\end{aligned}$$

Hence we reject the hypothesis of a fair coin at the significance level 1%.

Impact of H_1

In the above example, if we test

$$H_0 : \pi = 0.5 \qquad \text{against} \qquad H_1 : \pi > 0.5.$$

We should only reject H_0 if there is strong evidence against H_0 in favour of H_1. Having observed $Y = 17$, the more extreme values are 18, 19 and 20. Therefore the p-value is $\sum_{17 \leq i \leq 20} P_{H_0}(Y = i) = 0.0013$. Now the evidence against H_0 is even stronger.

On the other hand, if we test

$$H_0 : \pi = 0.5 \qquad \text{against} \qquad H_1 : \pi < 0.5.$$

The observation $Y = 17$ is more in favour of H_0 rather than H_1 now. We cannot reject H_0, as the p-value now is $\sum_{i \leq 17} P_{H_0}(Y = i) = 1 - 0.0013 = 0.9987$.

Remark. We only reject H_0 if there is significance evidence in favour of H_1.

Two types of errors

Statistical tests are often associated with two kinds of errors, which are displayed in the table below.

		Decision Made	
		H_0 not rejected	H_0 rejected
True State of Nature	H_0	Correct decision	Type I Error
	H_1	Type II Error	Correct decision

Remarks. 1. Ideally we would like to have a test that minimises the probabilities of making both types of errors, which unfortunately is not feasible.

2. The probability of making Type I error is the p-value and is not greater than α – the significance level. Hence it is under control.

3. We do not have an explicit control on the probability of Type II error. For a given significance level α, we choose a test statistic such that, hopefully, the probability of Type II error is small.

4. **Power**. The power function of the test is defined as

$$\beta(\theta) = P_\theta\{\ H_0 \text{ is rejected}\}, \quad \theta \in \Theta_1,$$

i.e. $\beta(\theta) = 1-$ Probability of Type II error.

5. **Asymmetry**: null hypothesis H_0 and alternative hypothesis H_1 are not treated equally in a statistical test. The choice of H_0 is based on the subject matter concerned and/or technical convenience.

6. It is more conclusive to end a test with H_0 rejected, as the decision of “Not Reject" does not imply that H_0 is accepted.

10.2 The Wald test

Suppose we would like to test $H_0 : \theta = \theta_0$, and $\widehat{\theta} = \widehat{\theta}(X_1, \cdots, X_n)$ is an estimator and is asymptotically normal, i.e.

$$(\widehat{\theta} - \theta)/\mathrm{SE}(\widehat{\theta}) \xrightarrow{D} N(0,1), \qquad \text{as } n \to \infty.$$

Then under H_0, $(\widehat{\theta} - \theta_0)/\mathrm{SE}(\widehat{\theta}) \sim N(0,1)$ approximately.

The Wald test at the significance levet α: Let $T = (\widehat{\theta} - \theta_0)/\mathrm{SE}(\widehat{\theta})$ be the test statistic. We reject H_0 against

$H_1 : \theta \neq \theta_0$ if $|T| > z_{\alpha/2}$ (i.e. the p-value $< \alpha$), or
$H_1 : \theta > \theta_0$ if $T > z_\alpha$ (i.e. the p-value $< \alpha$), or
$H_1 : \theta < \theta_0$ if $T < -z_\alpha$ (i.e. the p-value $< \alpha$),

where z_α is the top-α point of $N(0, 1)$, i.e. $P\{N(0, 1) > z_\alpha\} = \alpha$.

Remark. Since the Wald test is based on the asymptotic normality, it only works for reasonably large n.

Example 2. To deal with the customer complaint that the amount of coffee in a Hilltop coffee bottle is less than the advertised 3 pounds, 20 bottles were weighed, yielding observations

2.82, 3.01, 3.11, 2.71, 2.93, 2.68, 3.02, 3.01, 2.93, 2.56,
2.78, 3.01, 3.09, 2.94, 2.82, 2.81, 3.05, 3.01, 2.85, 2.79

The sample mean and standard deviation:

$$\bar{X} = 2.897, \qquad S = 0.148$$

Hence $\mathrm{SE}(\bar{X}) = 0.148/\sqrt{20} = 0.033$. By the CLT, $(\bar{X} - \mu)/\mathrm{SE}(\bar{X}) \xrightarrow{D} N(0, 1)$.

To test $H_0 : \mu = 3$ vs $H_1 : \mu < 3$, we apply the Wald test with $T = (\bar{X} - 3)/SE(\bar{X}) = -3.121 < -z_{0.01} = -2.326$. Hence we reject $H_0 : \mu = 3$ at the 1% significance level.

We conclude that there is significant evidence which supports the claim that the coffee in a Hilltop coffee bottle is less than 3 pounds.

10.3 χ^2-distribution and t-distribution

χ^2-Distributions

Background. χ^2-distribution is one of the important distributions in statistics. It is closely linked with normal, t- and F-distributions. Inference for variance parameter σ^2 relies on χ^2-distributions. More importantly most goodness-of-fit tests are based on χ^2-distributions.

Definition. Let $X_1, \cdots, X_k$ be independent $N(0,1)$ r.v.s. Let

$$Z = X_1^2 + \cdots + X_k^2 = \sum_{i=1}^{k} X_i^2.$$

The distribution of Z is called the χ^2-distribution with k degrees of freedom, denoted by $\chi^2(k)$ or χ_k^2.

We list some properties of the distribution χ_k^2 as follows.

1. χ_k^2 is a continuous distribution on $[0, \infty)$.

2. **Mean**: $EZ = kE(X_1^2) = k$.

3. **Variance**: $\mathrm{Var}(Z) = 2k$.

   Due to the independence among X_i's,

$$\mathrm{Var}(Z) = k\mathrm{Var}(X_1^2) = k[E(X_1^4) - \{E(X_1^2)\}^2] = k\{E(X_1^4) - 1\}.$$

$$
\begin{aligned}
E(X_1^4) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^4 e^{-x^2/2}dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^3 e^{-x^2/2}d(x^2/2) \\
&= -\frac{x^3}{\sqrt{2\pi}}e^{-x^2/2}\Big|_{-\infty}^{\infty} + \frac{3}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 e^{-x^2/2}dx = \frac{3}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 e^{-x^2/2}dx \\
&= \frac{3}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x e^{-x^2/2}d(x^2/2) = -\frac{3x}{\sqrt{2\pi}}e^{-x^2/2}\Big|_{-\infty}^{\infty} + \frac{3}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}dx \\
&= \frac{3}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}dx \ = \ 3.
\end{aligned}
$$

4. If $Z_1 \sim \chi_k^2$, $Z_2 \sim \chi_p^2$, and Z_1 and Z_2 are independent, then $Z_1 + Z_2 \sim \chi_{k+p}^2$.

According to the definition, we may write

$$Z_1 = \sum_{i=1}^{k} X_i^2, \qquad Z_2 = \sum_{j=k+1}^{k+p} X_j^2,$$

where all X_i's are independent $N(0,1)$ r.v.s. Hence

$$Z_1 + Z_2 = \sum_{i=1}^{k+p} X_i^2 \sim \chi^2_{k+p}.$$

5. The probability density function of χ_k^2 is

$$f(x) = \begin{cases} \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} & x > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where

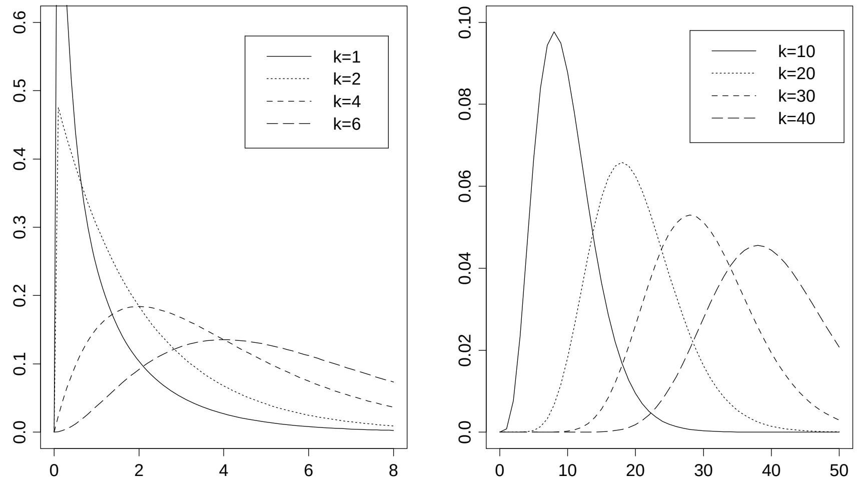
$$\Gamma(y) = \int_0^\infty u^{y-1} e^{-u} du.$$

For any integer k, $\Gamma(k) = (k-1)!$.

