L.1: Intro to R and Review of Probability 1

Econometrics 1: ver. 2024 Fall Semester

Naoki Awaya

Introduction to **R**

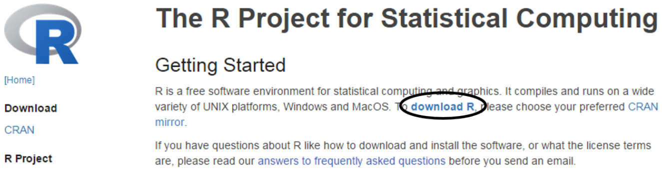
What is R?

- **R** is a free and open source software for statistical analysis.
- **R** has no license limitations. You can install and run it anytime and anywhere.
- One of the greatest advantages of **R** over other softwares is that **R** users can freely distribute their own original packages through CRAN (Comprehensive R Archive Network, "KRAN" or "SEE-RAN").
- We can implement a wide variety of brand new statistical methods quite easily just by downloading them from CRAN.
- For now, **R** and **Python** are two of the most popular programming languages used in statistical analysis.[1]

[1] Python is a general purpose programming language, which is not specialized for statistical analysis but can be used for a variety of purposes.

How to install R on Windows

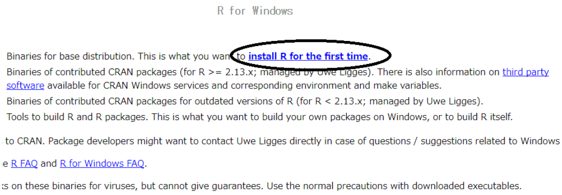
- Go to the website of **R** “The R Project for Statistical Computing”: https://www.r-project.org.
- Click on the link **download R**.



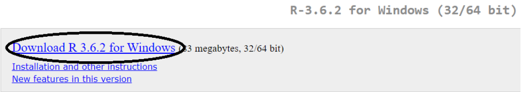
- Then, you will be asked to which server you want to connect. Choose “Japan - The Institute of Statistical Mathematics, Tokyo”
- Click on the link **Download R for Windows**.

How to install R on Windows

- Click on the link **install R for the first time**.



- Click on the link **Download R X.X.X for Windows**, where X.X.X gives the version of **R**.



- Then, the installer will be downloaded as “R-X.X.X-win.exe”.

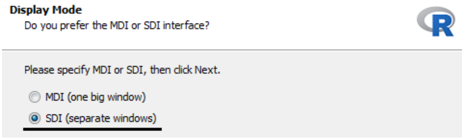
How to install R on Windows

- Double-click the downloaded installer to launch the installer.
- Click “Next” several times.
- (Optional) At the page that says “Startup options”, choose

  ”Yes (customized startup)”

  The default choice is “No (accept defaults)”.
- The next page says “Display Mode”. Choose

  ”SDI (separate windows)”



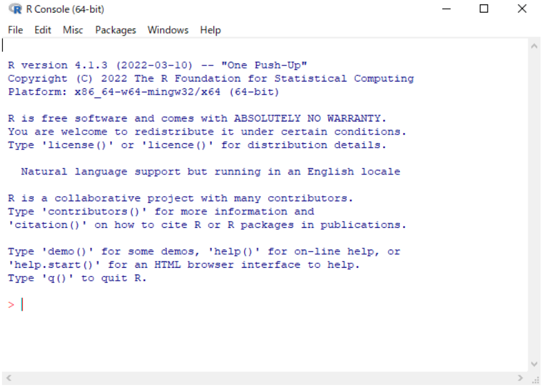
- The other options can be left as defaults.

R-Studio

- **R-Studio** is a software that provides a more efficient and user-friendly programming environment for using **R**.
- It includes a code editor, debugging, visualization tools, and R markdown editor.[2]
- Although it is not mandatory, I would recommend using **R-Studio**.

[2] R markdown is a tool that allows you to integrate R codes and their outputs in a document file and presentation slides. For example, these lecture slides are created using R markdown.

- When you start **R**, the window that first appears is called the R console.
- You can type or paste commands here. The console window also displays the results of the commands and error reports (if any).



