Psychol. Statistics using R

TAIPEI MEDICAL UNIVERSITY 1960

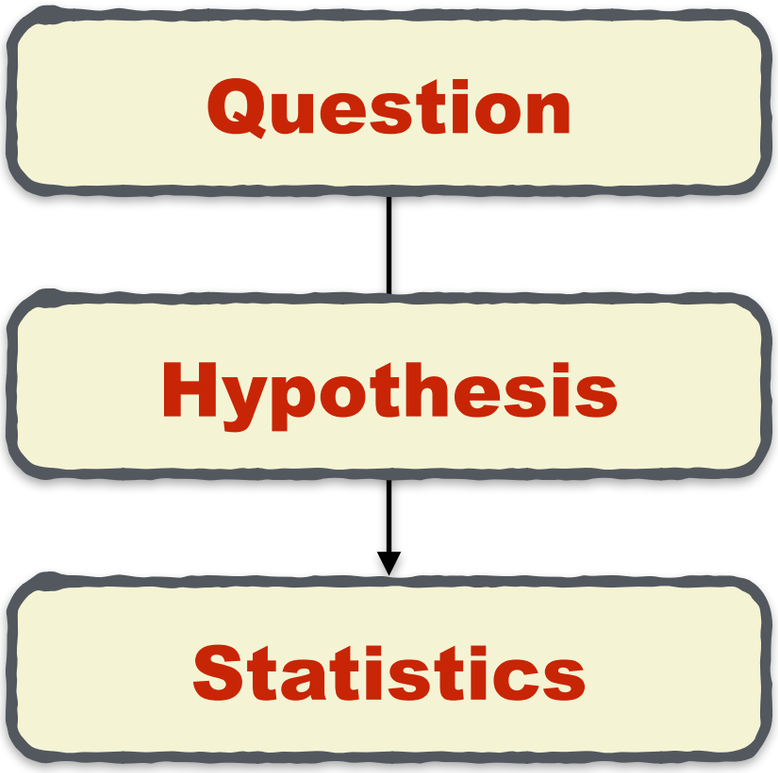
臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

Probability Distributions

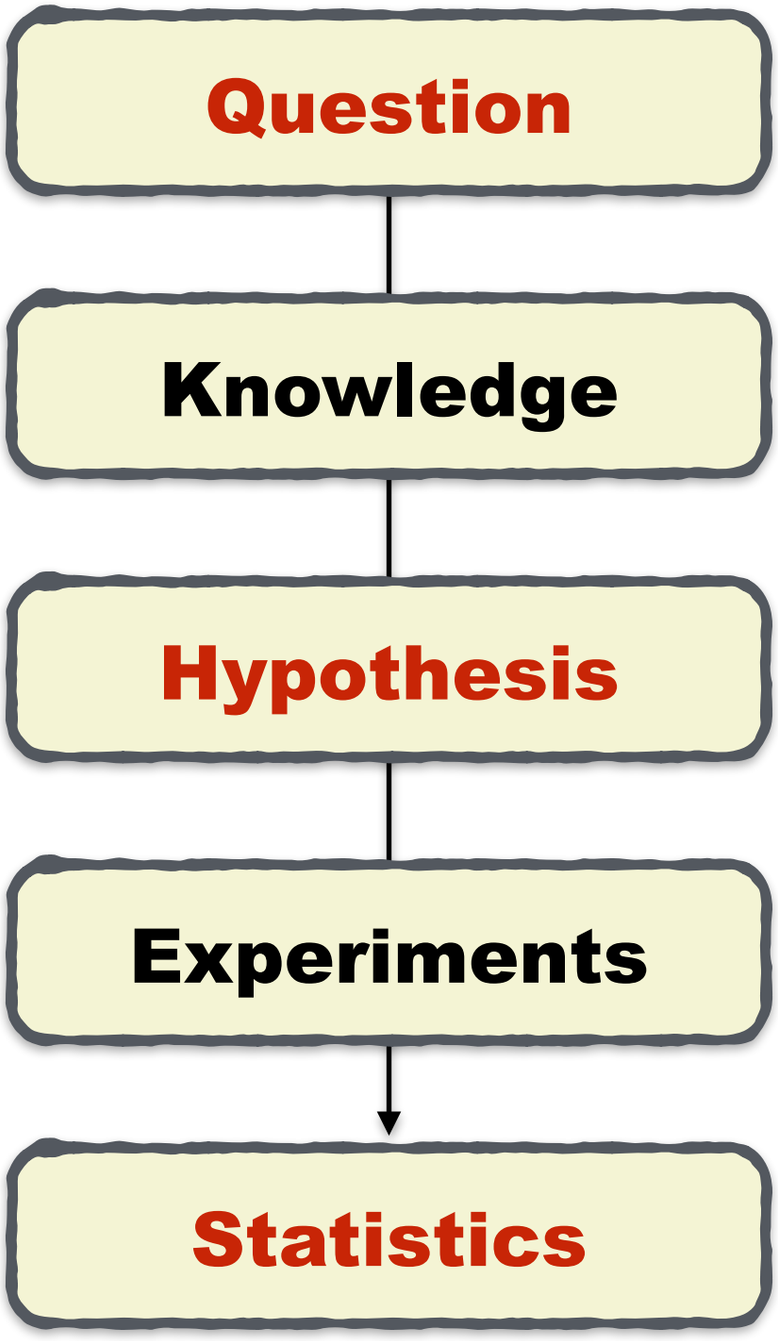
Changwei W. Wu, Ph.D.

Graduate Institute of Mind, Brain and Consciousness
Graduate Institute of Humanities in Medicine
Taipei Medical University

Process of Scientific Research (I)

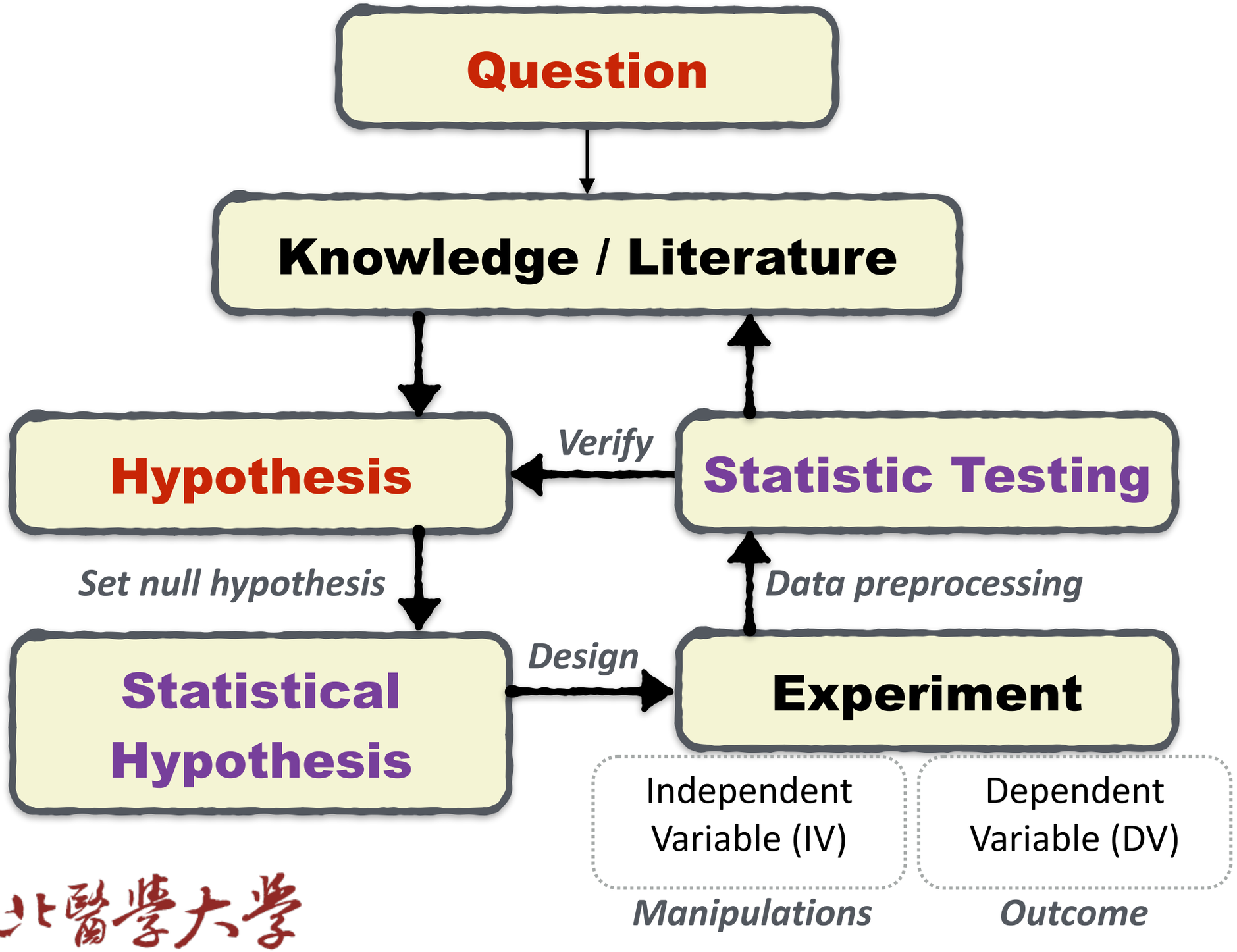


Process of Scientific Research (II)



TAIPEI MEDICAL UNIVERSITY
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

Process of Scientific Research (III)