Hence χ_2^2 is the exponential distribution with mean 2, as its pdf is

$$f(x) = \begin{cases} \frac{1}{2} e^{-x/2} & x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Probability density functions of χ_k^2-distributions



6. The values of distribution functions of χ_k^2-distributions (for different k) have been tabulated, and can also be easily obtained from statistical packages such as R.

Let $Y_1, \cdots, Y_n$ be independent $N(\mu, \sigma^2)$ r.v.s. Then

$$(Y_i - \mu)/\sigma \sim N(0, 1).$$

Hence

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)^2 \sim \chi^2_n.$$

Note that

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 + \frac{n}{\sigma^2}(\bar{Y} - \mu)^2. \tag{9}$$

Since $\bar{Y} \sim N(\mu, \sigma^2/n)$, $\frac{n}{\sigma^2}(\bar{Y} - \mu)^2 \sim \chi^2_1$. It may be proved that

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sim \chi^2_{n-1}.$$

Thus decomposition (9) may be formally written as

$$\chi^2_n = \chi^2_{n-1} + \chi^2_1.$$

Confidence Interval for σ^2

Let $\{X_1, \cdots, X_n\}$ be a random sample from population $N(\mu, \sigma^2)$.

Let $M = \sum_{i=1}^n (X_i - \bar{X})^2$. Then $M/\sigma^2 \sim \chi^2_{n-1}$.

For any given small $\alpha \in (0,1)$, we may find $0 < K_1 < K_2$ such that

$$P(\chi^2_{n-1} < K_1) = P(\chi^2_{n-1} > K_2) = \alpha/2,$$

where χ^2_{n-1} stands for a r.v. with χ^2_{n-1}-distribution. Then

$$1 - \alpha = P(K_1 < M/\sigma^2 < K_2) = P(M/K_2 < \sigma^2 < M/K_1)$$

Hence an $100(1-\alpha)$% confidence interval for σ^2 is

$$(M/K_2, \; M/K_1).$$

Suppose $n = 15$ and the sample variance $S^2 = 24.5$. Let $\alpha = 0.05$.

From a table of χ^2-distributions, we may find

$$P(\chi^2_{14} < 5.629) = P(\chi^2_{14} > 26.119) = 0.025.$$

Hence a 95% confidence interval for σ^2 is

$$\begin{aligned}&(M/26.119,\ M/5.629) = (14S^2/26.119,\ 14S^2/5.629)\\ &= (0.536S^2,\ 2.487S^2) = (13.132,\ 60.934).\end{aligned}$$

In the above calculation, we have used the formula

$$S^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2 = \frac{1}{n-1}M = M/14.$$

Student's t-distribution

Background. Another important distribution in statistics

- The t-test is perhaps the most frequently used statistical test in application.

- Confidence intervals for normal mean with unknown variance may be *accurately* constructed based on t-distribution.

Historical note. The t-distribution was first studied by W.S. Gosset (1876-1937), who worked as a statistician for Guinness, writing under the pen-name 'Student'.

Definition. Suppose $X \sim N(0,1)$ and $Z \sim \chi_k^2$, and X and Z are independent. Then the distribution of the random variable

$$T = X \Big/ \sqrt{Z/k}$$

is called the t-distribution with k degrees of freedom, denoted by t_k or $t(k)$.

We now list some properties of the t_k distribution below.

1. t_k is a continuous and symmetric distribution on $(-\infty, \infty)$.

   (T and $-T$ share the same distribution.)

2. $E(T) = 0$ provided $E|T| < \infty$.

3. **Heavy tails.** If $T \sim t_k$, $E\{|T|^k\} = \infty$. For $X \sim N(\mu, \sigma^2)$, $E\{|X|^p\} < \infty$ for any $p > 0$. Therefore, t-distributions have heavier tails. This is a useful properties in modelling abnormal phenomena in financial or insurance data.

   Note. $E\{|T|^{k-\varepsilon}\} < \infty$ for any small constant $\varepsilon > 0$.

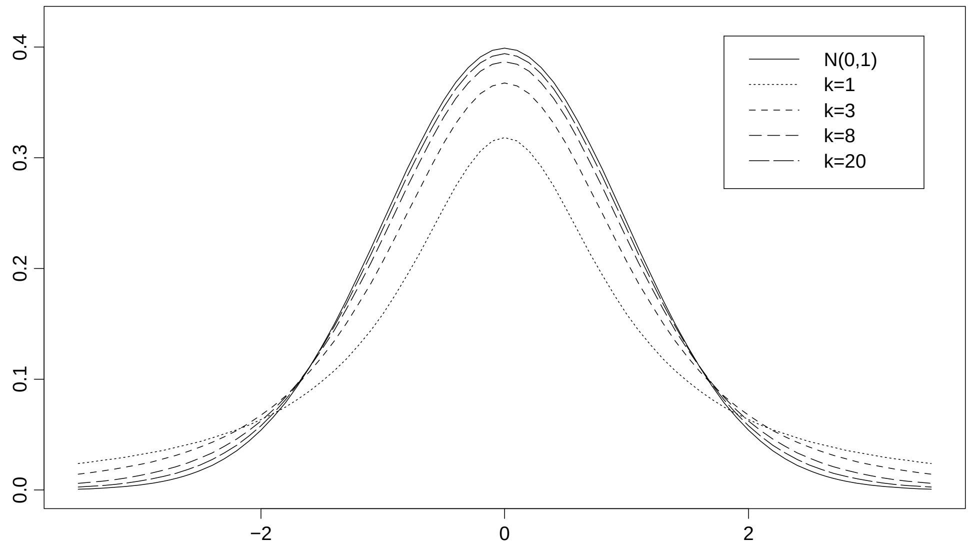
4. As $k \to \infty$, the distribution of t_k converges to the distribution of $N(0, 1)$.

   For $Z \sim \chi^2_k$, $Z = X_1^2 + \cdots + X_k^2$, where $X_1, \cdots, X_k$ are i.i.d. $N(0, 1)$. By the LLN, $Z/k \to E(X_1^2) = 1$. Thus $T = X/\sqrt{Z/k} \to X \sim N(0, 1)$.

5. The probability density function of t_k:

$$f(x) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})}\left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}.$$

Probability density functions of t_k-distributions



An important property of normal samples

Theorem. Let $\{X_1, \cdots, X_n\}$ be a sample from $N(\mu,\ \sigma^2)$. Let

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad \text{SE}(\bar{X}) = \frac{S}{\sqrt{n}}.$$

Then (i) $\bar{X} \sim N(\mu,\ \sigma^2/n)$, and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$,

(ii) $\bar{X}$ and S^2 are independent, and therefore

$$\frac{\sqrt{n}(\bar{X}-\mu)}{S} = \frac{\bar{X}-\mu}{\text{SE}(\bar{X})} \sim t_{n-1}.$$

The t-interval — an accurate $(1-\alpha)$ confidence interval for μ:

$$(\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}},\ \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}) = (\bar{X} - t_{\alpha/2,n-1}\cdot\text{SE}(\bar{X}),\ \bar{X} + t_{\alpha/2,n-1}\cdot\text{SE}(\bar{X})),$$

where $t_{\alpha/2,n-1}$ is a constant such that $P(t_{n-1} > t_{\alpha/2,n-1}) = \alpha/2$.

Proof of Theorem. Let $Y_i = (X_i - \mu)/\sigma$. Then $\bar{Y} = (\bar{X} - \mu)/\sigma$, and

$$S_y^2 \equiv \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y}) = S^2/\sigma^2.$$

Hence we only need to show that (a) $(n-1)S_y^2 \sim \chi^2_{n-1}$, and (b) $\bar{Y}$ and S_y^2 are independent.