Basics of R

- Anything following a hash (sharp) sign $\#$ is ignored and it is not processed by **R**. This can be used to include comments.
- If you want to write more than one command in a single line, you can use a semicolon ; as a command delimiter.

```
> 1 + 5 # addition
```

```
## [1] 6
```

```
> 8 - 2 # subtraction
```

```
## [1] 6
```

```
> 4 * 6 ; 3 / 7 # multiplication and division
```

```
## [1] 24

## [1] 0.4285714
```

Basics of R

- If you want to assign a number "a" to the variable "X", you can write

```
X <- a
```

- The assign symbol consists of two separate characters < and -, "less than" and "minus" with no space between them.

```
> sqrt(5*(1 + exp(2)) + tan(0.5))
```

```
## [1] 6.518557
```

```
> X <- sqrt(5*(1 + exp(2)) + tan(0.5))
> X
```

```
## [1] 6.518557
```

```
> 2*X
```

```
## [1] 13.03711
```

Basics of R

- Virtually any type of **R** objects (vector, matrix, data frames, functions, texts, etc) can be stored in a single object.

```
> A <- "Hoshino" # Texts must be enclosed in " ".
> A
```

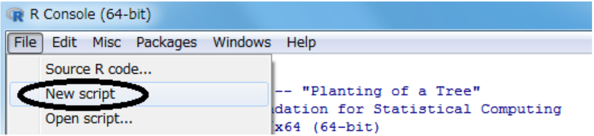
```
## [1] "Hoshino"
```

```
> A + 1 # You cannot add a number to a text.
```

```
## Error in A + 1: non-numeric argument to binary operator
```

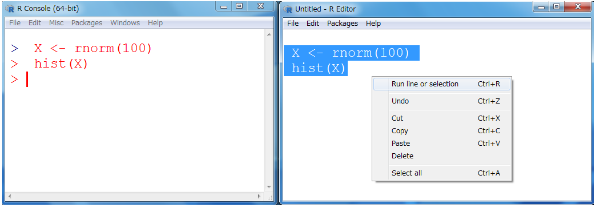
Script files

- When long and involved calculations are needed, typing each command directly into the console is inconvenient and error-prone.
- Besides, once the console window is closed all the commands you have executed will disappear
- A script file is a text file that contains a sequence of commands. You can execute the commands directly from the script file all at once.
- To create a new script file, click on "File" in the menu bar and then click "New Script".



Script file

- Once the commands are typed in the script, select the part you want to execute.
- Right click and choose "Run Line or Selection" or press [Ctrl] and [R] at the same time.[3]

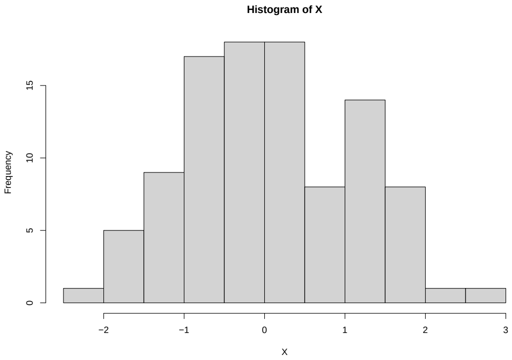


- `rnorm(100)`: draw 100 random numbers from the standard normal distribution, $N(0, 1)$.
- `hist(X)`: create a histogram of `X` (a new window will pop up).

[3]R-studio users: [Ctrl] + [Enter]

Basics of R

```
> X <- rnorm(100)
> hist(X)
```



Basics of R

Script file (cont.)

- To save the script file, click on "File" in the menu bar in the script editor, and then click "Save as". (The file extension is "r".)
- Note that double-clicking the script file does NOT open the **R** console window. To open the saved script file, launch **R** first, and choose "Open script" in the "File" in the menu bar.
- If you want to open the script file only, you can use any text editor like Notepad

If the .r-extension is associated with R-studio, you can launch R-studio just by double-clicking the r file.

Random Variables

Random Variables

- A random variable is a variable whose values are determined probabilistically.
- Examples of random variables are:
  - outcomes of rolling dice;
  - number of car accidents in a day;
  - income of a randomly sampled person.
- A more formal definition of random variable is as follows:

Random Variable

A random variable is a "function" that maps the outcomes of a random experiment to a numerical value.

Random Variables

- A random variable is a "rule" (i.e., function) that associates a number with each outcome in the set of possible outcomes.
- The set of possible outcomes is called the sample space.
  - Sample space is not necessarily a set of numbers.
- In the following, we use a capital letter, say X, to denote a random variable. A realized value of X is denoted by a small letter x.
- The set of possible realizations of X, $\{x_1, x_2, \dots,\}$, is called the support, and denoted by $\mathcal{X}$.