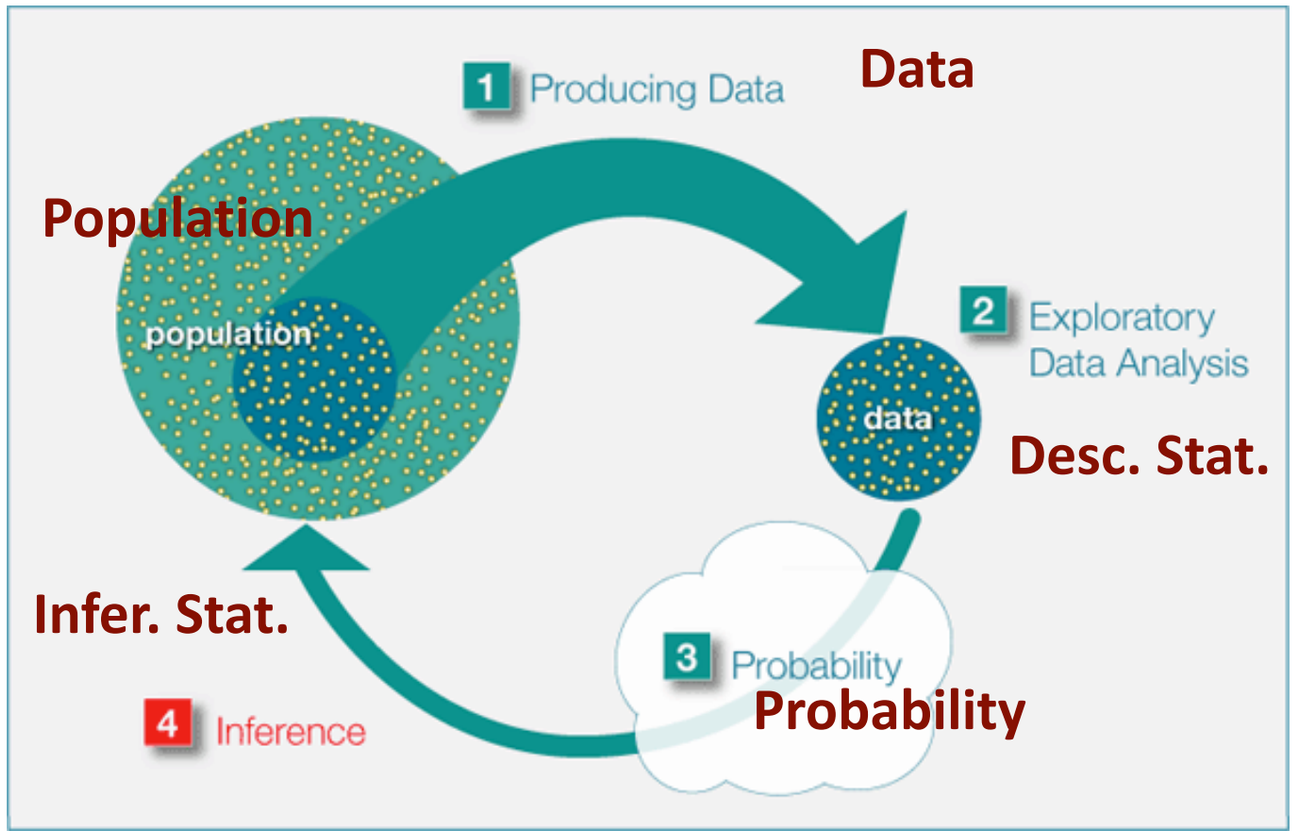


Statistics

- **Descriptive statistics*:***
  - Used for describing basic features of the collected data.

- **Inferential statistics*:***
  - Used for inferring the population (unknown) from the collected data (known).
  - (Why inference?) In reality, we do not know the truth (real distribution) of the entire population.

	day1	day2	day3
median	1.790	0.790	0.760
mean	1.771	0.961	0.977
SE.mean	0.024	0.044	0.064



https://bolt.mph.ufl.edu/6050-6052/unit-4/

臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

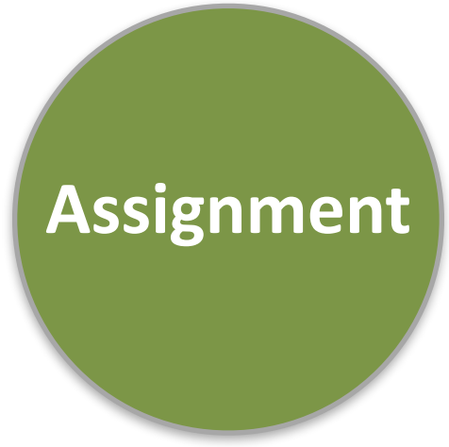
Statistics

1. Probability distributions

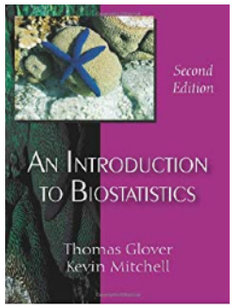
- Descriptive & Inferential statistics
- Sample space & Random variables
- Probability distribution: PDF & CDF

2. [R] Data preparation

- Data Input/Output
- Data restructure
- Plot probability distributions



TAIPEI MEDICAL UNIVERSITY
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY



Samples

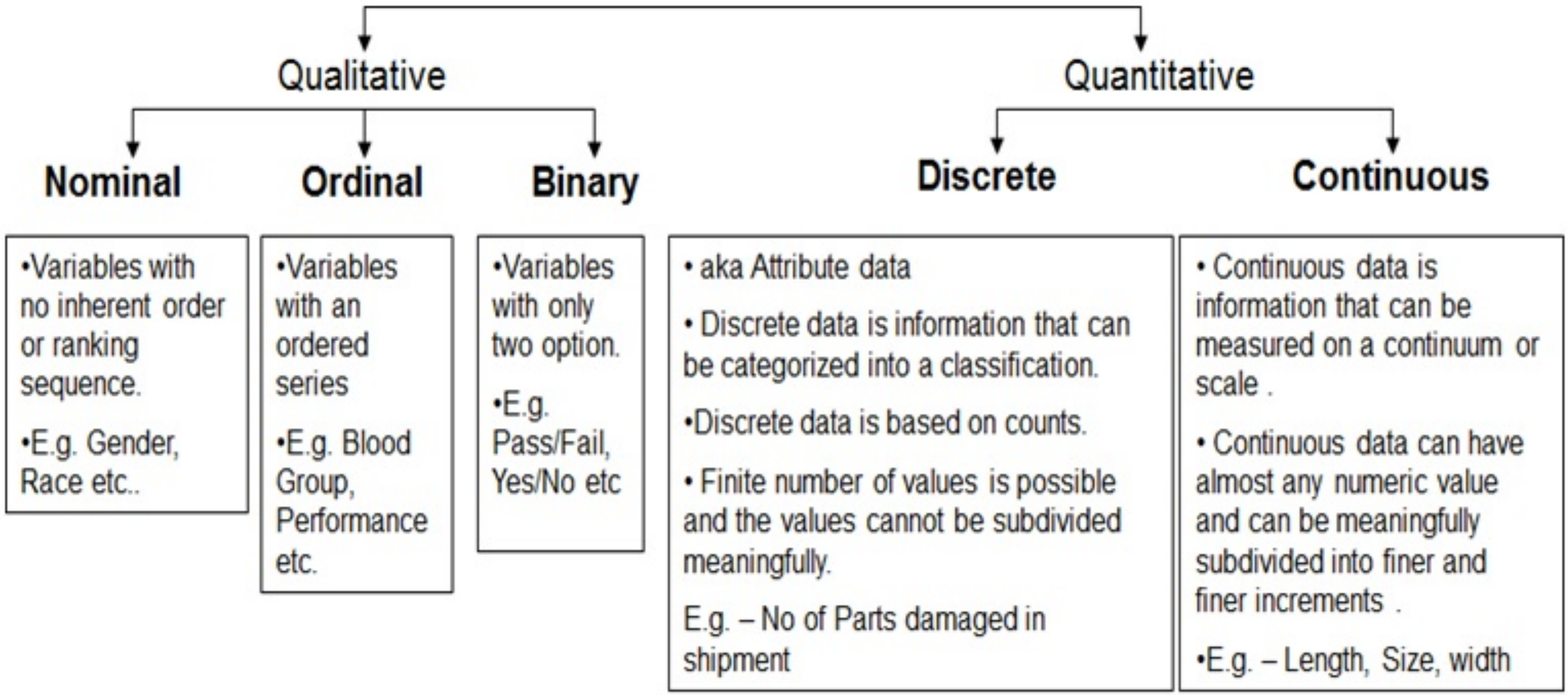
- Under specific research topics, collect the data from a portion of samples within the general population through observations/ measurements.
- **Random** samples will be representative <u>without bias</u>.
- **Sample space** (S): The set consisting of all possible outcomes in an experiment.
  - $S_1 = \{\text{heads, tails}\}$;
  - $S_2 = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$;
  - $S_3 = \{\texttt{A}\heartsuit, 2\heartsuit, \ldots, \texttt{K}\heartsuit, \texttt{A}\spadesuit, 2\spadesuit, \ldots, \texttt{K}\spadesuit, \texttt{A}\diamondsuit, 2\diamondsuit, \ldots, \texttt{K}\diamondsuit, \texttt{A}\clubsuit, 2\clubsuit, \ldots, \texttt{K}\clubsuit\}$;
  - $S_4 = \{x \geq 0\}$.

