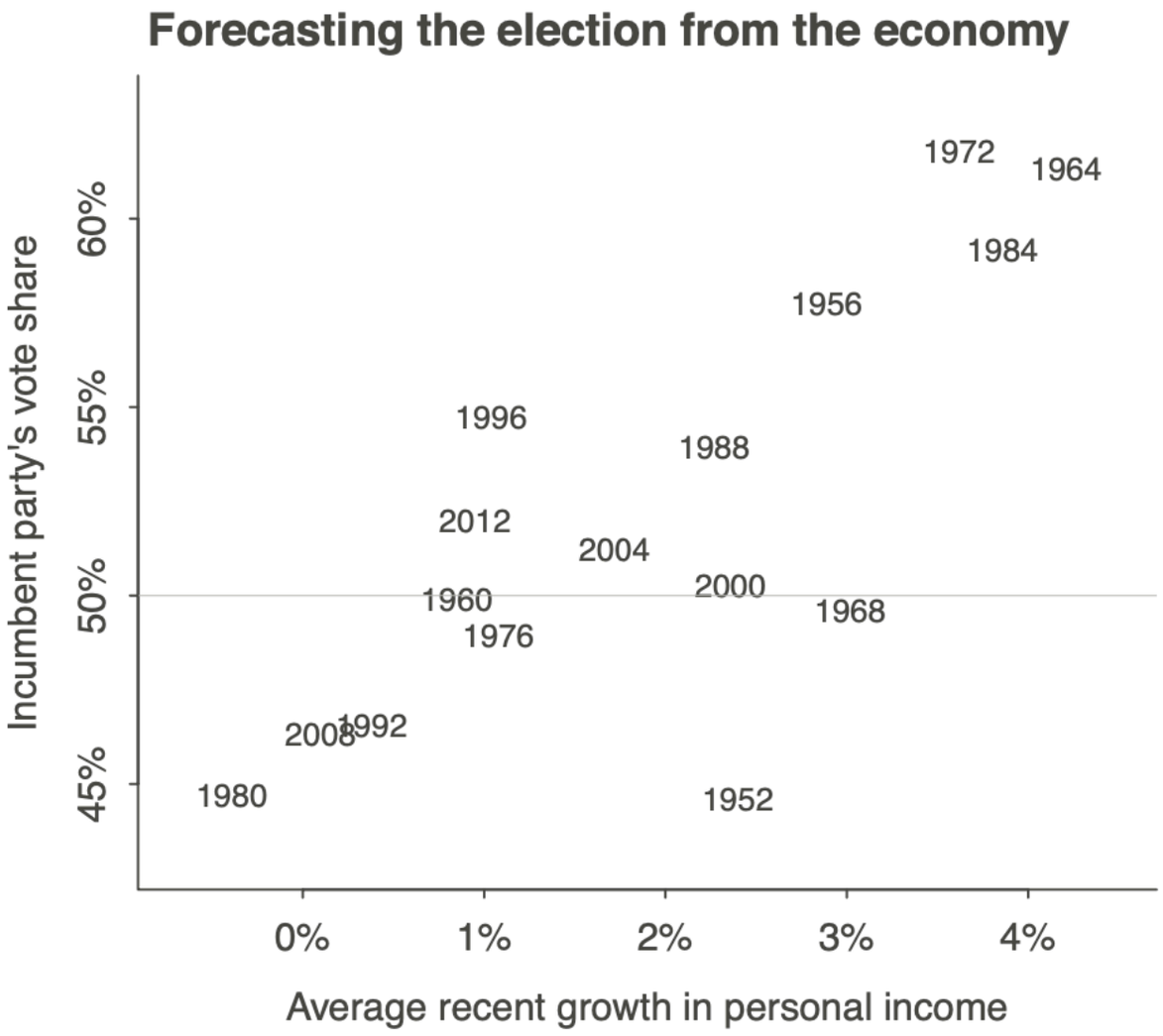