Random variable X	Sample space	Support $\mathcal{X}$
Outcome of rolling a dice	$\{1, 2, 3, 4, 5, 6\}$	$\{1, 2, 3, 4, 5, 6\}$
No. of heads when tossing a coin twice	$\{HH, TH, HT, TT\}$	$\{0, 1, 2\}$

Random Variables

Discrete Random Variable

A discrete random variable is a random variable where its support is a discrete set.

Continuous Random Variable

A continuous random variable is a random variable where its support is an interval (or a collection of intervals).

- Intuitively, a discrete random variable is a random variable whose possible values can be written down in a list.
- For example, letting X denote the height (in meters) of a randomly selected person, then X is a continuous random variable.
    - When X is continuous, there are infinitely many possible values in $\mathcal{X}$.

Random Variables

A dice roll simulation

- Dice roll outcome = Uniform random variable on $\mathcal{X} = \{1, 2, \ldots, 6\}$
  - Each element of $\mathcal{X}$ occurs with equal probability, $1/6$.

```
> #install.packages("extraDistr")
> library(extraDistr)
> rdunif(2, 1, 6)
```

```
## [1] 3 5
```

- `install.packages()`: install a new package. You need to run this code only once.
- We can use **extraDistr** package to compute discrete uniform distributions.
- `library()`: load the specified package.
- `rdunif(2, 1, 6)`: draw two X's from Uniform$\{1, 2, \ldots, 6\}$

Probability Distribution

Probability Distribution

Cumulative Distribution Function: CDF

The probability that the random variable X takes a value less than or equal to x, $\Pr(X \leq x)$, is called the cumulative distribution function (CDF) or simply the distribution function of X, and we denote

$$F(x) = \Pr(X \leq x).$$

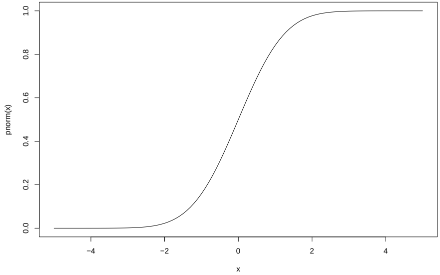
NOTE: x can be any value, even outside the support $\mathcal{X}$:

- E.g., $X =$ dice roll, $F(7) = 1$, $F(-1) = 0$, $F(3.1) = 0.5$, etc...

Probability Distribution

CDF of the standard normal distribution (i.e., normal distribution with mean 0 and standard deviation 1)

```
> curve(pnorm(x), xlim = c(-5,5))
```

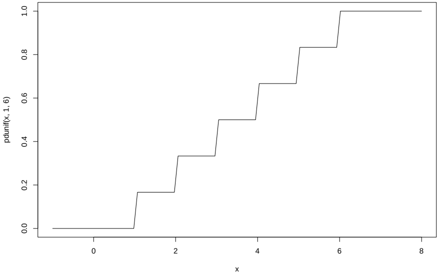


- `pnorm(x)`: standard normal CDF
- `curve`: depict the trajectory of the function given in the first argument.

Probability Distribution

CDF of dice roll $X \sim$ Uniform$\{1, 2, \ldots, 6\}$

```
> curve(pdunif(x, 1, 6), xlim = c(-1,8))
```



- [p] $+ \sim =$ distribution function of $\sim$
- [d] $+ \sim =$ density function of $\sim$
- [r] $+ \sim =$ draw a number from $\sim$

Probability Distribution

Properties of CDF

- For any X, the probability that X takes a value smaller (larger) than negative (positive) infinity is zero (one):

$$\lim_{x \to -\infty} F(x) = 0, \qquad \lim_{x \to \infty} F(x) = 1$$

- CDF is non-decreasing:

$$x_1 \leq x_2 \Rightarrow \Pr(X \leq x_1) \leq \Pr(X \leq x_2) \iff F(x_1) \leq F(x_2)$$

- In the following, we will mainly consider only continuous random variables (i.e., CDF is not a step function).