Types of Variables

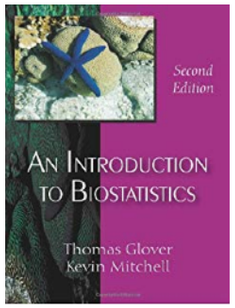
- **Numeric variables**
  - Equal-interval variables (interval scales)
    - Continuous variable: e.g., GPA; height
    - Discrete counts: e.g., the number of time visiting dentist
    - Proportions: e.g., percentage, rate
  - Rank-order variables (ordinal/discrete scales)
    - e.g., order of finishing a race; birth order of children
    - Physical activity level (low, moderate and high)
- **Nominal variables**
  - Gender (male, female)
  - Ethnicity (Caucasian, African American, Asian and Hispanic)
  - Profession (surgeon, doctor, nurse, dentist)

Variables

- **Levels of Measurement (& Types of Variables):**



https://stats.stackexchange.com/questions/159902/is-nominal-ordinal-binary-for-quantitative-data-qualitative-data-or-both



Probability

- $S_1 = \{\text{heads, tails}\}$;
- $S_2 = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$;
- $S_3 = \{\texttt{A}\heartsuit, 2\heartsuit, \ldots, \texttt{K}\heartsuit, \texttt{A}\spadesuit, 2\spadesuit, \ldots, \texttt{K}\spadesuit, \texttt{A}\diamondsuit, 2\diamondsuit, \ldots, \texttt{K}\diamondsuit, \texttt{A}\clubsuit, 2\clubsuit, \ldots, \texttt{K}\clubsuit\}$;
- $S_4 = \{x \geq 0\}$.

- i.e., to measure the **Uncertainty** from experiments

FORMULA 2.1. The empirical probability of an event A is defined as

$$P(A) = \frac{n_A}{n} = \frac{\text{number of times } A \text{ occurred}}{\text{number of trials run}}.$$

臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

Probability Density Function

DEFINITION 3.1. Let X be a discrete random variable. The **probability density function** or **probability distribution** f for X is

$$f(x) = P(X = x),$$

where x is any real number. (In words, this says that the value of $f(x)$ is the probability that the random variable X is equal to the value x.) Note that

- f is defined for all real numbers;
- $f(x) \geq 0$ since it is a probability;
- $f(x) = 0$ for most real numbers because X is discrete and cannot assume most real values;
- summing f over all possible values of X produces 1, i.e.,

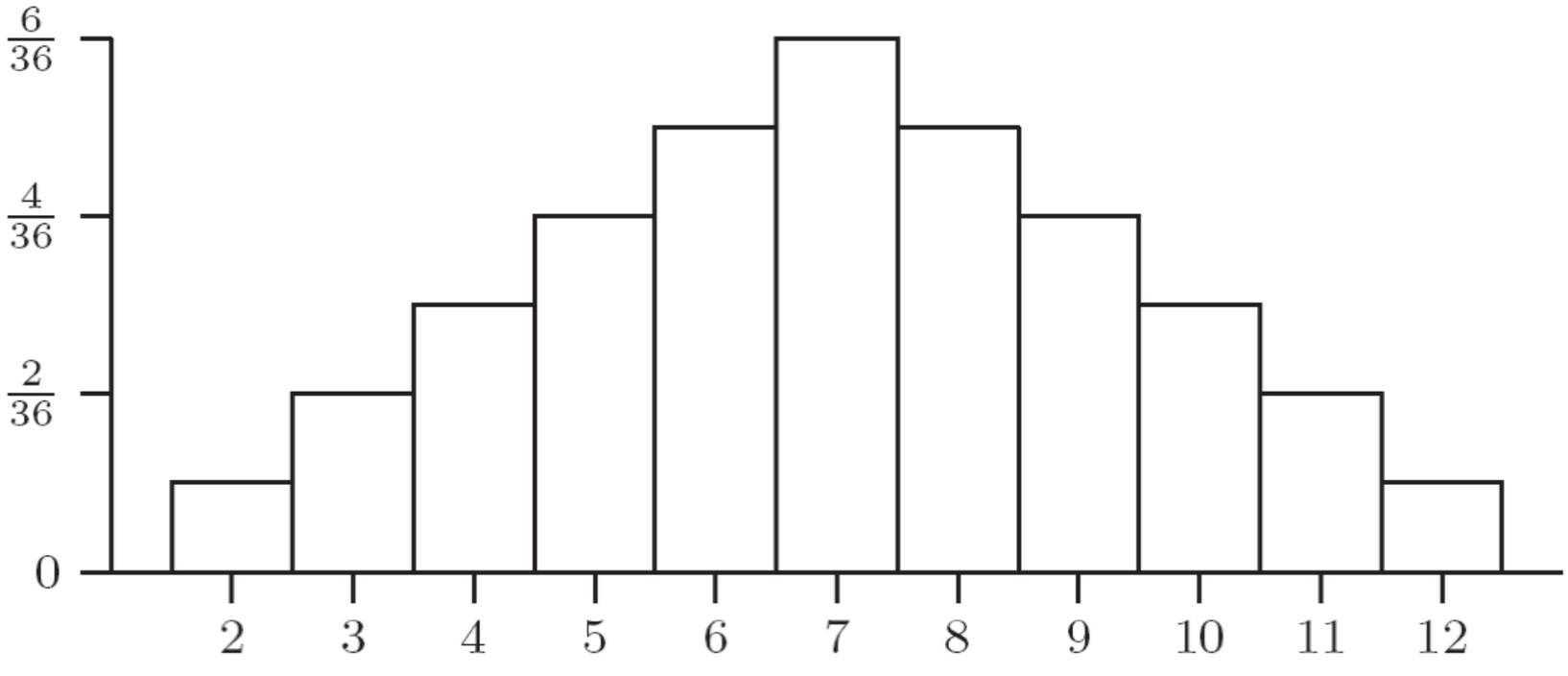
$$\sum_{\text{all } x} f(x) = 1.$$

A probability density function is also referred to as a **pdf**.

Probability Density Function

EXAMPLE 3.2. A fair 6-sided die is rolled twice with the discrete random variable X representing the sum of the numbers obtained on both rolls. Give the density function of this variable.

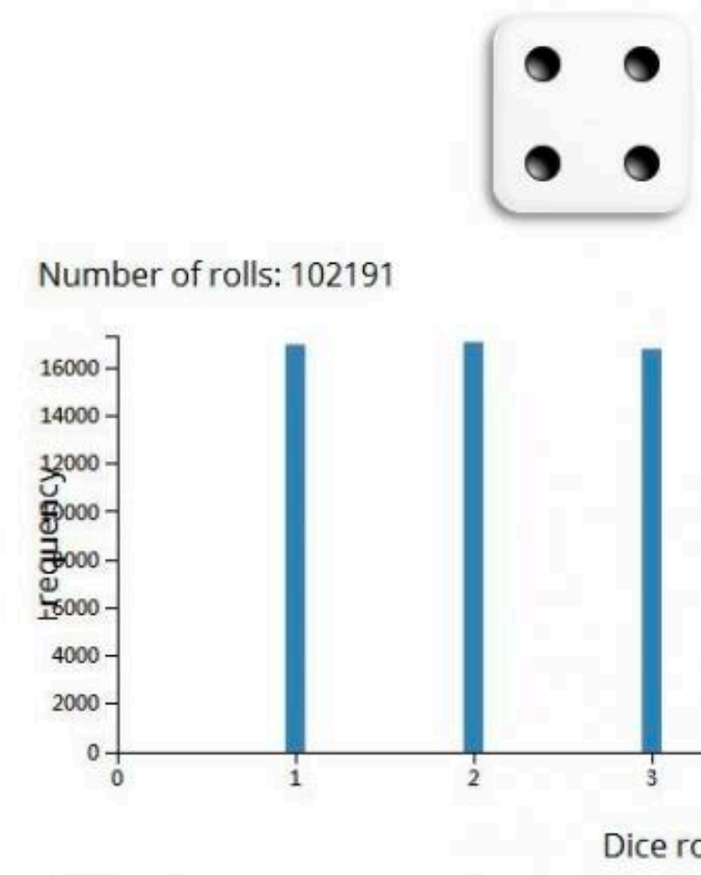
Random variable $X = x$	2	3	4	5	6	7	8	9	10	11	12
Density $f(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$