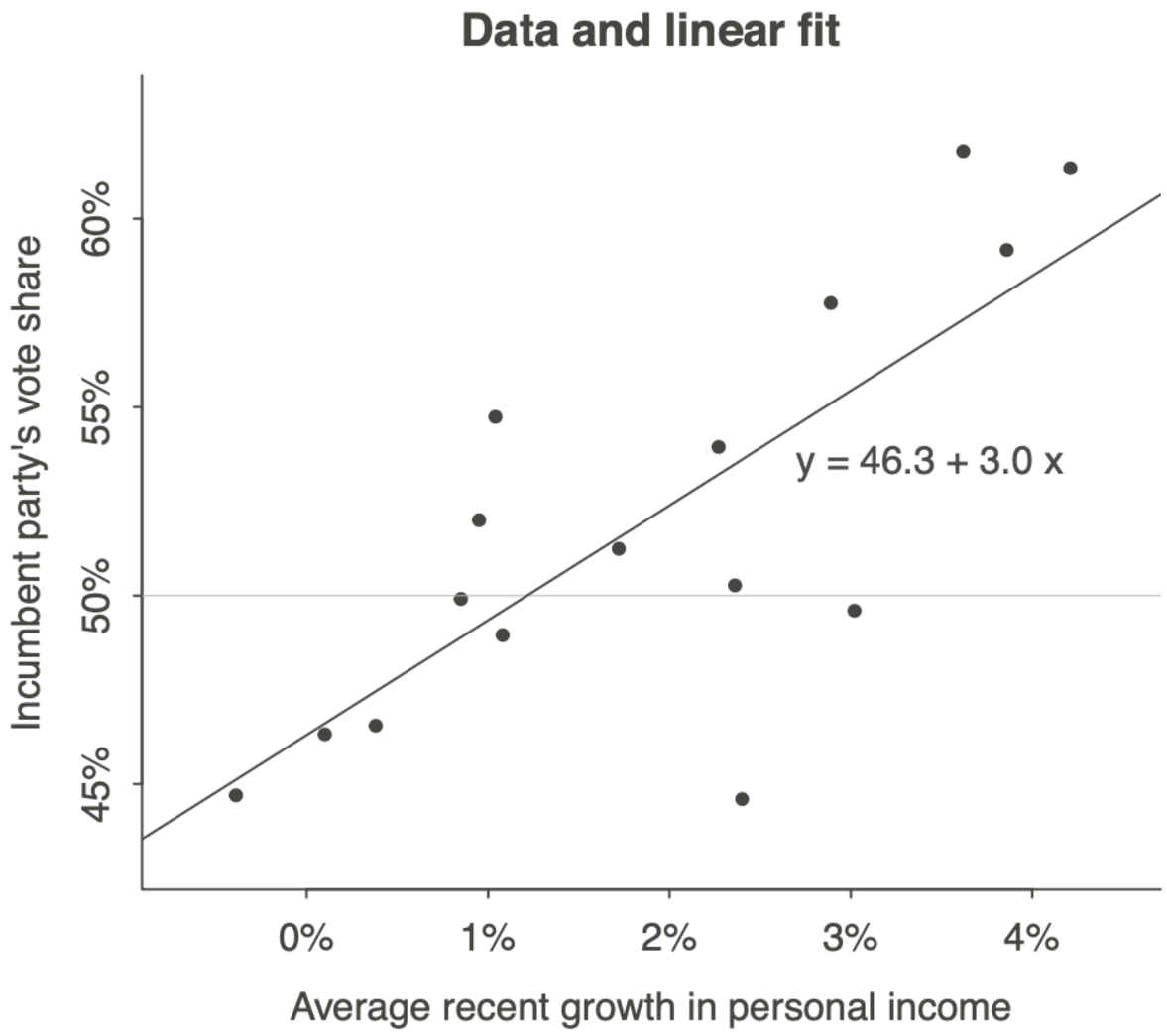
Regression Analysis