Probability Distribution

- The probability that X takes a value within an interval $[a, b]$ is

$$\begin{aligned}\Pr(X \in [a, b]) = \Pr(a \leq X \leq b) &= \Pr(X \leq b) - \Pr(X \leq a) \\ &= F(b) - F(a)\end{aligned}$$

- For a continuous random variable X, what is the probability that X takes a specific value, say $\Pr(X = 1)$?
- Intuitively, it is expected that $\Pr(X = 1)$ can be approximated by $\Pr(X \in [1, 1 + h])$ for a sufficiently small $h > 0$:[4] namely,

$$\begin{aligned}\Pr(X = 1) &\approx \Pr(X \in [1, 1 + h]) \\ &= F(1 + h) - F(1)\end{aligned}$$

[4] The probability that a person's height is 1.7 (m) should be approximately equal to the probability that his height is in between $[1.7, 1.700001]$ (m).

Probability Density

- For example, let F be the standard normal CDF. Then,

```
> Fdiff <- function(h) pnorm(1 + h) - pnorm(1)
> Fdiff(2); Fdiff(0.1); Fdiff(0.001)
```

```
## [1] 0.1573054

## [1] 0.02298919

## [1] 0.0002418497
```

(`function(xxx)`: create an original function of `xxx`)

- As h approaches to zero, the probability gets smaller and smaller; in the limit the probability will be exactly 0.
- This is due to the continuity of $F(\cdot)$: $a \to b \Rightarrow F(a) \to F(b)$
- Hence, if X is continuous, $\Pr(X = x) = 0$ holds <u>for any x</u>!

Probability Density

- On the other hand, if we divide $F(x+h) - F(x)$ by h, the value does not degenerate to zero but converges to a specific value:

```
> Fdiff2 <- function(h) (pnorm(1 + h) - pnorm(1))/h
> Fdiff2(0.1); Fdiff2(0.01); Fdiff2(0.001)
```

```
## [1] 0.2298919
```

```
## [1] 0.2407609
```

```
## [1] 0.2418497
```

Probability Density Function: PDF

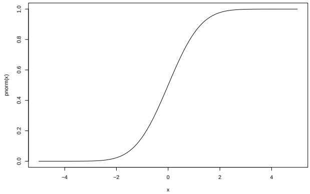
The probability density function (PDF) or simply the density function of X is defined as the "derivative" of the CDF of X:

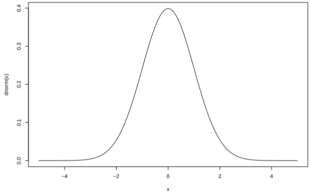
$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h} = \frac{\partial F(x)}{\partial x}$$

Probability Density

CDF and PDF of the standard normal distribution.

```
> curve(pnorm(x), xlim = c(-5,5)); curve(dnorm(x), xlim = c(-5,5))
```

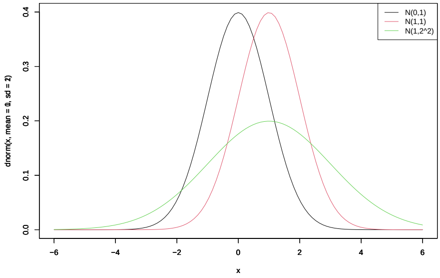




Probability Density

PDFs of normal distributions.

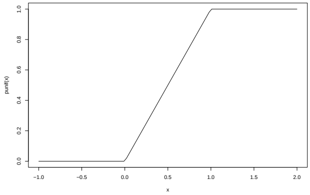
```
> curve(dnorm(x, mean = 0, sd = 1), xlim = c(-6,6), ylim = c(0, 0.4), col = 1)
> par(new = TRUE)
> curve(dnorm(x, mean = 1, sd = 1), xlim = c(-6,6), ylim = c(0, 0.4), col = 2)
> par(new = TRUE)
> curve(dnorm(x, mean = 1, sd = 2), xlim = c(-6,6), ylim = c(0, 0.4), col = 3)
> legend("topright", c("N(0,1)", "N(1,1)", "N(1,2^2)"),
+        lty = c(1,1,1), col = c(1,2,3))
```

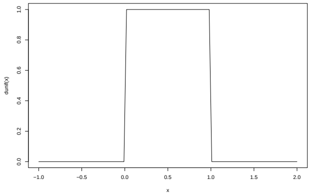


Probability Density

CDF and PDF of continuous Uniform$[0, 1]$.

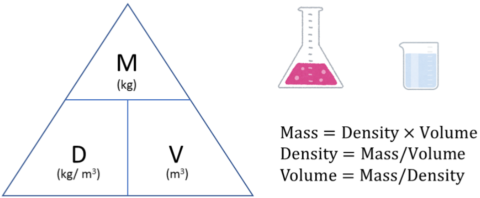
```
> curve(punif(x), xlim = c(-1,2)); curve(dunif(x), xlim = c(-1,2))
```





Probability Density

In what sense "density" function?