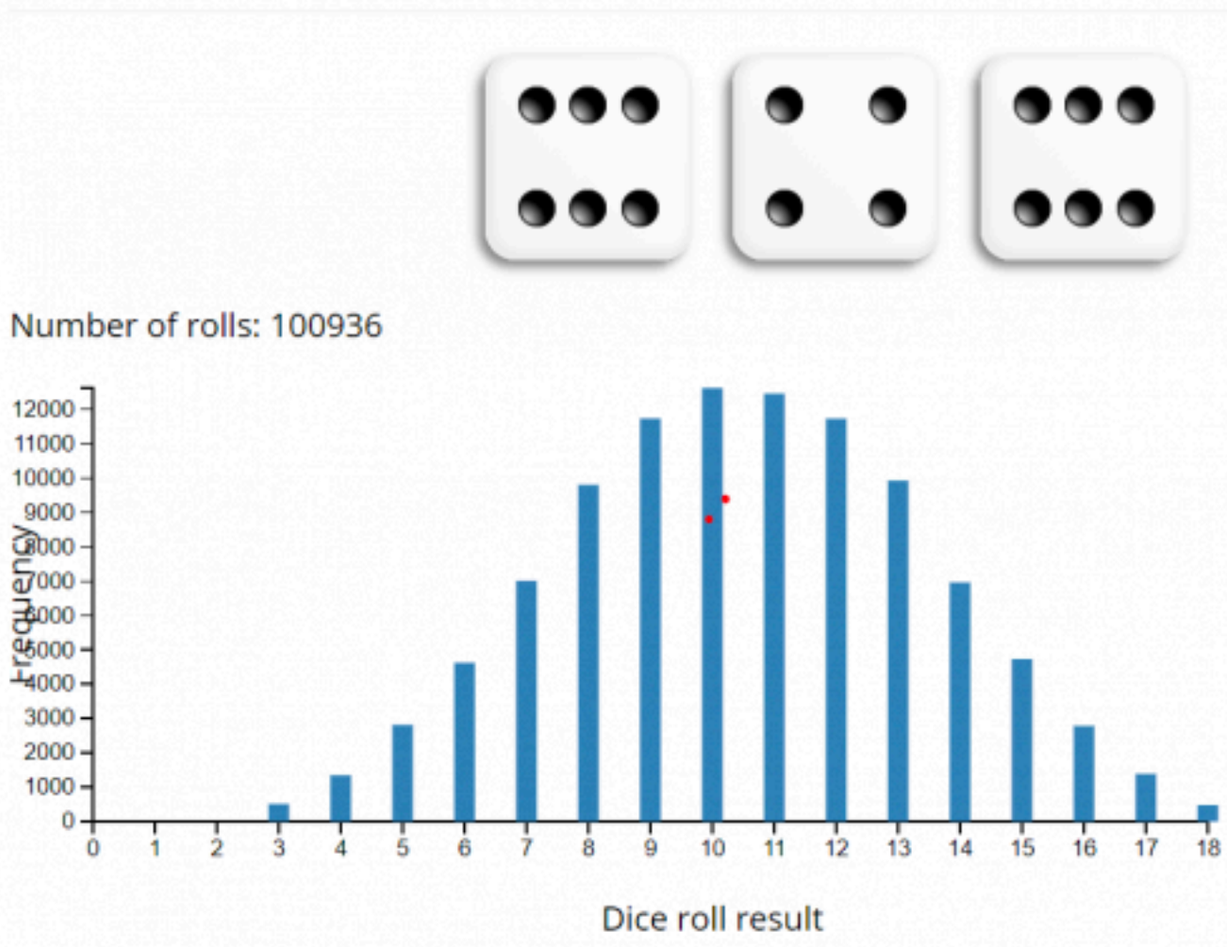
TAIPEI MEDICAL UNIVERSITY 臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

Probability Density Function

Rolling one die



Rolling 3 dice



https://academo.org/demos/dice-roll-statistics/

Cumulative Distribution Function

DEFINITION 3.5. Let X be a discrete random variable with density f. The **cumulative distribution function (CDF)** for X is denoted by F and is defined by

$$F(x) = P(X \leq x),$$

for all real x.

Random variable $X = x$	2	3	4	5	6	7	8	9	10	11	12
Density $f(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
CDF: $F(x)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	$\frac{36}{36}$

What is the probability of an 8 or less on a roll of a pair of dice?

SOLUTION. $P(X \leq 8) = F(8) = \frac{26}{36}$.

Binomial Distribution

DEFINITION 3.6. The probability density function of a **binomial random variable** with n trials and probability p of success on any trial is

$$f(x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}.$$

The binomial random variable probability density function is characterized by the two parameters n, the number of trials or sample size, and p, the probability of success on a trial.

1. A fixed number n of trials are carried out.

2. The outcome of each trial can be classified in precisely one of two mutually exclusive ways termed "success" and "failure." The term "binomial" literally means two names.

3. The probability of a success, denoted by p, remains constant from trial to trial. The probability of a failure is $1 - p$.

4. The trials are independent; that is, the outcome of any particular trial is not affected by the outcome of any other trial.

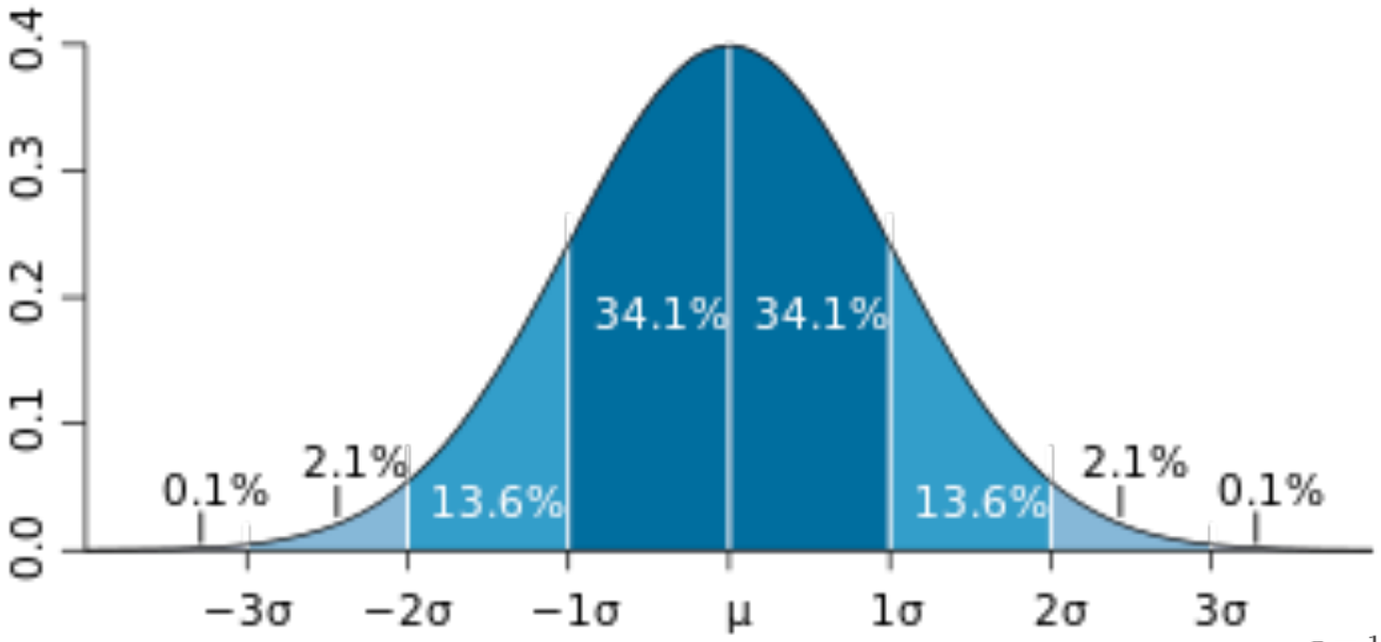
THEOREM 3.2. Let X be a binomial random variable with parameters n and p. Then $\mu = E(X) = np$ and $\sigma_X^2 = \text{Var}(X) = np(1-p)$.

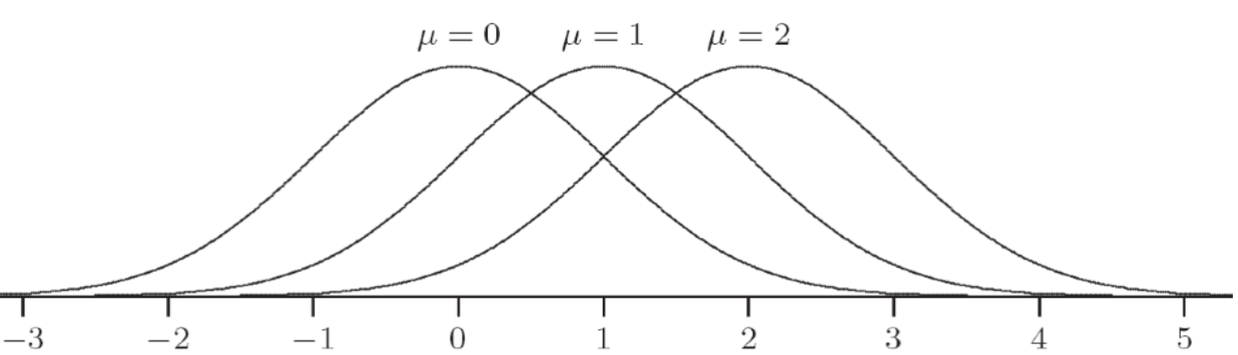
Normal Distribution

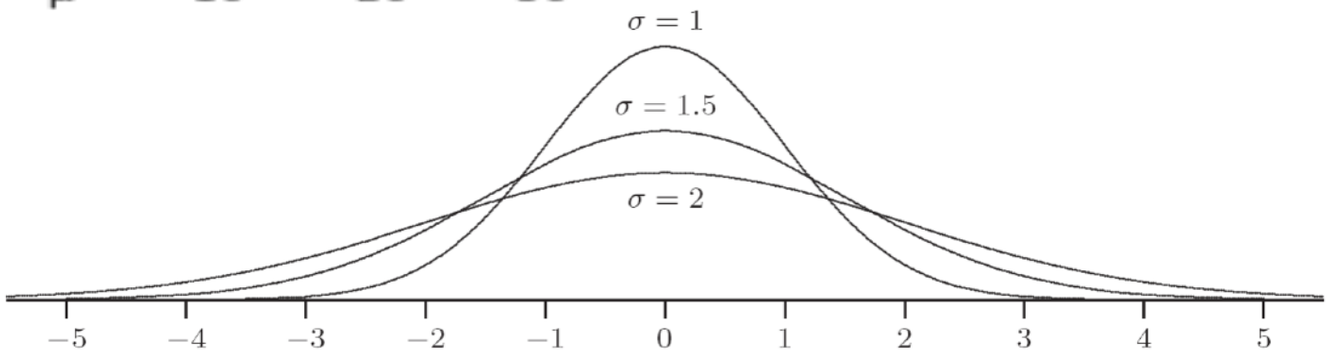
DEFINITION 3.9. The **probability density function for a normal random variable** has the form

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2},$$

where σ is the standard deviation of the random variable and μ is its mean.