As $\mathbf{Y} \equiv (Y_1, \cdots, Y_n)' \sim N(0, \mathbf{I}_n)$, it also holds that

$$\mathbf{Z} \equiv (Z_1, \cdots, Z_n)' \equiv \mathbf{\Gamma Y} \sim N(0, \mathbf{I}_n) \quad \text{for any orthogonal } \mathbf{\Gamma}.$$

Let $(\frac{1}{\sqrt{n}}, \cdots, \frac{1}{\sqrt{n}})$ be the first row of $\mathbf{\Gamma}$. Then $Z_1 = \sqrt{n}\bar{Y}$. Hence

$$(n-1)S_y^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 = \sum_{i=1}^{n} Z_i^2 - n\bar{Y}^2 = \sum_{i=2}^{n} Z_i^2 \sim \chi^2_{n-1},$$

and it is independent of $Z_1 = \sqrt{n}\bar{Y}$. □

The t-distributions with different degrees of freedom have been tabulated in all statistical tables.

The table below lists some values of C_α defined by the equation

$$P(t_k > C_\alpha) = \alpha$$

	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.005$
$k = 1$	6.314	12.706	63.657
$k = 2$	2.593	4.303	9.925
$k = 3$	2.353	3.182	5.841
$k = 10$	1.812	2.228	3.169
$k = 20$	1.725	2.086	2.845
$k = 120$	1.658	1.980	2.617
$\cdots$		$\cdots$	
$N(0,1)$	1,645	1.960	2.576

Remark. When $k \geq 120$, $t_k \approx N(0,1)$.

10.4 *t*-tests – one of the most frequently used tests in practice.

10.4.1 Tests for normal means – One-sample problems

Let $\{X_1, \cdots, X_n\}$ be a sample from $N(\mu,\ \sigma^2)$, where both μ and $\sigma^2 > 0$ are unknown. Test the hypotheses

$$H_0 : \mu = \mu_0 \qquad \text{against} \qquad H_1 : \mu \neq \mu_0,$$

where μ_0 is known.

The famous t-statistic:

$$T = \sqrt{n}(\bar{X} - \mu_0)/S = \sqrt{n}(\bar{X} - \mu_0)\Big/\Big\{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\Big\}^{1/2} = \frac{\bar{X} - \mu_0}{\text{SE}(\bar{X})},$$

where $\bar{X} = n^{-1}\sum_i X_i$ and $S^2 = \frac{1}{n-1}\sum_i(X_i - \bar{X})^2$. Note that under hypothesis H_0,

$$\sqrt{n}(\bar{X} - \mu_0)/\sigma \sim N(0, 1), \qquad (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}.$$

Therefore

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} = \frac{\sqrt{n}(\bar{X} - \mu_0)/\sigma}{\sqrt{S^2/\sigma^2}} \sim t_{n-1} \quad \text{under } H_0.$$

Hence we reject H_0 if $|T| > t_{\alpha/2,n-1}$, where α is the significance level of the test, and $t_{\alpha,k}$ is the top-α point of t_k-distribution, i.e. $P(t_k > t_{\alpha,k}) = \alpha$.

Remark. $H_0 : \mu = \mu_0$ is rejected against $H_1 : \mu \neq \mu_0$ at the α significance level iff μ_0 lies outside the $(1 - \alpha)$ t-interval $\bar{X} \pm t_{\alpha/2,n-1}\text{SE}(\bar{X})$.

Example 2. (Continue) We use t-test to re-examine this data set. Recall

$$n = 20, \quad \bar{X} = 2.897, \quad S = 0.148, \quad \mathrm{SE}(\bar{X}) = 0.033,$$

we are interested in testing hypotheses

$$H_0 : \mu = 3, \qquad H_1 : \mu < 3.$$

We reject H_0 at the level α if $T < -t_{\alpha,19}$. Since $T = (\bar{X} - 3)/\mathrm{SE}(\bar{X}) = -3.121 < -t_{0.01,19} = -2.539$, we reject the null hypothesis $H_0 : \mu = 3$ at 1% significance level.

10.4.2 Tests for normal means – two-sample problems

Available two independent samples: $\{X_1, \cdots, X_{n_x}\}$ from $N(\mu_x, \sigma_x^2)$ and $\{Y_1, \cdots, Y_{n_y}\}$ from $N(\mu_y, \sigma_y^2)$. We are interested in testing

$$H_0 : \mu_x - \mu_y = \delta \quad \text{against} \quad H_1 : \mu_x - \mu_y \neq \delta \text{ (or } \mu_x - \mu_y > \delta \text{ etc)},$$

where δ is a known constant. Let

$$\bar{X} = \frac{1}{n_x} \sum_{i=1}^{n_x} X_i, \quad S_x^2 = \frac{1}{n_x - 1} \sum_{i=1}^{n_x} (X_i - \bar{X})^2,$$

$$\bar{Y} = \frac{1}{n_y} \sum_{i=1}^{n_y} Y_i, \quad S_y^2 = \frac{1}{n_y - 1} \sum_{i=1}^{n_y} (Y_i - \bar{Y})^2.$$

Then

$$\bar{X} - \bar{Y} \sim N(\mu_x - \mu_y,\ \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}), \quad (n_x - 1)\frac{S_x^2}{\sigma_x^2} + (n_y - 1)\frac{S_y^2}{\sigma_y^2} \sim \chi^2_{n_x+n_y-2}.$$

With an addition assumption $\sigma_x^2 = \sigma_y^2$, it holds that

$$\sqrt{\frac{n_x + n_y - 2}{1/n_x + 1/n_y}} \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}} \sim t_{n_x + n_y - 2}$$

Define a t-statistic

$$T = \sqrt{\frac{n_x + n_y - 2}{1/n_x + 1/n_y}} \frac{\bar{X} - \bar{Y} - \delta}{\sqrt{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}}$$

The null hypothesis $H_0 : \mu_x - \mu_y = \delta$ is rejected against

$H_1 : \mu_x - \mu_y \neq \delta$ if $|T| > t_{\alpha/2,\, n_x + n_y - 2}$, or
$H_1 : \mu_x - \mu_y > \delta$ if $T > t_{\alpha,\, n_x + n_y - 2}$, or
$H_1 : \mu_x - \mu_y < \delta$ if $T < -t_{\alpha,\, n_x + n_y - 2}$,

where $t_{\alpha, k}$ is the top-α point of the t_k-distribution.

Example 3. Two types of razor, A and B, were compared using 100 men in an experiment. Each man shaved one side, chosen at random, of his face using one razor and the other side using the other razor. The times taken to shave, X_i and Y_i minutes, $i = 1, \cdots, 100$, corresponding to the razors A and B respectively, were recorded, yielding

$$\bar{X} = 2.84, \quad S_x^2 = 0.48, \quad \bar{Y} = 3.02, \quad S_y^2 = 0.42.$$

Also available is the sample variance of the differences $Z_i \equiv X_i - Y_i$: $S_z^2 = 0.6$.

Test, at the 5% significance level, if the two razors lead to different shaving times. State clearly the assumptions used in the test.

Assumption. Suppose $\{X_i\}$ and $\{Y_i\}$ are two samples from, respectively, $N(\mu_x,\ \sigma_x^2)$ and $N(\mu_y,\ \sigma_y^2)$.

The problem requires to test hypotheses

$$H_0 : \mu_x = \mu_y \qquad \text{vs} \qquad H_1 : \mu_x \neq \mu_y.$$

There are three approaches: a pairwise comparison method, two two-sample comparisons based on different assumptions. Since the data are recorded pairwisely, the pairwise comparison is most relevant and effective to analyse this data

Method I: Pairwise comparison — one sample t-test

Note $Z_i = X_i - Y_i \sim N(\mu_z,\ \sigma_z^2)$ with $\mu_z = \mu_x - \mu_y$. We test

$$H_0 : \mu_z = 0 \qquad \text{vs} \qquad H_1 : \mu_z \neq 0.$$

This is the standard one-sample t-test,

$$\sqrt{n}\frac{\bar{Z} - \mu_z}{S_z} = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{S_z/\sqrt{n}} \sim t_{n-1}.$$

H_0 is rejected if $|T| > t_{0.025,\,99} = 1.98$, where

$$T = \sqrt{n}\bar{Z} \,/\, S_z = \sqrt{100}(\bar{X} - \bar{Y})/S_z.$$

With the given data, we observe $T = 10(2.84 - 3.02)/\sqrt{0.6} = -2.327$. Hence we reject the hypothesis that the two razors lead to the same shaving time.

A 95% confidence interval for $\mu_x - \mu_y$:

$$\bar{X} - \bar{Y} \pm t_{0.025,\, n-1} S_z \big/ \sqrt{n} = -0.18 \pm 0.154 = (-0.334,\ -0.026).$$

Remark. (i) Zero is not in the confidence interval for $\mu_x - \mu_y$.