- The probability density of the event $\{X \in [x, x+h]\}$:

$$\frac{\Pr(X \in [x, x+h])}{h} = \frac{F(x+h) - F(x)}{h}$$

- Hence, $f(x)$ is the probabilistic mass density of X at x.

Probability Density

- More intuitively, one can interpret the value of the density $f(x)$ as

$$f(x): \text{ How likely } X \text{ takes a value in the neighborhood of } x \in \mathcal{X}$$

- Note that $f(x)$ is NOT a probability, and thus its value can be larger than one but never be negative.

- When X is a discrete random variable, we can easily compute the probability $\Pr(X = x)$ for any $x \in \mathcal{X}$.
- The function $p(x) = \Pr(X = x)$ is called the probability mass function.

X	Measurement of the "likelihood" of the event $\{X = x\}$
Discrete	probability mass function $p(x)$
Continuous	probability density function $f(x)$

Probability Density

Since PDF is the derivative of CDF, the following properties are straightforward.

Properties of PDF

- $F(x)$ is non-decreasing $\Rightarrow f(x) \geq 0$ for all $x \in \mathcal{X}$
- $F(a) = \int_{-\infty}^{a} f(x)dx \Rightarrow$

$$\begin{aligned} \Pr(X \in [a, b]) &= F(b) - F(a) \\ &= \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx = \int_{a}^{b} f(x)dx \end{aligned}$$

- $F(\infty) = 1 \Rightarrow \int_{-\infty}^{\infty} f(x)dx = 1$

Probability Density

```r
> pnorm(1) # Standard normal CDF
```

```
## [1] 0.8413447
```

```r
> integrate(dnorm, -Inf, 1)
```

```
## 0.8413448 with absolute error < 1.5e-05
```

```r
> punif(0.2) # Uniform[0,1] CDF
```

```
## [1] 0.2
```

```r
> integrate(dunif, -Inf, 0.2)
```

```
## 0.2 with absolute error < 9.4e-05
```

- `integrate(h,a,b)` $= \int_a^b h(x)dx$
- Note: when $X \sim$ Uniform$[0,1]$, $\Pr(X \leq a) = a$ for any $a \in [0,1]$.

Expectation and Variance

Expectation

Expectation

Let X be a continuous random variable with density function $f(x)$. The expectation of X, also called the mean of X, is defined as

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx.$$

- NOTE $\int_{-\infty}^{\infty} x f(x) dx = \int_{\mathcal{X}} x f(x) dx$ (because $f(x) = 0$ for $x \notin \mathcal{X}$).

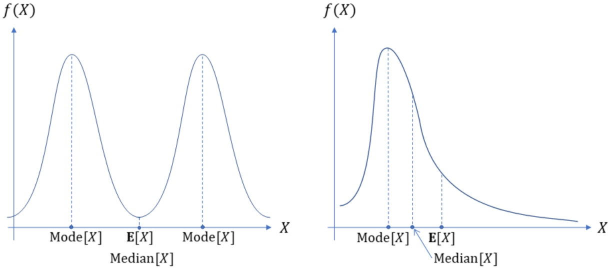
In the case of discrete random variable X, the expectation of X is given by

$$\mathbb{E}[X] = \sum_{i=1}^{k} x_i p(x_i),$$

where $p(x_i) = \Pr(X = x_i)$, and $\{x_1, \dots, x_k\} = \mathcal{X}$.

Expectation

- Note that the expectation is NOT the value that can be "expected" to occur.



- Median$[X] = x^*$ such that $F(x^*) = 0.5$ (i.e., the 50-percentile point of X).
- Mode$[X] = x^*$ such that $f(x^*) = \max_x f(x)$.

Expectation

Example: If $X \sim$ Uniform$[a, b]$, then $\mathbb{E}[X] = (b + a)/2$.

Proof The PDF of uniform distribution on $[a, b]$ is

$$f(x) = \begin{cases} (b-a)^{-1} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

By the definition of expectation, we have

$$\begin{aligned} \mathbb{E}[X] &= \int_a^b x \cdot (b-a)^{-1} \cdot dx \\ &= [0.5 \cdot x^2]_a^b \cdot (b-a)^{-1} \\ &= 0.5 \cdot (b^2 - a^2) \cdot (b-a)^{-1} = (b+a)/2 \quad \blacksquare \end{aligned}$$

- In general, when a random variable X follows a symmetric distribution, the expectation $\mathbb{E}[X]$ coincides with the center of symmetry.