Chapter 1. Introduction and Basic Prerequisite Knowledge

Outline

1.1 What is Regression Analysis and Why Learn Regression

1.2 Selected Examples of Regression

1.3 Steps in Regression Analysis

1.4 Prerequisite I: Probability Distributions

1.5 Prerequisite II: Confidence intervals and t-Test

1.6 Prerequisite III: Elements of Matrix Algebra

1.1. What is Regression Analysis and Why Learn Regression

1.1 What is Regression Analysis and Why Learn Regression

What is Regression Analysis ?

Regression analysis *is a conceptually simple method for investigating functional relationships among variables.*

Example

A real estate appraiser may wish to relate the sale price of a home from selected physical characteristics of the building and taxes (local, school, county) paid on the building.

variables involved: ***sale price, floor size of building, location, number of bedrooms, etc.***

Example

We may wish to examine whether cigarette consumption is related to various socioeconomic and demographic variables such as age, education, income, and price of cigarettes.

variables involved: ***cigarette consumption, age, education, income, price, etc.***

1.1 What is Regression Analysis and Why Learn Regression

How to represent the relationship among variables?

*The relationship is expressed in the form of an equation or a model connecting the **response** or **dependent** variable and one or more **explanatory** or **predictor** variables.*

$$\text{response variable} = f(\text{predictor variables})$$

a function

Example

A real estate appraiser may wish to relate the sale price of a home from selected physical characteristics of the building and taxes (local, school, county) paid on the building.

*The **response** variable is the price of a home and the explanatory or **predictor** variables are the characteristics of the building and taxes paid on the building.*

Example

We may wish to examine whether cigarette consumption is related to various socioeconomic and demographic variables such as age, education, income, and price of cigarettes.

*The **response** variable is cigarette consumption (measured by the number of packs of cigarette sold in a given state on a per capita basis during a given year) and the explanatory or **predictor** variables are the various socioeconomic and demographic variables.*

1.1 What is Regression Analysis and Why Learn Regression

General Mathematical Formulation

We denote the response variable by Y and the set of predictor variables by $X_1, X_2, \cdots, X_p$, where p denotes the number of predictor variables.

Example $Y = \text{cigarette consumption}$

$X_1 = \text{age}, X_2 = \text{education}, X_3 = \text{income}, X_4 = \text{price}$

*The true relationship between Y and $X_1, X_2, \cdots, X_p$ can be approximated by a **regression** model*

$$Y = f(X_1, X_2, \cdots, X_p) + \varepsilon, \qquad (1.1)$$

where ε is assumed to be a random error representing the discrepancy in the approximation. It accounts for the failure of the model to fit the data exactly.

The function $f(X_1, \cdots, X_p)$ describes the relationship between Y and $X_1, X_2, \cdots, X_p$.

1.1 What is Regression Analysis and Why Learn Regression

General Mathematical Formulation

*An example is the **linear regression** model*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon, \tag{1.2}$$

*where $\beta_0, \beta_1, \cdots, \beta_p$ called the **regression parameters** of **coefficients**, are **unknown** constants to be determined (estimated) from the data. We follow the commonly used notational convention of denoting unknown parameters by Greek letters.*

*The **predictor** or **explanatory** variables are also called by other names such as **independent** variables, **covariates**, **regressors**, **factors**, and **carriers**.*

The name independent variable, though commonly used, is the least preferred, because in practice the predictor variables are rarely independent of each other.

1.1 What is Regression Analysis and Why Learn Regression

Why Learn Regression?

Some of the most important uses of regression are:

Prediction: *Modeling existing observations or forecasting new data.*

Example *In the real estate price example, we can use the regression model to **predict** the sale price of a **new** home based on its floor size, location, number of bedrooms.*

Decision making: *make policy or decision by the relationship among variables.*

Example *In the cigarette consumption example, if people get how the price influences cigarette consumption, **decision makers** can adjust the price to **control** the cigarette consumption.*

Exploring associations: *Summarizing how well one variable, or set of variables, predicts the outcome.*

Examples include identifying risk factors for a disease, attitudes that predict voting, and characteristics that make someone more likely to be successful in a job. More generally, one can use a model to explore associations, stratifications, or structural relationships between variables. Examples include associations between pollution levels and disease incidence, differential police stop rates of suspects by ethnicity, and growth rates of different parts of the body.