Standard Normal Distribution

3 Probability

DEFINITION 3.10. Let X be a normal random variable with mean μ, and standard deviation σ. The transformation

$$Z = \frac{X - \mu}{\sigma}$$

expresses X as the **standard normal random variable** Z with $\mu = 0$ and $\sigma = 1$.

EXAMPLE 3.16. Suppose that the scores on an aptitude test are normally distributed with a mean of 100 and a standard deviation of 10. (Some of the original IQ tests were purported to have these parameters.) What is the probability that a randomly selected score is below 90?

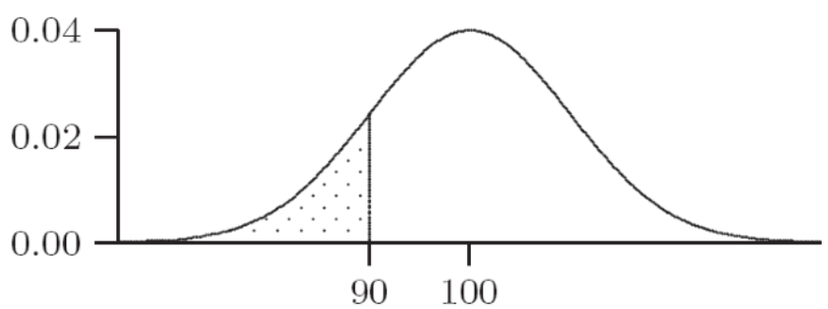
Transform X into a standard normal variable. $\mu = 100$ and $\sigma = 10$, so

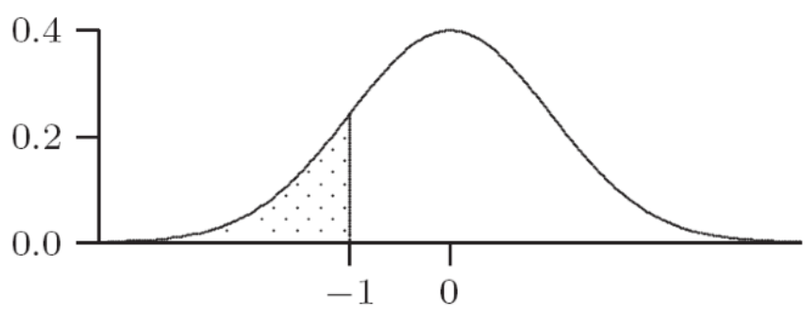
$$Z = \frac{X - \mu}{\sigma} = \frac{90 - 100}{10} = -1.0.$$

Thus a score of 90 can be represented as 1 standard deviation below the mean,

$$P(X < 90) = P(Z < -1.0).$$

probability of a score less than 90 is 0.1587.





Practice

Standard Normal Distribution

EXAMPLE 4.4. The mean blood cholesterol concentration of a large population of adult males (50–60 years old) is 200 mg/dl with a standard deviation of 20 mg/dl. Assume that blood cholesterol measurements are normally distributed. What is the probability that a randomly selected individual from this age group will have a blood cholesterol level below 250 mg/dl?

SOLUTION. Apply the standard normal transformation

$$P(X < 250) = P\left(Z < \frac{250 - 200}{20}\right) = P(Z < 2.5) = F(2.5).$$

Try using R to calculate the probability…

What is the probability that a randomly selected individual from this age group will have a blood cholesterol level above 225 mg/dl?

Random Variable

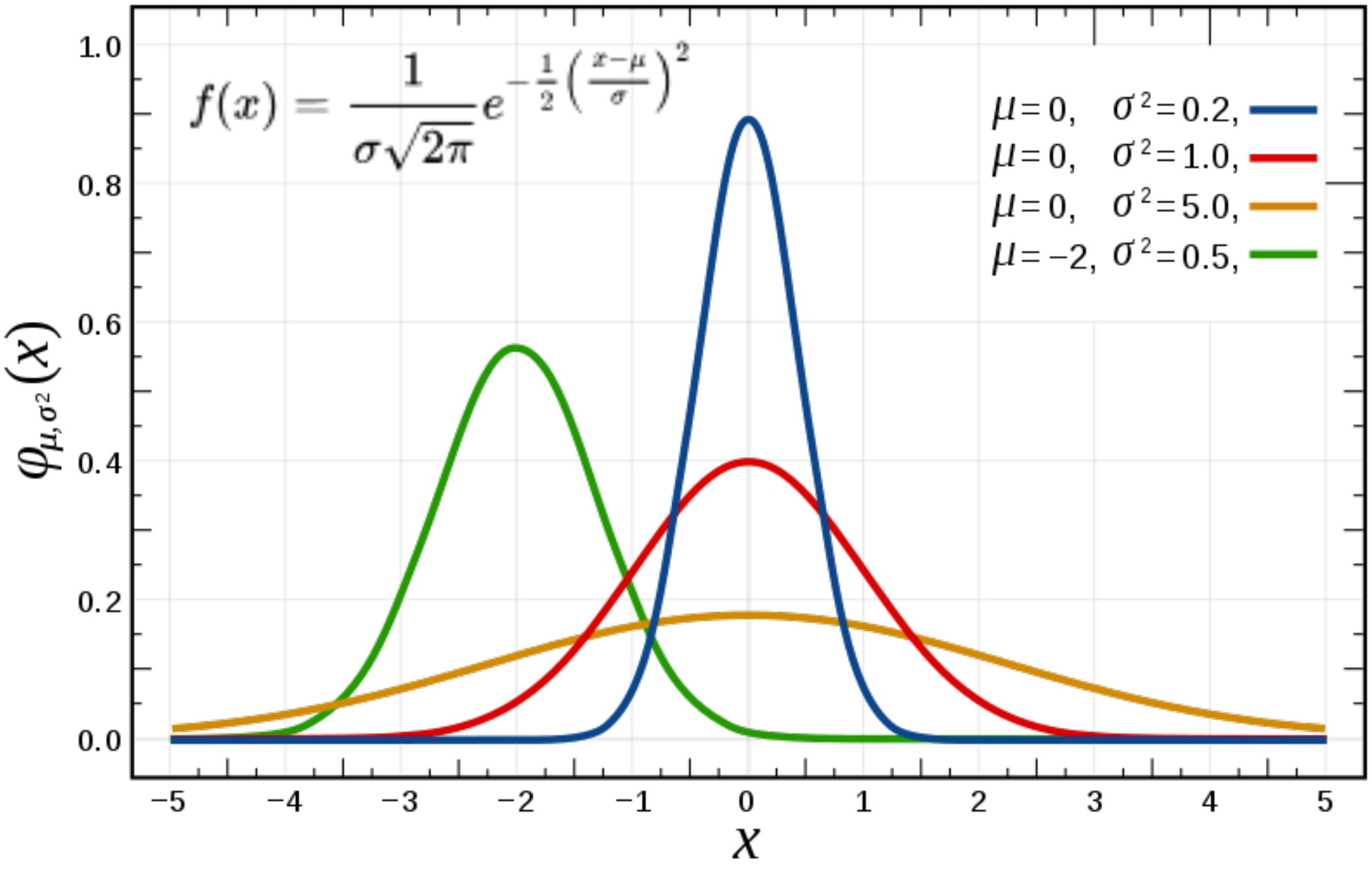
- A **random variable** is a variable whose actual value is determined by chance.

- **Probability distributions of random variables**
  - **Probability density functions (PDF)**
    - Expected values
    - Mean
    - Sum of Squares (SS)
    - Variance
    - Degree of freedom
  - **Cumulative distribution functions (CDF)**