(ii) $t_{0.025,\,99} = 1.98$ is pretty close to $z_{0.025} = 1.96$. **Indeed when n is large, the t-test and the Wald test are almost the same.**

Method II: Two sample t-test with equal but unknown variance

Additional assumption: two samples are independent, $\sigma_x^2 = \sigma_y^2$.

Now $\bar{X} - \bar{Y} \sim N(\mu_x - \mu_y,\ \sigma_x^2/50)$, $99(S_x^2 + S_y^2)/\sigma_x^2 \sim \chi^2_{198}$. Hence

$$\frac{\sqrt{50}\{\bar{X} - \bar{Y} - (\mu_x - \mu_y)\}}{\sqrt{99(S_x^2 + S_y^2)/198}} = 10 \times \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{S_x^2 + S_y^2}} \sim t_{198}$$

Hence we reject H_0 if $|T| > t_{0.025,\,198} = 1.97$ where

$$T = 10(\bar{X} - \bar{Y})\Big/\sqrt{S_x^2 + S_y^2}.$$

For the given data, $T = -1.897$. Hence we cannot reject H_0.

A 95% confidence interval for $\mu_x - \mu_y$ contains 0:

$$(\bar{X} - \bar{Y}) \pm \frac{t_{0.025,\,198}}{10}\sqrt{S_x^2 + S_y^2} = -0.18 \pm 0.1870 = (-0.367,\ 0.007),$$

Method III: The Wald test — The normality assumption is not required. But the two samples are assumed to be independent. Note

$$\mathrm{SE}(\bar{X} - \bar{Y}) = \sqrt{S_x^2/n_1 + S_y^2/n_2}.$$

Hence it holds approximately that

$$\{\bar{X} - \bar{Y} - (\mu_x - \mu_y)\}/\mathrm{SE}(\bar{X} - \bar{Y}) \sim N(0,\ 1).$$

Hence, we reject H_0 when $|T| > 1.96$ at the 95% significance level, where

$$T = (\bar{X} - \bar{Y})/\sqrt{S_x^2/100 + S_y^2/100}.$$

For the given data, $T = -0.18/\sqrt{0.009} = -1.9$. Hence we <u>cannot</u> reject H_0.

An approximate 95% confidence interval for $\mu_x - \mu_y$ is

$$\bar{X} - \bar{Y} \pm 1.96 \times \sqrt{S_x^2/100 + S_y^2/100} = -0.18 \pm 0.186 = (-0.366,\ 0.006).$$

The value 0 is contained in the interval now.

Remarks. (i) Different methods lead to different but *not contradictory* conclusions, as

Not reject $\neq$ Accept

(ii) The pairwise comparison is intuitively most relevant, and leads to most conclusive inference (i.e. rejection). It also produces the shortest confidence interval.

(iii) Methods II and III ignore the pairing of the data, and therefore fail to take into account the variation due to the different individuals. Consequently the inference is less conclusive and less accurate.

(iv) A general observation: H_0 is rejected iff the hypothesized value by H_0 is not in the corresponding confidence interval.

(v) It is much more challenging to compare two normal means with unknown and different variances, which is not discussed in this course. On the other hand, the Wald test provides an easy alternative when both n_x and n_y are large.

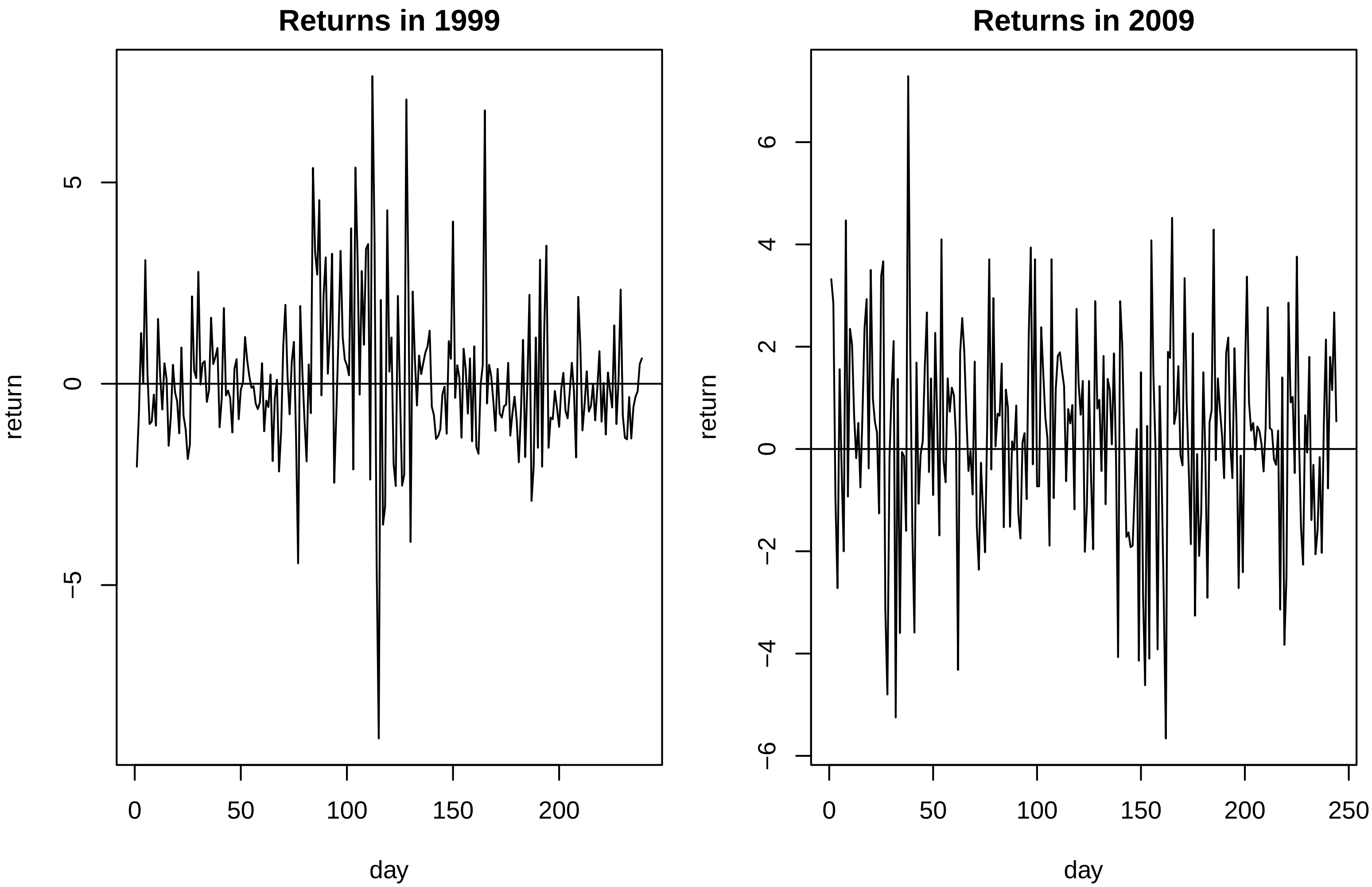
10.4.3. *t*-tests with *R*

The *R*-function `t.test` performs one-sample, or two-sample *t*-tests with one-sided or two-sided alternatives. We illustrate it via an example.

Example 4. The daily returns of the Shanghai Stock Exchange Composite Index in 1999 and 2009: two subsets of the data analysed in Example 1 of Chapter 8.