1.2 Selected Examples of Regression

1.2 Selected Examples of Regression

Agricultural Sciences

We Dairy Herd Improvement Cooperative (DHI) in upstate New York collects and analyzes data on milk production. One question of interest here is how to develop a suitable model to ***predict current milk*** *production from* ***a set of measured variables****. The response variable (current milk production in pounds) and the predictor variables are given in table.*

Samples are taken once a month during milking. The period that a cow gives milk is called lactation. Number of lactations is the number of times a cow has calved or given milk. The recommended management practice is to have the cow produce milk for about 305 days and then allow a 60day rest period before beginning the next lactation. The data set, consisting of ***199*** *observations, was compiled from the DHI milk production records.*

Table 1.1 Variables in Milk Production Data

Variable	Definition
Current	Current month milk production in pounds
Previous	Previous month milk production in pounds
Fat	Percent of fat in milk
Protein	Percent of protein in milk
Days	Number of days since present lactation
Lactation	Number of lactations
I79	Indicator variable (0 if Days $\leq$ 79 and 1 if Days $>$ 79)

response

predictor

at each of 199 factories, values of all these variables are collected.

1.2 Selected Examples of Regression

Industrial and Labor Relations

In 1947, the United States Congress passed the Taft-Hartley Amendments to the Wagner Act. The original Wagner Act had permitted the unions to use a Closed Shop Contract unless prohibited by state law. The Taft-Hartley Amendments made the use of Closed Shop Contract illegal and gave individual states the right to prohibit union shops as well. These right-to-work laws have caused a wave of concern throughout the labor movement.

A question of interest here is: What are the effects of these laws on the cost of living for a four-person family living on an intermediate budget in the United States? To answer this question a data set consisting of ***38*** *geographic locations has been assembled from various sources. The variables used are defined in the table.*

Table 1.2 Variables in Right-To-Work Laws Data

Variable	Definition
COL	Cost of living for a four-person family
PD	Population density (person per square mile)
URate	State unionization rate in 1978
Pop	Population in 1975
Taxes	Property taxes in 1972
Income	Per capita income in 1974
RTWL	Indicator variable (1 if there are right-to-work laws in the state and 0 otherwise)

response

predictor

at each of 38 locations, values of all these variables are collected.

1.2 Selected Examples of Regression

Environmental Sciences

*In a 1976 study exploring the relationship between water quality and land use, Haith (1976) obtained the measurements (shown in Table) on **20** river basins in New York State. A question of interest here is how the land use around a river basin contributes to the water pollution as measured by the mean nitrogen concentration (mglliter).*

Table 1.8 Variables in Study of Water Pollution in New York Rivers

Variable	Definition
Y	Mean nitrogen concentration (mg/liter) based on samples taken at regular intervals during the spring, summer, and fall months
X_1	Agriculture: percentage of land area currently in agricultural use
X_2	Forest: percentage of forest land
X_3	Residential: percentage of land area in residential use
X_4	Commercial/Industrial: percentage of land area in either commercial or industrial use

response

predictor

at each of 20 basins, values of all these variables are collected.