$$\textbf{sum of squares } \sum (y - \bar{y})^2$$

$$\text{variance} = \frac{\text{sum of squares}}{\text{degrees of freedom}}$$

PDF of Normal Distribution

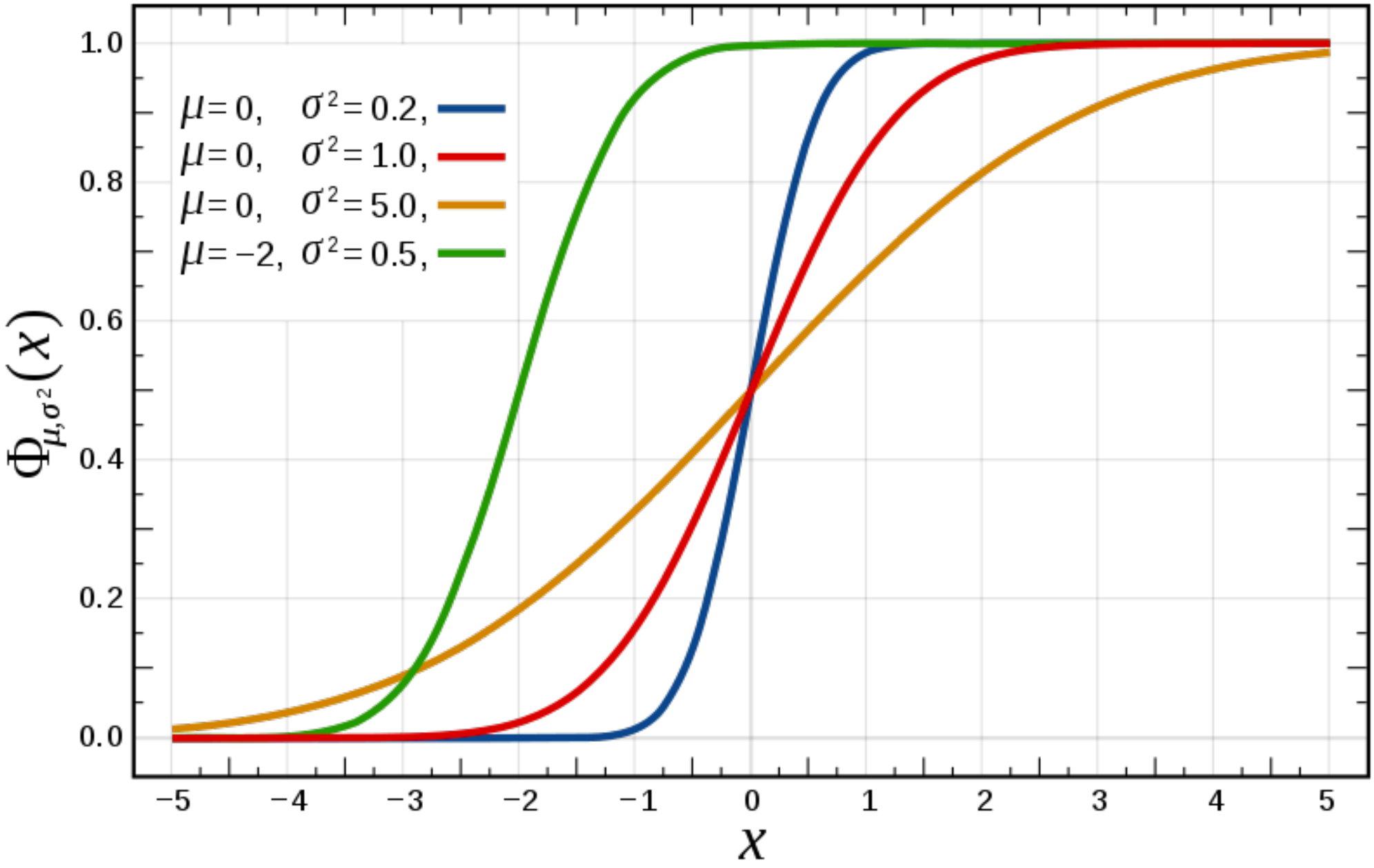


臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

https://en.wikipedia.org/wiki/Normal_distribution

CDF of Normal Distribution

CDF is what we use most of the time.



https://en.wikipedia.org/wiki/Normal_distribution

Descriptive Statistics

DEFINITION 3.2. The long-term **expected value** or **mean** for a discrete random variable X with density function f is given by

$$\mu = E(X) = \sum_{\text{all } x} x f(x).$$

In other words, in the sum each value x is weighted by its density or probability.

$$\mu = E(X) = \sum_{x=2}^{12} x f(x) = 2 \cdot \tfrac{1}{36} + 3 \cdot \tfrac{2}{36} + 4 \cdot \tfrac{3}{36} + \cdots + 11 \cdot \tfrac{2}{36} + 12 \cdot \tfrac{1}{36} = 7.$$

DEFINITION 3.4. Let X be any discrete random variable. Then the **variance** of X, denoted by σ^2, σ_X^2, or $\text{Var}(X)$, is defined by

$$\sigma^2 = E\left[(X-\mu)^2\right].$$

$$\sigma^2 = E(X^2) - [E(X)]^2 = 54.833 - (7)^2 = 5.833.$$

Descriptive Statistics

	Population	Samples
Means	$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$ where N is the population size.	$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ where n is the sample size. The sample mean is usually reported to one more decimal place than the data and always has appropriate units associated with it.
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}.$	$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}.$

Key Functions

1. **Data Input/Output**
   - read csv files (*read.delim* / *read.csv*)
   - read Excel files (*read_excel*)
   - save R data (*write.table* / *write.csv*)

2. **Data restructure**
   - combine column/rows (*cbind* / *rbind*)
   - restructure between long/wide format (*melt* / *dcast*)

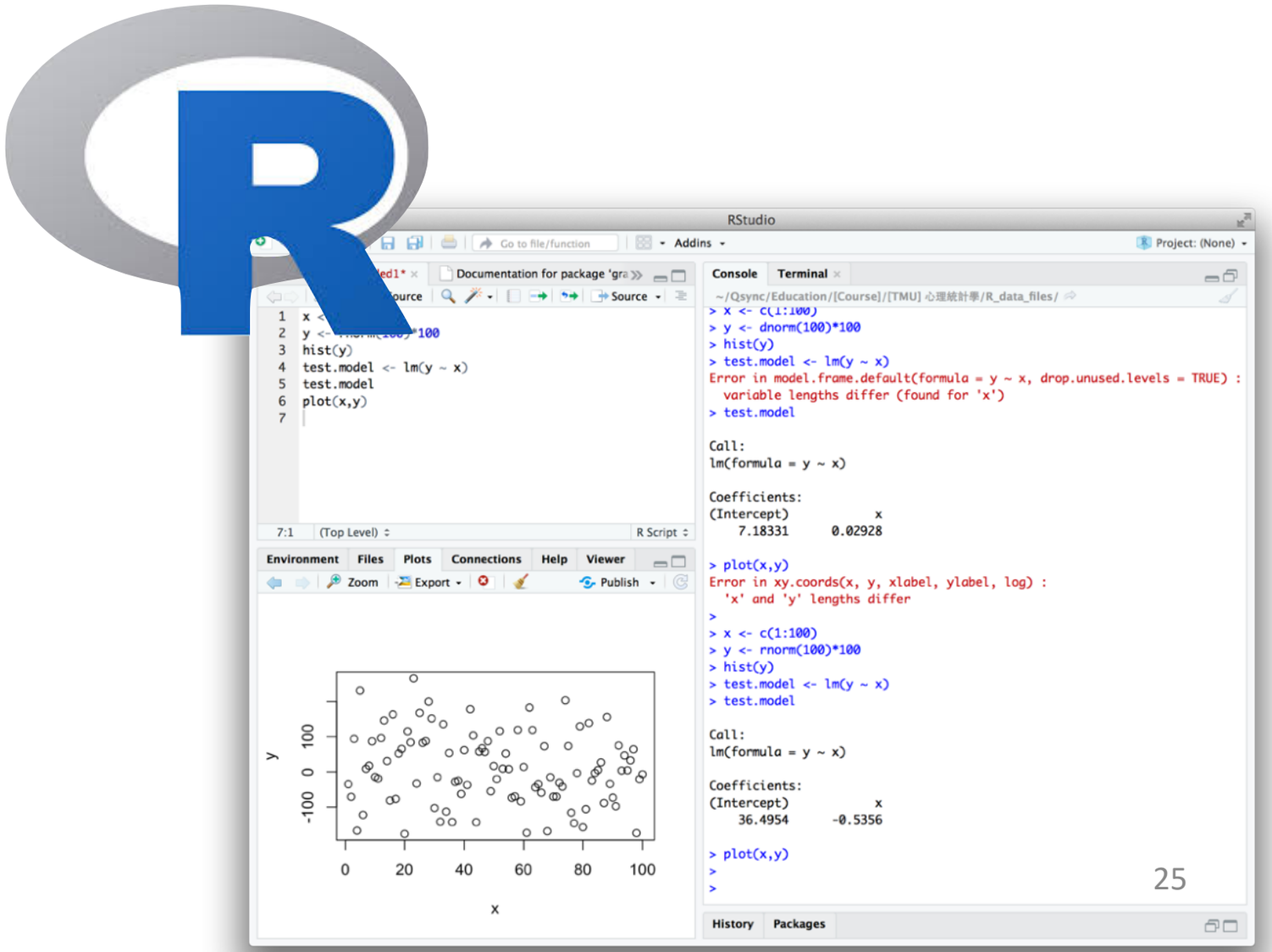
3. **Plot distributions**
   - Histogram (*hist*) / Boxplot (*boxplot*) / Scatter plot (*plot*)
   - check descriptive statistics (*stat.desc*)

TAIPEI MEDICAL UNIVERSITY 臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

Let's check if you have all these functions!

- Data Input/Output
- Data restructure
- Plot probability distributions

HANDS-ON PRACTICE OF R



TAIPEI MEDICAL UNIVERSITY
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

A. Input / Output

General form:

- *dataframe <-* ***read.table(file****, header=FALSE,* ***sep = " ")***
- *dataframe <-* ***read.csv(file****, header=FALSE)*

General form:

- ***write.table****(dataframe, "filename.txt", sep = " ")*
- ***write.csv****(dataframe, "filename.txt")*

- **Parameters**:
  - header=(TRUE / FALSE): whether the file contains var. names
  - sep: field separator character.
  - dec: character of decimal points.

	sep	dec
read.table	<empty>	.
read.csv	,	.
read.csv2	;	,
read.delim	\t	.
read.delim2	\t	.

TAIPEI MEDICAL UNIVERSITY 臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

Demo

A. Input / Output

- Import data from external files
  - *test<-read.csv("Ex1.csv",header=T)*
  - *dim(test)*
  - *colnames(test)*
  - *mean(test$latitude)*
  - *var(test$latitude)*
  - *dim(test[test$depth>40 & test$latitude>0,])[1]/dim(test)[1]*
- Export data to files
  - *write.table(test, "TEST.txt", sep="\t", row.names = FALSE)*
  - *write.csv(test, "TEST.csv")*

Load: **Ex1.csv**

Demo

A. Input Excel file

- Import data from external files
  - Library: **readxl**, can work on both .xls and .xlsx
  - *library(readxl)*
  - *excel_sheets('ExcelExample.xlsx')*
  - *tomatoXL <- read_excel('ExcelExample.xlsx')*
  - *wineXL1 <- read_excel('ExcelExample.xlsx', sheet=2)*

Load:
ExcelExample.xlsx

B. Restructuring Data

'long' data structure

ID	Y	time	X_4
1	3.5	1	1
1	3.7	2	1
1	3.9	3	1
1	3.0	4	1
1	3.2	5	1
1	3.2	6	1
2	4.1	1	1
2	4.1	2	1
.			
.			
N	5.0	5	2
N	4.7	6	2

'broad' data structure

ID	Y_{t1}	Y_{t2}	Y_{t3}	Y_{t4}	Y_{t5}	Y_{t6}	X_4
1	3.5	3.7	3.9	3.0	3.2	3.2	1
2	4.1	4.1	4.2	4.6	3.9	3.9	1
3	3.8	3.5	3.5	3.4	2.9	2.9	2
4	3.8	3.9	3.8	3.8	3.7	3.7	1
.							
.							
N	4.0	4.6	4.7	4.3	4.7	5.0	2

Usually, we use "long" format for statistical analysis.