(i) First we extract the two subsets and conduct some preliminary data analysis.

```
> x <- read.table("shanghaiSECI.txt", skip=3, header=T)
> y <- x[,4]*100   # daily returns in percentages
> y1999 <- y[1005:1243] # extract daily returns in 1999
> y2009 <- y[3415:3658] # extract daily returns in 2009
> par(mar=c(4,4,2,1),mfrow=c(1,2))
> plot(y1999, type='l', xlab='day', ylab='return',
     main='Returns in 1999')
> plot(y2009, type='l', xlab='day', ylab='return',
     main='Returns in 2009')
> length(y1999); length(y2009)
[1] 239  # sample size of returns in 1999
[1] 244  # sample size of returns in 2009
> summary(y1999)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-8.8100 -0.8800 -0.1300  0.1037  0.7400  7.6400
> summary(y2009)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-5.6600 -0.8150  0.3650  0.2561  1.4780  7.2900
> var(y1999); var(y2009)
[1] 3.493598
[1] 3.922712
```


Returns in 1999
return
5
0
−5
0
50
100
150
200
day
Returns in 2009
return
6
4
2
0
−2
−4
−6
0
50
100
150
200
250
day

(ii) *One sample t-test*. Let X_i denote the returns in 1999, and Y_i denote the returns in 2009. Then $n_x = 239$, $n_y = 244$, and

$$\bar{X} = 0.1037, \quad \bar{Y} = 0.2561, \quad S_x^2 = 3.4936, \quad S_y^2 = 3.9227.$$

We test $H_0 : \mu_x = 0$ vs $H_1 : \mu_x > 0$ first.

```
> t.test(y1999, alternative='greater', mu=0, conf.level=0.95)
     # use alternative='two.sided' for two-sided alternative
        One Sample t-test
data:  y1999
t = 0.8576, df = 238, p-value = 0.196
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 -0.09596304         Inf
sample estimates:
mean of x
 0.103682
```

Since the p-value is 0.196, we cannot reject $H_0 : \mu_x = 0$, i.e. the returns in 1999 are not significantly different from 0.

Corresponding the one-sided alternative, *R* also gives a corresponding one-sided confidence interval for μ: $(-0.096, \infty)$, which contains 0. (Note that the setting indicates that we believe μ is either 0 or positive. Therefore reasonable confidence intervals are in the form (a, ∞).)

Now we test $H_0 : \mu_y = 0$ vs $H_1 : \mu_y > 0$.

```
> t.test(y2009, alternative='greater', mu=0, conf.level=0.99)
        One Sample t-test
data:  y2009
t = 2.0202, df = 243, p-value = 0.02223
alternative hypothesis: true mean is greater than 0
99 percent confidence interval:
 -0.04077725        Inf
sample estimates:
mean of x
0.2561475
```

For the returns in 2009, the p-value of the t-test is 0.022. Hence we reject $H_0 : \mu_y = 0$ at the 5% significance level, but cannot reject H_0 at the 1%

level. We conclude that there exists evidence indicating that the returns in 2009 tend to greater than 0, although the evidence is not overwhelming.

Remark. With the sample sizes over 200, the above t-tests yield practically the same results as the Wald test.

(iii) *Two-sample t-tests*. We now test $H_0 : \mu_x - \mu_y = 0$ against $H_1 : \mu_x - \mu_y \neq 0$ or $H_1 : \mu_x - \mu_y < 0$.

```
> t.test(y1999, y2009, mu=0, alternative='two.sided', var.equal=T)
      # without flag "var.equal=T", the WelchâĂŞSatterthwaite approximate
      # test will be used instead
        Two Sample t-test
data:  y1999 and y2009
t = -0.8697, df = 481, p-value = 0.3849
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4969197  0.1919887
sample estimates:
mean of x mean of y
0.1036820 0.2561475

> t.test(y1999, y2009, mu=0, alternative='less', var.equal=T)
        Two Sample t-test
data:  y1999 and y2009
t = -0.8697, df = 481, p-value = 0.1924
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf 0.1364386
sample estimates:
mean of x mean of y
0.1036820 0.2561475
```

Both the tests indicate that there is no significant evidence against the hypothesis that the average returns in the two years are the same.

10.5 Most Powerful Tests and Neyman-Pearson Lemma

Ideally we would choose, among those tests of size α, the test which minimises the probability of Type II error, i.e. that maximises the power $\beta(\theta)$ over $\theta \in \Theta_1$. If such a test exists, it is called *the most powerful test* (MPT).

Neyman-Pearson Lemma. If a test of size α for

$$H_0 : \ \theta = \theta_0 \quad \text{against} \quad H_1 : \ \theta = \theta_1$$

rejects H_0 when

$$L(\theta_1; \mathbf{x}) > K L(\theta_0; \mathbf{x}),$$

and does not reject H_0 when

$$L(\theta_1; \mathbf{x}) < K L(\theta_0; \mathbf{x}),$$

then it is a most powerful test of size α, where $K > 0$ is a constant.

Note. Both H_0 and H_1 are simple hypotheses.

Example 5 Let $X_1, X_2, \ldots, X_n$ be a sample from $N(\mu, 1)$. To test

$$H_0 : \mu = 0 \quad \text{against} \quad H_1 : \mu = 5,$$

the likelihood ratio is

$$\text{LR} = \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\sum_{i=1}^n (X_i - 5)^2/2\right)}{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\sum_{i=1}^n X_i^2/2\right)} \propto \exp(5n\bar{X}).$$

Thus $LR > K$ is equivalent to $\bar{X} > K_1$, K_1 is determined by the size of the test. Thus the MPT of size α rejects H_0 iff $\bar{X} > z_\alpha/\sqrt{n}$, where z_α is a top-α point of $N(0, 1)$.

Question: If we change the alternative hypothesis to $H_1 : \mu = 10$, what is the MPT then?

Uniformly Most Powerful Tests

Suppose that the MPT for testing

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta = \theta_1$$

does not change its form for all $\theta_1 \in \Theta_1$. Then it is the *Uniformly Most Powerful Test* (UMPT) for testing

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1.$$

Note. Typically, $\Theta_1 = (-\infty, \theta_0)$ or $\Theta_1 = (\theta_0, \infty)$.

Example 5 (continue). For

$$H_0 : \mu = 0 \quad \text{against} \quad H_1 : \mu > 0,$$

the UMPT of size α rejects H_0 iff $\bar{X} > z_\alpha$.

A more general case

Let $\mathbf{X} = (X_1, \cdots, X_n)^T$ be random variables with joint pdf $f(\mathbf{x}, \theta)$. We test the hypotheses

$$H_0: \ \theta \leq \theta_0 \quad \text{vs} \quad H_1: \ \theta > \theta_0. \tag{10}$$

Denoted by $T \equiv T(\mathbf{X})$ the MPT of size α for simple hypotheses

$$H_0: \ \theta = \theta_0 \quad \text{vs} \quad H_1: \ \theta = \theta_1,$$

exists, where $\theta_1 > \theta_0$.

Then T is the UMPT of the same size α for hypotheses (10) provided that

(i) T remains unchaged for all values of $\theta_1 > \theta_0$, and

(ii) $P_\theta(T \text{ rejects } H_0) \leq P_{\theta_0}(T \text{ rejects } H_0) = \alpha$ for all $\theta < \theta_0$.

Note. For hypotheses $H_0 : \theta \geq \theta_0$ vs $H_1 : \theta < \theta_0$, the UMPT may be obtained in the similar manner.

Example 6. Let $(X_1, \cdots, X_n)$ be a random sample from an exponential distribution with mean $1/\lambda$. We are interested in testing

$$H_0 : \lambda \leq \lambda_0 \quad \text{vs} \quad H_1 : \lambda > \lambda_0.$$

For

$$H_0 : \lambda = \lambda_0 \quad \text{vs} \quad H_1 : \lambda = \lambda_1,$$

the MPT rejects H_0 iff $\sum_{i=1}^n X_i \leq K$ for any $\lambda_1 > \lambda_0$, where K is determined by $P_{\lambda_0}\{\sum_{i=1}^n X_i < K\} = \alpha$.

It is easy to verify that for $\lambda < \lambda_0$, $P_\lambda\{\sum_{i=1}^n X_i < K\} < \alpha$.

Hence the MPT for the simple null hypothesis against simple alternative is also the UMPT for the composite hypotheses.

Chapter 11. Hypothesis Testing (II)

11.1 Likelihood Ratio Tests — one of the most popular ways of constructing tests when both null and alternative hypotheses are composite (i.e. not a single point).

Let $\mathbf{X} \sim f(\cdot, \boldsymbol{\theta})$. Consider hypotheses

$$H_0 : \ \boldsymbol{\theta} \in \Theta_0 \quad \text{vs} \quad H_1 : \ \boldsymbol{\theta} \in \Theta - \Theta_0.$$

The likelihood ratio test will reject H_0 for the large values of the statistic

$$LR \ = \ LR(\mathbf{X}) \ \equiv \ \frac{\sup_{\boldsymbol{\theta}\in\Theta} f(\mathbf{X}, \boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}\in\Theta_0} f(\mathbf{X}, \boldsymbol{\theta})} = f(\mathbf{X}, \widehat{\boldsymbol{\theta}}) \big/ f(\mathbf{X}, \widetilde{\boldsymbol{\theta}}),$$

where $\widehat{\boldsymbol{\theta}}$ the (unconstrained) MLE, and $\widetilde{\boldsymbol{\theta}}$ is the constrained MLE under hypothesis H_0.