1.2 Selected Examples of Regression

Political Sciences

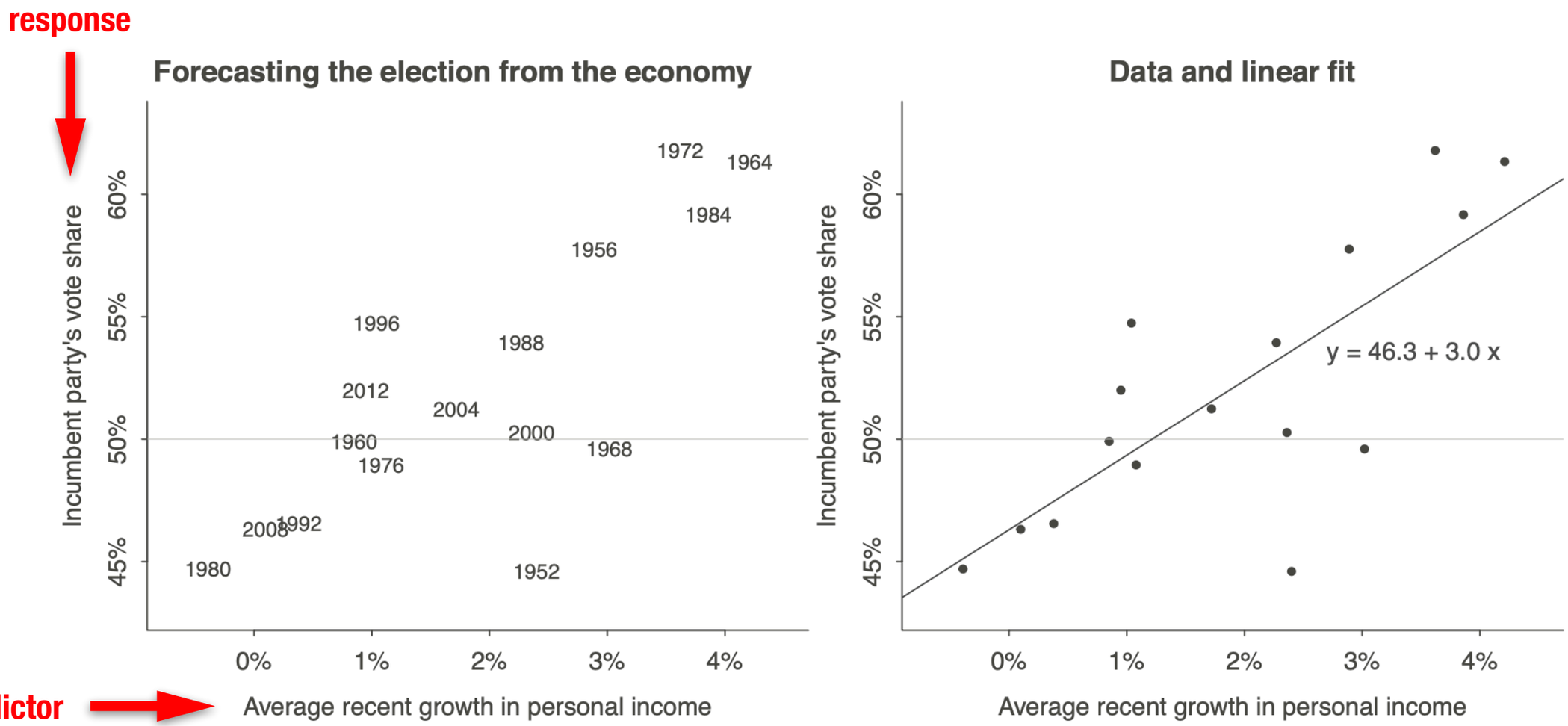


Figure 1.1: *Predicting elections from the economy: (a) the data, (b) the linear fit,* $y = 46.3 + 3.0x$.

Left figure shows the incumbent party's vote share in a series of U.S. presidential elections, plotted vs. a measure of economic growth in the period leading up to each election year. Right figure shows a linear regression fit to these data. The model allows us to predict the vote—with some uncertainty—given the economy and under the assumption that future elections are in some way like the past.

1.2 Selected Examples of Regression

Cost of Health Care

*The cost of delivery of health care has become an important concern. Getting data on this topic is extremely difficult because it is highly proprietary. These data were collected by the Department of Health and Social Services of the State of New Mexico and cover **52** of the 60 licensed facilities in New Mexico in 1988. The variables in these data are the characteristics which describe the facilities size, volume of usage, expenditures, and revenue. The location of the facility is also indicated, whether it is in the rural or non-rural area. Specific definitions of the variables are given in Table.site.*

There are several ways of looking at a body of data and extracting various kinds of information. For example, (a) Are rural facilities different from non-rural facilities? and (b) How do the hospital characteristics affect the total patient care revenue?