B. Restructuring Data

General form:

- ***cbind(vector.1, vector.2)***
- ***rbind(dataframe.1, dataframe.2)***
- ***melt(data, id.vars**=c("X", "Y"), **variable.name**="Z", **value.name**="V")*** - become long format
- ***dcast(data, X + Y ~ Z, value.var="V")*** - become wide format

- **Parameters**:
  - cbind: bind the columns (vectors)
  - rbind: bind the data.frame/matrix by rows (same column#)
  - id.vars: specify variables that defines group(s) without values.
  - variable.name: specify name of variables in different columns.
  - value.name(.var): specify the name for stored values.
  - left side of "~": variables to be maintained; (right) to be distributed.

DEMO

B. Restructuring Data

Load: **US_Foreign_Aid_00s.csv**

- Changing between long/wide form
  - *library(reshape2); library(stringr)*

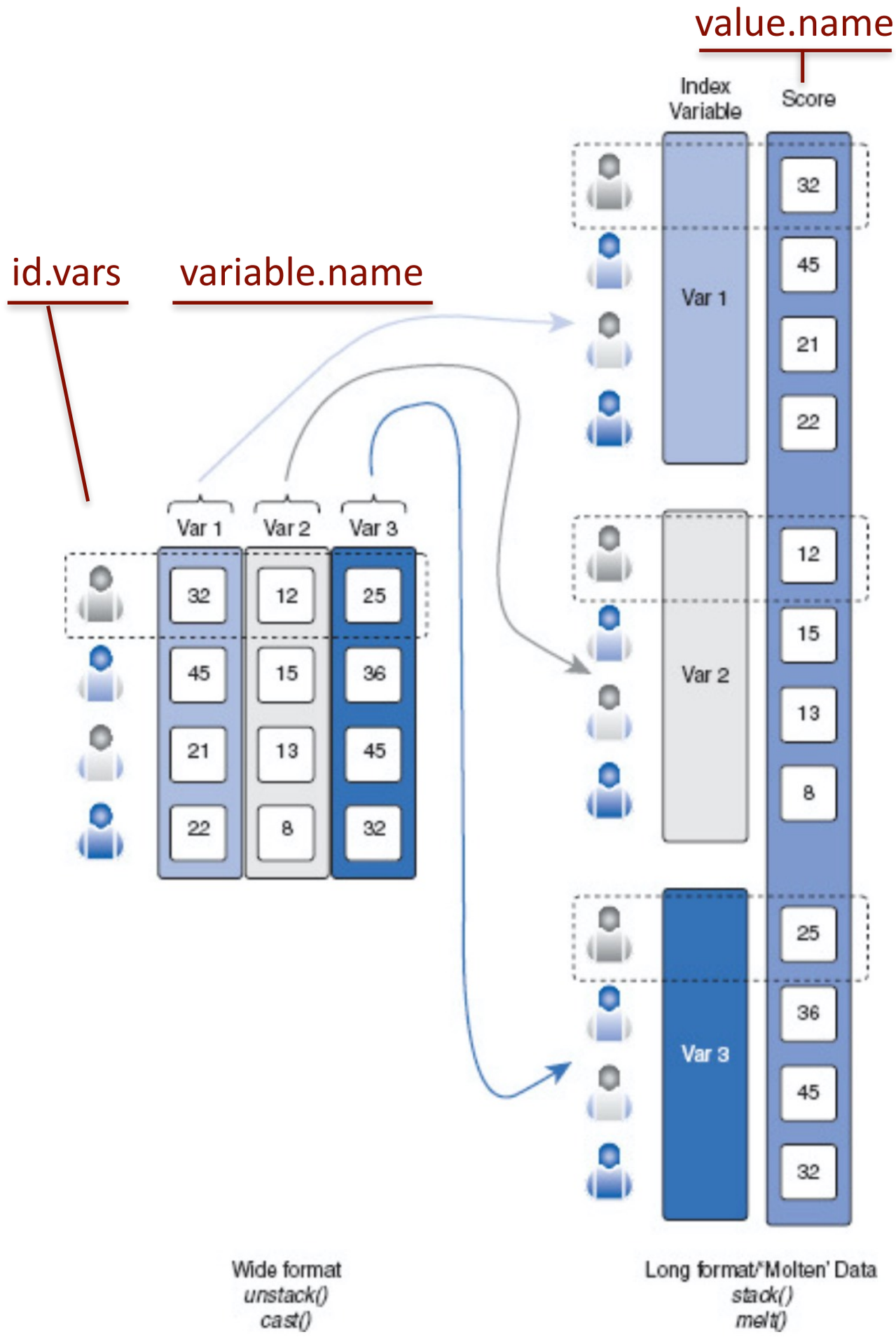
  - *Aid_00s <- read.csv(file)*
  - *melt00 <- melt(Aid_00s, id.vars=c("Country.Name", "Program.Name"), variable.name="Year", value.name="Dollars")*

MELT()

gather()

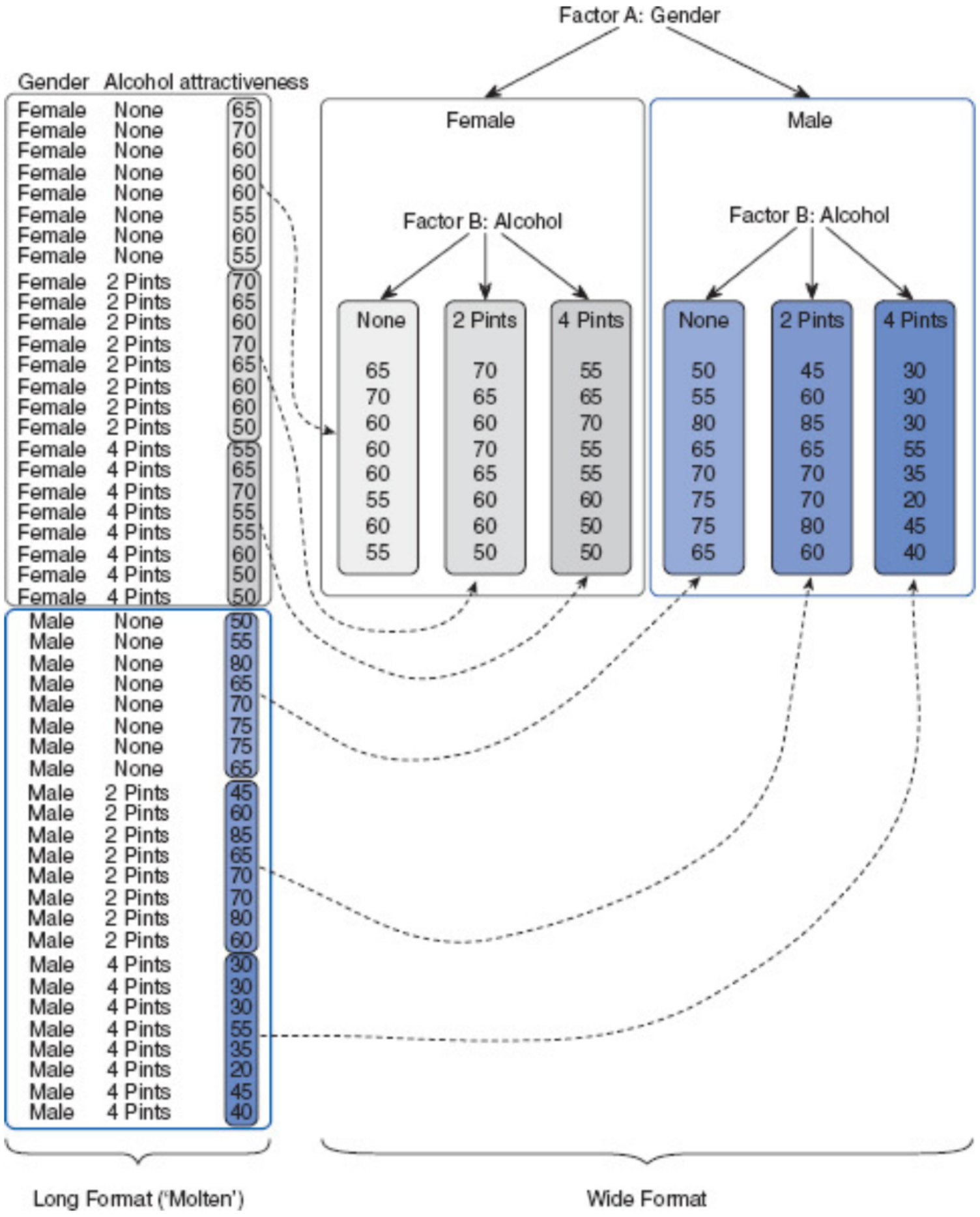
Transform from **Wide form** to **Long form**

- ***newdata <- melt(data,***
  id.vars*=c("Person"),*
  variable.name*="Index Variable",*
  value.name*="Score")*