Remark. (i) It is easy to see that $LR \geq 1$.

(ii) The exact sampling distributions of LR are usually unknown, except in a few special cases.

Example 1. **(One-sample t-test)**
Let $\mathbf{X} = (X_1, \cdots, X_n)^\tau$ be a random sample from $N(\mu, \sigma^2)$. We are interested in testing hypotheses

$$H_0: \ \mu = \mu_0 \quad \text{against} \quad H_1: \ \mu \neq \mu_0,$$

where μ_0 is given, and σ^2 is unknown and is a nuisance parameter. Now both H_0 and H_1 are composite. The likelihood function is

$$L(\mu, \sigma^2) = C\sigma^{-n} \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n} (X_i - \mu)^2 \Big\}.$$

The unconstrained MLEs are

$$\widehat{\mu} = \bar{X}, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(X_i - \bar{X})^2,$$

and the constrained MLE is

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(X_i - \mu_0)^2.$$

The LR-ratio statistic is then

$$LR = \frac{L(\widehat{\mu}, \widehat{\sigma}^2)}{L(\mu_0, \widetilde{\sigma}^2)} = (\widetilde{\sigma}^2/\widehat{\sigma}^2)^{n/2}.$$

Since

$$n\widetilde{\sigma}^2 = n\widehat{\sigma}^2 + n(\bar{X} - \mu_0)^2,$$

it holds that $\widetilde{\sigma}^2/\widehat{\sigma}^2 = 1 + T^2/(n-1)$, where

$$T = \sqrt{n}(\bar{X} - \mu_0) \Big/ \Big\{ \frac{1}{n-1} \sum_{j=1}^{n} (X_i - \bar{X})^2 \Big\}^{1/2}.$$

Note that $T \sim t_{n-1}$ under H_0. The LRT will reject H_0 iff $|T| > t_{n-1,\alpha/2}$, where $t_{k,\alpha}$ is the upper α-point of the t-distribution with k degrees of freedom.

Asymptotic Distribution of Likelihood ratio test statistic

Let $\mathbf{X} = (X_1, \cdots, X_n)^\tau$, and assume certain regularity conditions. Then as $n \to \infty$, the distribution of $2\log(LR)$ under H_0 converges to the χ^2-distribution with $d - d_0$ degrees of freedom, where d is the ‘dimension’ of Θ and d_0 is the ‘dimension’ of Θ_0.

To make the computation of ‘dimension’ easy, **reparametrisation** is often adopted. Suppose that the parameter θ may be written in two parts

$$\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$$

where $\boldsymbol{\psi}$ is $k \times 1$ parameter of interest, and $\boldsymbol{\lambda}$ is of little interest and is called *nuisance parameters*. The hypotheses to be tested may be expressed as

$$H_0: \ \boldsymbol{\psi} = \boldsymbol{\psi}_0 \quad \text{vs} \quad H_1: \ \boldsymbol{\psi} \neq \boldsymbol{\psi}_0.$$

Now the LR-statistic is of the form

$$LR = \frac{L(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\lambda}}; \mathbf{X})}{L(\boldsymbol{\psi}_0, \widetilde{\boldsymbol{\lambda}}; \mathbf{X})},$$

where $(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\lambda}})$ is unconstrained MLE while $\widetilde{\boldsymbol{\lambda}}$ is the constrained MLE of $\boldsymbol{\lambda}$ subject to $\boldsymbol{\psi} = \boldsymbol{\psi}_0$. Then as $n \to \infty$,

$$2\log(LR) \xrightarrow{D} \chi_k^2 \quad \text{under } H_0.$$

Example 2. Let $X_1, \cdots, X_n$ be independent, and $X_j \sim N(\mu_j, 1)$. Consider the null hypothesis

$$H_0: \ \mu_1 = \cdots = \mu_n.$$

The likelihood function is

$$L(\mu_1, \cdots, \mu_n) = C \ \exp\Big\{ -\frac{1}{2}\sum_{j=1}^{n}(X_j - \mu_j)^2\Big\},$$

where $C > 0$ is a constant independent of μ_j. Then the unconstrained MLE are $\widehat{\mu}_j = X_j$ and the constrained MLE is $\widetilde{\mu} = \bar{X}$. Hence

$$LR = \frac{L(\widehat{\mu}_1, \cdots, \widehat{\mu}_n)}{L(\widetilde{\mu}, \cdots, \widetilde{\mu})} = \exp\Big\{\frac{1}{2}\sum_{j=1}^{n}(X_j - \bar{X})^2\Big\}.$$

Hence

$$2\log(LR) = \sum_{j=1}^{n}(X_j - \bar{X})^2 \sim \chi^2_{n-1} \quad \text{under } H_0,$$

which is true for any finite n as well.

How to *calculate the degree of freedom*?

Since $d = n,\ d_0 = 1$, the d.f. is $d - d_0 = n - 1$.

Alternatively we may adopt the following reparametrisation:

$$\mu_j = \mu_1 + \psi_j \quad \text{for } 2 \le j \le n.$$

Then the null hypothesis can be expressed as

$$H_0: \ \psi_2 = \cdots = \psi_n = 0.$$

Therefore $\boldsymbol{\psi} = (\psi_2, \cdots, \psi_n)^\tau$ has $n - 1$ component, i.e. $k = n - 1$.

11.2 The permutation test — a nonparametric method for testing if two distributions are the same. It is particularly appealing when sample sizes are small, as it does not rely on any asymptotic theory.

Let $X_1, \cdots, X_m$ be sample from distribution F_x and $Y_1, \cdots, Y_n$ be a sample from distribution F_y. We are interested in testing

$$H_0 : F_x = F_y \qquad \text{versus} \qquad H_1 : F_x \neq F_y.$$

Key idea: under H_0, $\{X_1, \cdots, X_m, Y_1, \cdots, Y_n\}$ form a sample of size $m + n$ from a single distribution.

Choose a test statistic

$$T = T(X_1, \cdots, X_m, Y_1, \cdots, Y_n)$$

which is capable to tell the difference between the two distribution, e.g. $T = |\bar{X} - \bar{Y}|$, or $T = |\bar{X} - \bar{Y}|^2 + |S_x^2 - S_y^2|$.

Consider all $(m+n)!$ permutations of $(X_1, \cdots, X_m, Y_1, \cdots, Y_n)$, compute the test statistic T for each permutation, yielding the values $T_1, \cdots, T_{(m+n)!}$.

The p-value of the test is defined as

$$p = \frac{1}{(m+n)!} \sum_{j=1}^{(m+n)!} I(T_j > t_{obs}),$$

where $t_{obs} = T(X_1, \cdots, X_m, Y_1, \cdots, Y_n)$. We reject H_0 at the significance level α if $p \leq \alpha$.

Note. When H_0 holds, all those $(m+n)!$ T_j's are on the equal footing, and $t_{obs} = T(X_1, \cdots, X_m, Y_1, \cdots, Y_n)$ is one of them. Therefore t_{obs} is unlikely to be an extreme value among T_j's.

Algorithm for Permutation Tests:

1. Compute $t_{obs} = T(X_1, \cdots, X_m, Y_1, \cdots, Y_n)$.

2. Randomly permute the data. Compute T again using the permuted date.

3. Repeat Step 2 B times, and let $T_1, \cdots, T_B$ denote the resulting values.

4. The approximate p-value is $B^{-1} \sum_{1 \le j \le B} I(T_j > t_{obs})$.

Remark. Let $Z = (X_1, \cdots, X_m, Y_1, \cdots, Y_n)$ (`Z <- c(X,Y)`). A permutation of Z may be obtained in R as

```
Zp <- sample(Z, n+m)
```

You may also use the R-function sample.int:

```
k <- sample.int(n+m, n+m)
```

Now k is a permutation of $\{1, 2, \cdots, n+m\}$.

Example 3. Class A was taught using detailed PowerPoint slides. The marks in the final exam are

$$45,\ 55,\ 39,\ 60,\ 64,\ 85,\ 80,\ 64,\ 48,\ 62,\ 75,\ 77,\ 50.$$

Students in Class B were required to read books and answer questions in class discussions. The marks in the final exam are

$$45,\ 59,\ 48,\ 74,\ 73,\ 78,\ 66,\ 69,\ 79,\ 81,\ 60,\ 52.$$

Can we infer that the marks from the two classes are significantly different?