Table 1.12 Variables in Cost of Health Care Data

Variable	Definition
RURAL	Rural home (1) and nonrural home (0)
BED	Number of beds in home
MCDAYS	Annual medical in-patient days (hundreds)
TDAYS	Annual total patient days (hundreds)
PCREV	Annual total patient care revenue ($100)
NSAL	Annual nursing salaries ($100)
FEXP	Annual facilities expenditures ($100)
NETREV	PCREV – NSAL – FEXP

at each of 52 facilities, values of all these variables are collected.

1.3 Steps in Regression Analysis

1.3 Steps in Regression Analysis

Regression analysis includes the following steps:

- Statement of the problem
- Selection of potentially relevant variables
- Data collection
- Model specification
- Choice of fitting method
- Model fitting
- Model validation and criticism
- Using the chosen model(s) for the solution of the posed problem.

1.3 Steps in Regression Analysis

Step 1: Statement of the Problem

Regression analysis usually starts with a formulation of the problem. This includes the determination of the question(s) to be addressed by the analysis. The problem statement is the first and perhaps the most important step in regression analysis. It is important because an ill-defined problem or a misformulated question can lead to wasted effort. It can lead to the selection of irrelevant set of variables or to a wrong choice of the statistical method of analysis. A question that is not carefully formulated can also lead to the wrong choice of a model.

Suppose we wish to determine whether or not an employer is discriminating against a given group of employees, say women. Data on salary, qualifications, and gender are available from the company's record to address the issue of discrimination. There are several definitions of employment discrimination in the literature. For example, discrimination occurs when on the average (a) women are paid less than equally qualified men, or (b) women are more qualified than equally paid men. To answer the question: "On the average, are women paid less than equally qualified men?" we choose salary as a response variable, and qualification and gender as predictor variables. But to answer the question: "On the average, are women more qualified than equally paid men?" we choose qualification as a response variable and salary and gender as predictor variables, that is, the roles of variables have been switched.

1.3 Steps in Regression Analysis

Step 2: Selection of Potentially Relevant Variables

The next step after the statement of the problem is to select a set of variables that are thought by the experts in the area of study to explain or predict the response variable. The response variable is denoted by Y and the explanatory or predictor variables are denoted by $X_1, X_2, \cdots, X_p$, where p denotes the number of predictor variables. An example of a response variable is the price of a single-family house in a given geographical area. A possible relevant set of predictor variables in this case is: area of the lot, area of the house, age of the house, number of bedrooms, number of bathrooms, type of neighborhood, style of the house, amount of real estate taxes, and so forth.

1.3 Steps in Regression Analysis

Step 3: Data Collection

*The next step after the selection of potentially relevant variables is to collect the data from the environment under study to be used in the analysis. Sometimes the data are collected in a controlled setting so that factors that are not of primary interest can be held constant. More often the data are collected under nonexperimental conditions where very little can be controlled by the investigator. In either case, the collected data consist of observations on **n subjects**. Each of these n observations consists of measurements for each of the potentially relevant variables. The data are usually recorded as in Table.*

Table 1.14 Notation for Data Used in Regression Analysis

Observation	Response Variable	Predictors			
Number	Y	X_1	X_2	$\cdots$	X_p
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2p}
3	y_3	x_{31}	x_{32}	$\cdots$	x_{3p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}

*A **column** represents a variable, whereas a **row** represents an **observation**, which is a set of p+1 values for a single subject (e.g., a house); one value for the response variable and one value for each of the p predictors. The notation x_{ij} refers to the **i**-th value of the **j**-th variable. The first subscript refers to observation number and the second refers to variable number.*

1.3 Steps in Regression Analysis

Step 3: Data Collection

*Each of the variables in Table can be classified as either **quantitative** or **qualitative**. Examples of quantitative variables are the house price, number of bedrooms, age, and taxes. Examples of qualitative variables are neighborhood type (e.g., good or bad neighborhood) and house style (e.g., ranch, colonial, etc.). We deal **mainly** with the cases where the **response** variable is **quantitative**.*