Expectation

```
> h <- function(x) x * dunif(x, -2, 4) # X ~ Uniform[-2, 4]
> integrate(h, -Inf, Inf) # E(X)
```

```
## 1 with absolute error < 8.3e-05
```

```
> h <- function(x) x * dnorm(x, mean = 2, sd = 1) # X ~ Normal(2, 1)
> integrate(h, -Inf, Inf) # E(X)
```

```
## 2 with absolute error < 1.2e-05
```

Expectation

Expectation of a function of a random variable

For a random variable X, the expectation of $g(X)$, $\mathbb{E}[g(X)]$, is given by

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

The proof is complicated and is omitted.

Expectation

- An important special case is when $g(X)$ is an indicator function:

$$g(X) = \mathbf{1}\{X \leq a\} = \begin{cases} 1 & \text{if } X \leq a \\ 0 & \text{if } X > a \end{cases}$$

- In this case, we have $\mathbb{E}(\mathbf{1}\{X \leq a\}) = F(a)$:

$$\begin{aligned} \mathbb{E}(\mathbf{1}\{X \leq a\}) &= \int_{-\infty}^{\infty} \mathbf{1}\{x \leq a\} f(x)dx \\ &= \int_{-\infty}^{a} 1 \cdot f(x)dx + \int_{a}^{\infty} 0 \cdot f(x)dx = F(a). \end{aligned}$$

- In general, for an event A, $\mathbb{E}[\mathbf{1}\{A\}] = \Pr(A)$.

Expectation

```
> g <- function(x) ifelse(x <= 1, 1, 0) # 1{x <= 1}
> f <- function(x) g(x)*dnorm(x)
> integrate(f, -Inf, Inf) # E(g(x))
```

```
## 0.8413447 with absolute error < 4.7e-05
```

```
> pnorm(1)
```

```
## [1] 0.8413447
```

- `ifelse(A, x, y)`: if the condition `A` is true, return `x`, otherwise return `y`.

Expectation

Linearity of expectation

For a random variable X and constants a and b,

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

This property is called the linearity of expectation.

Proof Since $aX + b$ is a function of X,

$$\begin{aligned}
\mathbb{E}[aX + b] &= \int_{-\infty}^{\infty} (ax + b) f(x) dx \\
&= a \underbrace{\int_{-\infty}^{\infty} x f(x) dx}_{\mathbb{E}[X]} + b \underbrace{\int_{-\infty}^{\infty} f(x) dx}_{1} \\
&= a\mathbb{E}[X] + b \quad \blacksquare
\end{aligned}$$

Variance

Variance

- For a random variable X, the (population) variance of X is defined by

$$\mathbb{V}(X) = \mathbb{E}[\{X - \mathbb{E}(X)\}^2] = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

- In particular, if $\mathbb{E}(X) = 0$, $\mathbb{V}(X) = \mathbb{E}(X^2)$.

When a sample $\{X_1, \dots, X_n\}$ of n observations is available, the sample variance is given by

$$\text{Sample}\mathbb{V}(X_1, \dots, X_n) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$$

The definition of the variance of a random variable is obtained by replacing the sample average $n^{-1}\sum_{i=1}^{n}$ with the expectation $\mathbb{E}$.

Standard Deviation

Standard deviation

The square root of the variance of X, $\sqrt{\mathbb{V}(X)}$, is called the standard deviation of X.

It is often convenient to standardize X by subtracting its expectation $\mathbb{E}(X)$ and dividing by its standard deviation $\sqrt{\mathbb{V}(X)}$. The standardized random variable has mean zero and variance one:

$$Z = \frac{X - \mathbb{E}(X)}{\sqrt{\mathbb{V}(X)}} \text{ : standardization of } X$$

$$\mathbb{E}(Z) = \mathbb{E}\left[\frac{X - \mathbb{E}(X)}{\sqrt{\mathbb{V}(X)}}\right] = \frac{\mathbb{E}(X) - \mathbb{E}(X)}{\sqrt{\mathbb{V}(X)}} = 0$$

$$\mathbb{V}(Z) = \frac{\mathbb{E}[\{X - \mathbb{E}(X)\}^2]}{\mathbb{V}(X)} = \frac{\mathbb{V}(X)}{\mathbb{V}(X)} = 1$$