We conduct the permutation test using the test statistic $T = |\bar{X} - \bar{Y}|$ in R:

```
> x <- c(45, 55, 39, 60, 64, 85, 80, 64, 48, 62, 75, 77, 50)
> y <- c(45, 59, 48, 74, 73, 78, 66, 69, 79, 81, 60, 52)
> length(x); length(y)
[1] 13
```

```
[1] 12
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  39.00   50.00   62.00   61.85   75.00   85.00
> summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  45.00   57.25   67.50   65.33   75.00   81.00
> Tobs <- abs(mean(x)-mean(y))
> z <- c(x,y)
> k <- 0
> for(i in 1:5000) {
+ zp <- sample(z, 25)       # zp is a permutation of z
+ T <- abs(mean(zp[1:13])-mean(zp[14:25]))
+ if(T>Tobs) k <- k+1
+ }
cat("p-value:", k/5000, "\n")
p-value: 0.5194
```

Since *p*-value is 0.5194, we cannot reject the null-hypothesis that the mark distributions of the two classes are the same.

We also apply the *t*-sample, obtaining the similar results:

```
> t.test(x, y, var.equal=T)   #  mu=0 is the default
        Two Sample t-test
data:  x and y
t = -0.6472, df = 23, p-value = 0.5239
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-14.632967   7.658608
```

11.3 χ^2-tests

11.3.1 Goodness-of-fit tests: to test if a given distribution fits the data.

Let $\{X_1, \cdots, X_n\}$ be a random sample from a discrete distribution of k categories denoted by $1, \cdots, k$. Denote the probability function

$$p_j = P(X_i = j), \qquad j = 1, \cdots, k.$$

Then $p_j \ge 0$ and $\sum_{j=1}^k p_j = 1$.

Typically $n >> k$. Therefore the data are often compressed into a table:

Category	1	2	$\cdots$	k
Frequency	Z_1	Z_2	$\cdots$	Z_k

where

$$Z_j = \text{ No. of } X_i\text{'s equal to } j, \qquad j = 1, \cdots, k.$$

Obviously $\sum_{j=1}^{k} Z_j = n$.

To test the null hypothesis

$$H_0: \; p_i = p_i(\theta), \qquad i = 1, \cdots, k,$$

where the function forms of $p_i(\theta)$ are known but the parameter θ is unknown. For example, $p_i(\theta) = \theta^{i-1} e^{-\theta}/(i-1)!$ (i.e. Poisson distribution).

We first estimate θ by, for example, its MLE $\widehat{\theta}$. The expected frequencies under H_0 are

$$E_i = n p_i(\widehat{\theta}), \quad i = 1, \cdots, k.$$

Listing them together with observed frequencies, we have

Category	1	2	$\cdots$	k
Frequency	Z_1	Z_2	$\cdots$	Z_k
Expected frequency	E_1	E_2	$\cdots$	E_k

If H_0 is true, we expect $Z_j \approx E_j = np_j(\widehat{\theta})$ when n is large, as, by the LLN, it holds

$$\frac{Z_j}{n} = \frac{1}{n}\sum_{i=1}^{n} I(X_i = j) \to E\{I(X_i = j)\} = P(X_i = j) = p_j(\theta).$$

Test statistic: $T = \sum_{j=1}^{k}(Z_j - E_j)^2/E_j$.

Theorem. Under H_0, $T \xrightarrow{D} \chi^2_{k-1-d}$ as $n \to \infty$, where d is the number of components in θ.

Remark. (i) It is important that $E_i \geq 5$ at least. This may be achieved by combining together the categories with smaller expected frequencies.

(ii) When p_j are completely specified (i.e. known) under H_0, $d = 0$.

Example 4. A supermarket recorded the numbers of arrivals over 100 one-minute intervals. The data were summarized as follows

No. of arrivals	0	1	2	3	4	5	7
Frequency	13	29	32	20	4	1	1

Do the data match a Poisson distribution?

The null hypothesis is $H_0: \ p_i = \lambda^i e^{-\lambda}/i!$ for $i = 0, 1, \cdots$. We find the MLE for λ first.

The likelihood function: $L(\lambda) = \prod_{i=1}^{100} \frac{\lambda^{X_i}}{X_i!} e^{-\lambda} \propto \lambda^{\sum_{i=1}^{100} X_i} e^{-100\lambda}$.

The log-likelihood function: $l(\lambda) = \log(\lambda) \sum_{i=1}^{100} X_i - 100\lambda$.

Let $\frac{d}{d\lambda} l(\lambda) = 0$, leading to $\widehat{\lambda} = \frac{1}{100} \sum_{i=1}^{100} X_i = \bar{X}$.

Since we are only given the counts Z_j instead of X_i, we need to compute $\bar{X}$ from Z_j. Recall Z_j = no. of X_i equal to j. Hence

$$\begin{aligned} \bar{X} &= \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} \sum_{j=1}^{k} j \cdot Z_j \\ &= \frac{1}{100} (0 \times 13 + 1 \times 29 + 2 \times 32 + 3 \times 20 + 4 \times 4 \\ &\quad +5 \times 1 + 7 \times 1) = 1.81. \end{aligned}$$

With $\widehat{\lambda} = 1.81$, the expected frequencies are

$$E_i = n \cdot p_i(\widehat{\lambda}) = 100 \times \frac{(1.81)^i}{i!} e^{-1.81}, \quad i = 0, 1, \cdots .$$

We combine the last three categories to make sure $E_i \geq 5$.

No. of arrivals	0	1	2	3	≥ 4	Total
Frequency Z_i	13	29	32	20	6	100
$p_i(\widehat{\lambda}) = \widehat{\lambda}^i e^{-\widehat{\lambda}}/i!$	0.164	0.296	0.268	0.162	0.110	1
Expected frequency E_i	16.4	29.6	26.8	16.2	11.0	100
Difference $Z_i - E_i$	-3.4	-0.6	5.2	3.8	-5	0
$(Z_i - E_i)^2/E_i$	0.705	0.012	1.01	0.891	2.273	4.891

Note under H_0, $T = \sum_{i=0}^{4}(Z_i - E_i)^2/E_i \sim \chi^2_{5-1-1} = \chi^2_3$. Since $T = 4.891 < \chi^2_{0.10,3} = 6.25$, we cannot reject the assumption that the data follow a Poisson distribution.

Remark. (i) The goodness-of-fit test has been widely used in practice. However we should bear in mind that when H_0 cannot be rejected, *we are not in the position to conclude that the assumed distribution is true,* as

$$\text{“not reject”} \neq \text{“accept"}$$

(ii) The above test may be used to test the goodness-of-fit of a continuous distribution via discretization. However there exist more appropriate methods such as *Kolmogorov-Smirnov test* and *Cramér-von Mises test*, which deal with the goodness-of-fit for continuous distributions directly.

11.3.2 Tests for contingency tables

Tests for independence of two discrete random variables

Let (X, Y) be two discrete random variables, and X have r categories and Y have c categories. Let

$$p_{ij} = P(X = i,\ Y = j), \quad i = 1, \cdots, r,\ j = 1, \cdots, c.$$

Then $p_{ij} \geq 0$ and $\sum_{i,j} p_{ij} \equiv \sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} = 1.$

Let $p_{i\cdot} = P(X = i)$ and $p_{\cdot j} = P(Y = j).$ It is easy to see that

$$p_{i\cdot} = \sum_{j=1}^{c} P(X = i,\ Y = j) = \sum_{j=1}^{c} p_{ij} = \sum_{j} p_{ij}$$

Similarly, $p_{\cdot j} = \sum_i p_{ij}$

X and Y are independent iff

$$p_{ij} = p_{i\cdot}p_{\cdot j} \text{ for } i = 1, \cdots, r \text{ and } j = 1, \cdots, c.$$

Suppose we have n pairs of observations from (X, Y). The data are presented in a contingency table below

		Y			
		1	2	$\cdots$	c
	1	Z_{11}	Z_{12}	$\cdots$	Z_{1c}
X	2	Z_{21}	Z_{22}	$\cdots$	Z_{2c}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	r	Z_{r1}	Z_{r2}	$\cdots$	Z_{rc}

where Z_{ij} = no. of the pairs equal to (i, j).