*A technique used in cases where the **response** variable is binary is called **logistic regression**.*
*This is introduced in the final chapters. In regression analysis, the **predictor** variables can be either **quantitative** and/or **qualitative**. For the purpose of computations, however, the **qualitative** variables, if any, have to be coded into a set of **indicator** or **dummy** variables as discuss later.*

1.3 Steps in Regression Analysis

Step 4: Model Specification

*The form of the model that is thought to relate the response variable to the set of predictor variables can be specified initially by the experts in the area of study based on their knowledge or their **objective** and/or **subjective** judgments. The **hypothesized** model can then be either **confirmed** or **refuted** by the analysis of the collected data. Note that the model needs to be specified **only in form**, but it can still depend on **unknown** parameters.*

We need to select the form of the function $f(X_1, X_2, \cdots, X_p)$.

*This function can be classified into two types: **linear** and **nonlinear**. An example of a **linear** function is*

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \tag{1.3}$$

while an example of nonlinear function is

$$Y = \beta_0 + e^{\beta_1 X_1} + \varepsilon. \tag{1.4}$$

1.3 Steps in Regression Analysis

Step 4: Model Specification

*Note that the term linear (nonlinear) here does **not** describe the relationship between* Y *and* $X_1, X_2, \cdots, X_p$

*It is related to the fact that the **regression parameters** enter the equation **linearly (nonlinearly)**. Each of the following models are **linear**:*

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon, \\ Y &= \beta_0 + \beta_1 \ln X + \varepsilon, \end{aligned}$$

Linear in the regression coefficients.

*because in each case the parameters **enter linearly** although the relationship between Y and X is **nonlinear**. This can be seen if the two models are reexpressed, respectively, as follows:*

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \\ Y &= \beta_0 + \beta_1 X_1 + \varepsilon, \end{aligned}$$

where in the first equation $X_1 = X$ and $X_2 = X^2$ and in the second equation $X_1 = \ln X$

Variables here are reexpressed or transformed. Transformation is dealt with in later chapters.

1.3 Steps in Regression Analysis

Step 4: Model Specification

*All nonlinear functions that can be **transformed** into linear functions are called **linearizable functions**. Accordingly, the class of linear models is actually **wider** than it might appear at first sight because it includes all **linearizable** functions. Note, however, that **not** all nonlinear functions are linearizable. For example, it is not possible to linearize the nonlinear function* $e^{\beta_1 X_1}$.
*Some authors refer to nonlinear functions that are **not** linearizable as **intrinsically** nonlinear functions.*

*Regression equation containing only **one** predictor variable is called a **simple** regression equation.*
*An equation containing **more than one** predictor variable is called a **multiple** regression equation.*
*An example of **simple** regression would be an analysis in which the time to repair a machine is studied in relation to the number of components to be repaired. Here we have one **response** variable (time to repair the machine) and one **predictor** variable (number of components to be repaired).*
*An example of a very complex **multiple** regression situation would be an attempt to explain the age-adjusted mortality rates prevailing in different geographic regions (**response** variable) by a large number of environmental and socioeconomic factors (**predictor** variables).*
***Both** types of problems are treated in this book.*

1.3 Steps in Regression Analysis

Step 4: Model Specification

Table 1.15 Various Classifications of Regression Analysis

Type of Regression	Conditions
Univariate	Only one quantitative response variable
Multivariate	Two or more quantitative response variables
Simple	Only one predictor variable
Multiple	Two or more predictor variables
Linear	All parameters enter the equation linearly, possibly after transformation of the data
Nonlinear	The relationship between the response and some of the predictors is nonlinear or some of the parameters appear nonlinearly, but no transformation is possible to make the parameters appear linearly
Analysis of variance	All predictors are qualitative variables
Analysis of covariance	Some predictors are quantitative variables and others are qualitative variables
Logistic	The response variable is qualitative

When we deal with with one response variable, regression analysis is called uni- variate regression and in cases where we have two or more response variables, the regression is called multivariate regression. Simple and multiple regressions should not be confused with univariate versus multivariate regressions. The distinction between simple and multiple regressions is determined by the number of predictor variables (simple means one predictor variable and multiple means two or more predictor variables), whereas the distinction between univariate and multivariate regressions is determined by the number of response variables (univariate means one response variable and multivariate means two or more response variables). In this book we consider only univariate regression (both simple and multiple, linear and nonlinear).