DCAST()

spread()



DEMO

B. Restructuring Data

- Changing between long/wide form
  - *library(reshape2); library(stringr)*

Load: **US_Foreign_Aid_00s.csv**

  - *Aid_00s <- read.csv(file)*
  - *melt00 <- melt(Aid_00s, id.vars=c("Country.Name", "Program.Name"), variable.name="Year", value.name="Dollars")*
  - *melt00$Year <- as.numeric(str_sub(melt00$Year, start=3, 6))*
  - *cast00 <- dcast(melt00, Country.Name + Program.Name ~ Year, value.var = "Dollars")*

C. Built-in Graph Functions

- Load the built-in data in R
  - *data(diamonds)*

- Histogram -

- ***hist**(diamonds$carat, main = "Carat Histogram", xlab = "Carat")*

- Box plot -

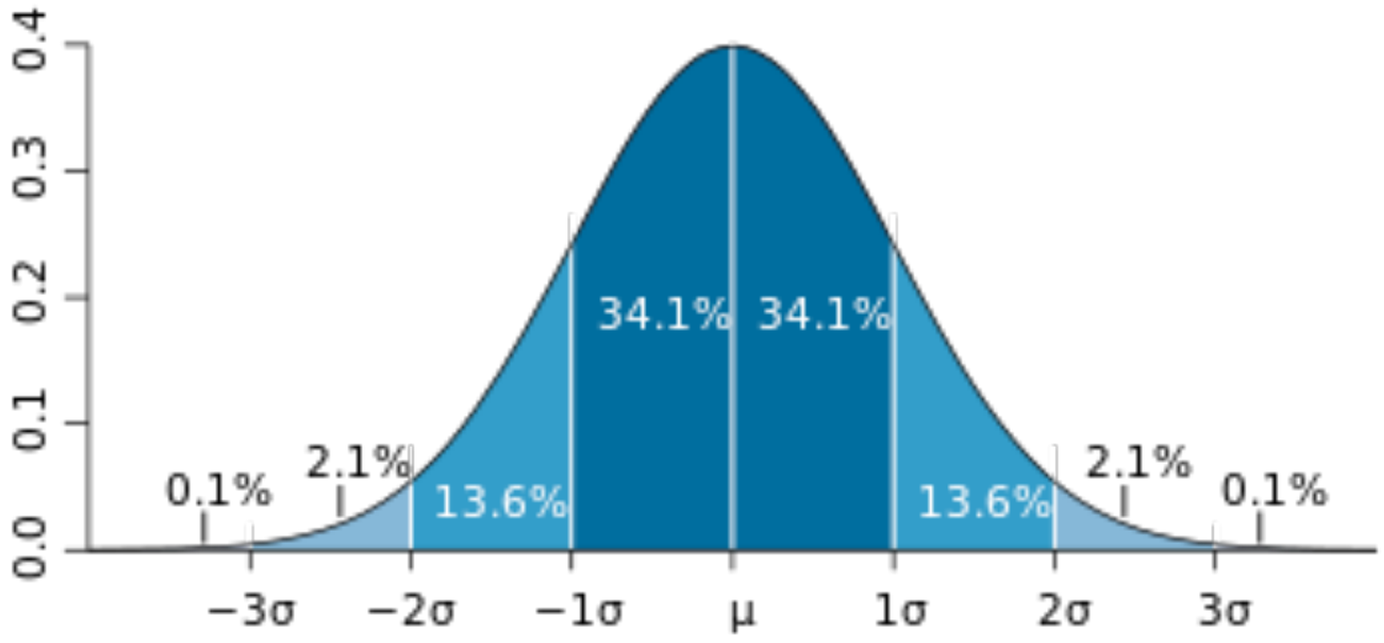
- ***boxplot**(diamonds$carat)*

- Scatter plot -

- ***plot**(price ~ carat, data = diamonds)*
- ***plot**(diamonds$carat, diamonds$price)*

R Functions for Probability Distributions

	Random variables	*Probability density function (PDF)*	*Cumulative distribution function (CDF)*	*Quantile*
Normal distribution	***rnorm***	***dnorm***	***pnorm***	***qnorm***
Binomial distribution	***rbinom***	***dbinom***	***pbinom***	***qbinom***

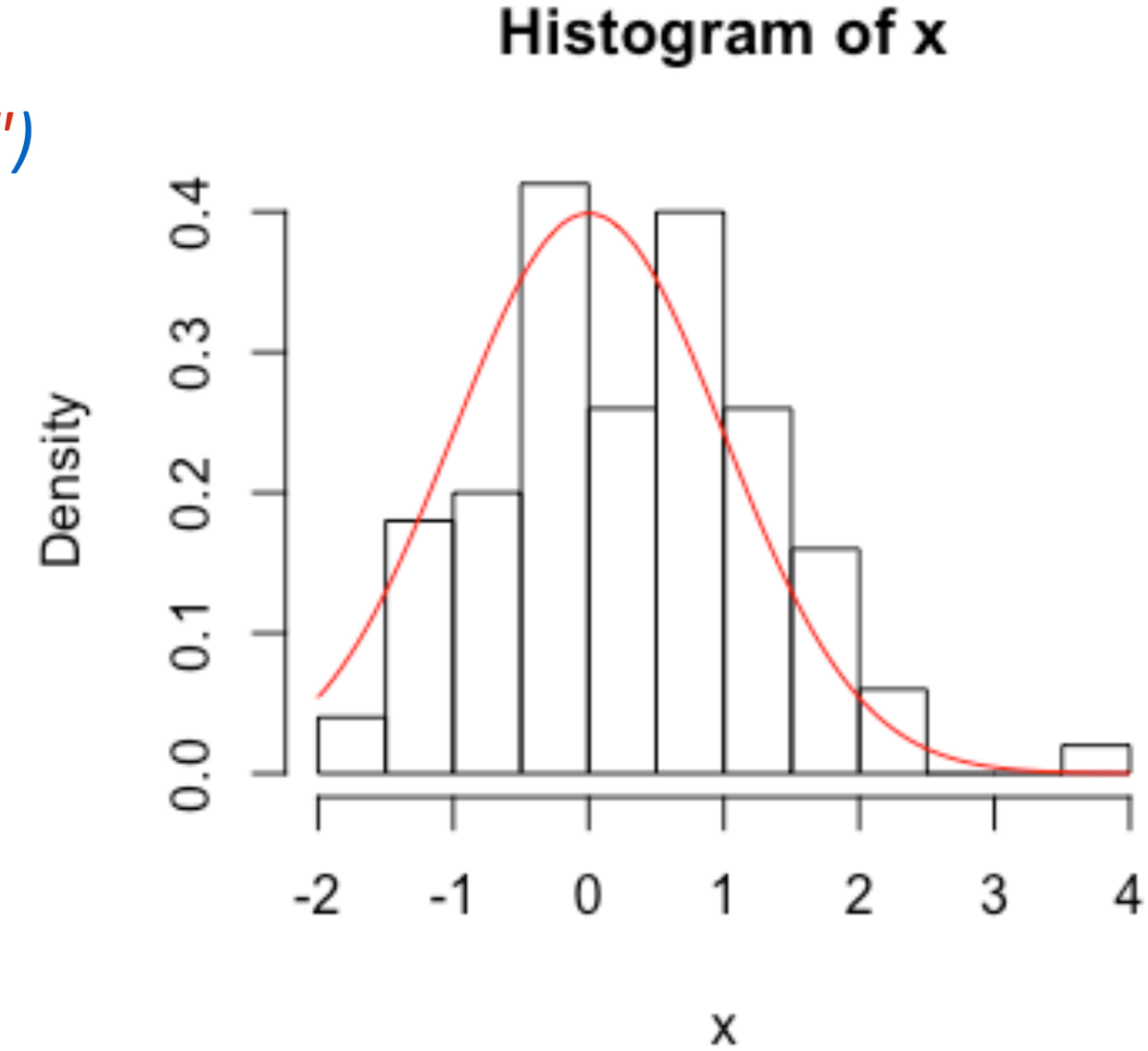


臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

DEMO

Plot Normal Distribution

- Function of “norm”
  - *x<-rnorm(100)*
  - *hist(x,freq=F)*
  - *curve(dnorm(x),add=T, col="red")*



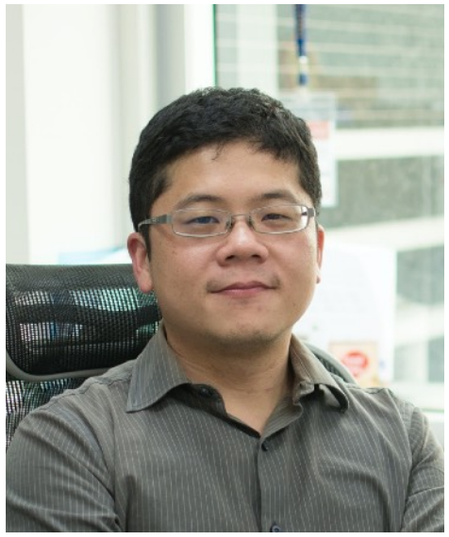
- Plot both PDF & CDF
  - *par(mfrow=c(1,2))*
  - *x<-seq(-4,4,0.1)*
  - *plot(x,dnorm(x),type="b")*
  - *plot(x,pnorm(x),type="b")*

臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

Discussion

1. **Probability and Distributions**
   - Binomial distribution
   - Normal distribution
   - Chi-square distribution

2. **[R] Data preparation**
   - Data Input/Output
   - Data restructure
   - Plot probability distributions



THANK YOU FOR YOUR ATTENTION

E-mail: sleepbrain@tmu.edu.tw