It is often useful to add the marginals into the table:

$$Z_{i\cdot} = \sum_{j=1}^{c} Z_{ij}, \quad Z_{\cdot j} = \sum_{i=1}^{r} Z_{ij}, \quad Z_{\cdot\cdot} = \sum_{i=1}^{r} Z_{i\cdot} = \sum_{j=1}^{c} Z_{\cdot j} = n$$

		Y				
		1	2	$\cdots$	c	
X	1	Z_{11}	Z_{12}	$\cdots$	Z_{1c}	$Z_{1\cdot}$
	2	Z_{21}	Z_{22}	$\cdots$	Z_{2c}	$Z_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	r	Z_{r1}	Z_{r2}	$\cdots$	Z_{rc}	$Z_{r\cdot}$
		$Z_{\cdot 1}$	$Z_{\cdot 2}$	$\cdots$	$Z_{\cdot c}$	$Z_{\cdot\cdot} = n$

We are interested in testing the independence

$$H_0: \; p_{ij} = p_{i\cdot}p_{\cdot j}, \; i = 1, \cdots, r, \; j = 1, \cdots, c.$$

Under H_0, a natural estimator for p_{ij} is

$$\widetilde{p}_{ij} = \widehat{p}_{i\cdot}\widehat{p}_{\cdot j} = \frac{Z_{i\cdot}}{n}\frac{Z_{\cdot j}}{n}$$

Hence the expected frequency at the (i, j)-th cell is

$$E_{ij} = n\widetilde{p}_{ij} = Z_{i\cdot}Z_{\cdot j}/n = Z_{i\cdot}Z_{\cdot j}/Z_{\cdot\cdot}, \quad i = 1, \cdots, r, \;\; j = 1, \cdots, c.$$

If H_0 is true, we expect $Z_{ij} \approx E_{ij}$. The goodness-of-fit test statistic is defined as

$$T = \sum_{i=1}^{r}\sum_{j=1}^{c}(Z_{ij} - E_{ij})^2/E_{ij}.$$

We reject H_0 for large values of T.

Under H_0, $T \sim \chi^2_{p-d}$, where

- $p =$ no. of cells -1 $= rc - 1$
- $d =$ no. of estimated 'free' parameters $= r + c - 2$.

Note. 1. The sum of $r \times c$ counts Z_{ij} is n fixed. So knowing $rc - 1$ of them, the other one is also known. This is why $p = rc - 1$.

2. The estimated parameters are $p_{i\cdot}$ and $p_{\cdot j}$. But $\sum_{i=1}^r p_{i\cdot} = 1$ and $\sum_{j=1}^c p_{\cdot j} = 1$. Hence $d = (r-1) + (c-1) = r + c - 2$.

3. For testing independence, it always holds that

$$Z_{i\cdot} - E_{i\cdot} = 0 \qquad \text{and} \qquad Z_{\cdot j} - E_{\cdot j} = 0.$$

Those are useful facts in checking for computational errors. The proofs are simple, as, for example,

$$Z_{i\cdot} - E_{i\cdot} = Z_{i\cdot} - \sum_j E_{ij} = Z_{i\cdot} - \sum_j \frac{Z_{i\cdot} Z_{\cdot j}}{Z_{\cdot\cdot}} = Z_{i\cdot} - \frac{Z_{i\cdot} Z_{\cdot\cdot}}{Z_{\cdot\cdot}} = 0.$$

Theorem. Under H_0, the limiting distribution of T is χ^2 with $(r-1)(c-1)$ degrees of freedom, as $n \to \infty$.

Example. The table below lists the counts on the beer preference and gender of beer drinker from randomly selected 150 individuals. Test at the 5% significance level the hypothesis that the preference is independent of the gender.

		Beer preference			
		Light ale	Lager	Bitter	Total
Gender	Male	20	40	20	80
	Female	30	30	10	70
	Total	50	70	30	150

The expected frequencies are:

$$E_{11} = \frac{80 \cdot 50}{150} = 26.67, \quad E_{12} = \frac{80 \cdot 70}{150} = 37.33, \quad E_{13} = \frac{80 \cdot 30}{150} = 16,$$

$$E_{21} = \frac{70 \cdot 50}{150} = 23.33, \quad E_{22} = \frac{70 \cdot 70}{150} = 32.67, \quad E_{33} = \frac{70 \cdot 30}{150} = 14.$$

E_{ij}				
	26.67	37.33	16	80
	23.33	32.67	14	70
	50	70	30	150

$Z_{ij} - E_{ij}$				
	-6.67	2.67	4	0
	6.67	-2.67	-4	0
	0	0	0	0

$(Z_{ij} - E_{ij})^2/E_{ij}$				
	1.668	0.191	1.000	2.859
	1.907	0.218	1.142	3.267
				6.126

Under the null hypothesis of independence, $T = \sum_{i,j}(Z_{ij} - E_{ij})^2/E_{ij} \sim \chi^2_2$. Note the degree freedom is $(2-1)(3-1) = 2$.

Since $T = 6.126 > \chi^2_{0.05,\,2} = 5.991$, we reject the null hypothesis, i.e. there is significant evidence from the data indicating that the beer preference and the gender of beer drinker are not independent.

Tests for several binomial distributions

Consider a real example: Three independent samples of sizes 80, 120 and 100 are taken respectively from single, married, and widowed or divorced persons. Each individual was asked to if “friends and social life” or “job and primary activity” contributes most to their general well-being. The counts from the three samples are summarized in the table below.

	Single	Married	Widowed or divorced
Friends and social life	47	59	56
Job or primary activity	33	61	44
Total	80	120	100

Conditional Inference: Sometimes we conduct inference under the assumption that all the row (or column) margins are fixed.

Different from the tables for independent tests, now

$$Z_{1j} \sim Bin(Z_{\cdot j},\ p_{1j}), \qquad j = 1, 2, 3,$$

where $Z_{\cdot j}$ are fixed constants — sample sizes. Furthermore, $p_{2j} = 1 - p_{1j}$.

We are interested in testing hypothesis

$$H_0 :\ p_{11} = p_{12} = p_{13}.$$

Under H_0, the three independent samples may be seen from the same population. Furthermore,

$$Z_{11} + Z_{12} + Z_{13} \sim Bin(Z_{\cdot 1} + Z_{\cdot 2} + Z_{\cdot 3},\ p),$$

where p denotes the common value of p_{11}, p_{12} and p_{13}.

Therefore the MLE is

$$\widehat{p} = \frac{Z_{11} + Z_{12} + Z_{13}}{Z_{\cdot 1} + Z_{\cdot 2} + Z_{\cdot 3}} = \frac{47 + 59 + 56}{80 + 120 + 100} = 0.54.$$

The expected frequencies are

$$E_{1j} = \widehat{p} Z_{\cdot j} \quad \text{and} \quad E_{2j} = Z_{\cdot j} - E_{1j}, \qquad j = 1, 2, 3.$$

E_{ij}			
	43.2	64.8	54.0
	36.8	55.2	46.0
Total	80	120	100

$Z_{ij} - E_{ij}$			
	3.8	-5.8	2.0
	-3.8	5.8	-2.0
Total	0	0	0

$(Z_{ij} - E_{ij})^2/E_{ij}$				
	0.334	0.519	0.074	
	0.392	0.609	0.087	
Total	0.726	1.128	0.161	2.015

Under H_0, $T = \sum_{i,j}(Z_{ij} - E_{ij})^2/E_{ij} \sim \chi^2_{p-d} = \chi^2_2$, where

- p = no. of free counts $Z_{ij} = 3$
- d = no. of estimated free parameters = 1.

Since $T = 2.015 < \chi^2_{0.10,2} = 4.605$, we cannot reject H_0, i.e. there is no significant difference among the three populations in terms of choosing between F&SL and J&PA as the more important factor towards their general well-being.

Remark. Similar to the independence tests, it holds that $Z_{i\cdot} - E_{i\cdot} = 0$ and $Z_{\cdot j} - E_{\cdot j} = 0$.

Tests for $r \times c$ tables – a general description

In general, we may test for different types of the structure in a $r \times c$ table, for example, the symmetry ($p_{ij} = p_{ji}$).

The key is to compute expected frequencies E_{ij} under null hypothesis H_0.

Under H_0, the test statistic

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(Z_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{p-d},$$

- p = no. of 'free' counts among Z_{ij} ,
- d = no. of the estimated 'free' parameters.

We reject H_0 if $T > \chi^2_{\alpha,\, p-d}$.

Remark. The R-function `chisq.test` performs both the goodness-of-fit test and the contingency table test.