The various classifications of regression analysis are shown in Table 1.15.

1.3 Steps in Regression Analysis

Step 5: Method of Fitting

After the model has been defined and the data have been collected, the next task is to estimate the parameters of the model based on the collected data. This is also referred to as *parameter estimation* or *model fitting*. The most commonly used method of estimation is called the *least squares* method. Under certain assumptions (to be discussed in detail in this course), least squares method produce estimators with desirable properties. In this book we will deal mainly with the least squares method and its variants (e.g., weighted least squares). In some instances (e.g., when one or more of the assumptions does not hold) other estimation methods may be superior to least squares. The other estimation methods that we consider in this book are the *maximum likelihood* method, the *ridge regression*

1.3 Steps in Regression Analysis

Step 6: Model Fitting

*The next step in the analysis is to estimate the regression parameters or to fit the model to the collected data using the chosen estimation method (e.g., **least squares**).*
The estimates of the regression parameters $\beta_0, \beta_1, \cdots, \beta_p$ *are denoted by* $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$.
The estimated regression equation then becomes

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p. \tag{1.5}$$

if linear model

*A **hat** on top of a parameter denotes an estimate of the parameter. The value* $\hat{Y}$ *(pronounced as Y-hat) is called the **fitted** value.*

Using this equation, we can compute n fitted values, one for each of the n observations in our data.

For example, the ith fitted value $\hat{y}_i$ is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}, \quad i = 1, 2, \cdots, n, \tag{1.6}$$

where $x_{i1}, \cdots, x_{ip}$ are the values of the p predictor variables for the ith observation.

1.3 Steps in Regression Analysis

Step 6: Model Fitting

The estimated equation can be used to predict the response variable for any values of the predictor variables not observed in our data. In this case, the obtained $\hat{Y}$ is called the *predicted* value. The difference between fitted and predicted values is that the fitted value refers to the case where the values used for the predictor variables correspond to one of the n observations in our data, but the predicted values are obtained for any set of values of the predictor variables. It is generally not recommended to predict the response variable for a set of values of the predictor variables far outside the range of our data. In cases where the values of the predictor variables represent future values of the predictors, the predicted value is referred to as the *forecasted* value.

1.3 Steps in Regression Analysis

Step 7: Model Criticism and Selection

*The **validity** of a statistical method, such as **regression** analysis, depends on certain **assumptions**. Assumptions are usually made about the **data** and the **model**. The accuracy of the analysis and the conclusions derived from an analysis depends **crucially** on the **validity** of these assumptions.*

*Before using **the estimated equation** for any purpose, we first need to determine whether the specified assumptions hold. We need to address the following questions:*

1. What are the required assumptions?
2. For each of these assumptions, how do we determine whether or not the assumption is valid?
3. What can be done in cases where one or more of the assumptions does not hold?

*The **standard** regression assumptions will be specified and the above questions will be addressed in great detail in various parts of this course. We emphasize here that validation of the assumptions **must** be made **before** any conclusions are drawn from the analysis.*

1.3 Steps in Regression Analysis

Step 7: Model Criticism and Selection

*Regression analysis is viewed here as a **iterative** process, a process in which the outputs are used to **diagnose**, **validate**, **criticize**, and possibly **modify** the inputs. The process has to be **repeated** until a **satisfactory** output has been obtained. A satisfactory output is an estimated model that satisfies the assumptions and fits the data reasonably well. This iterative process is illustrated schematically in the following figure.*

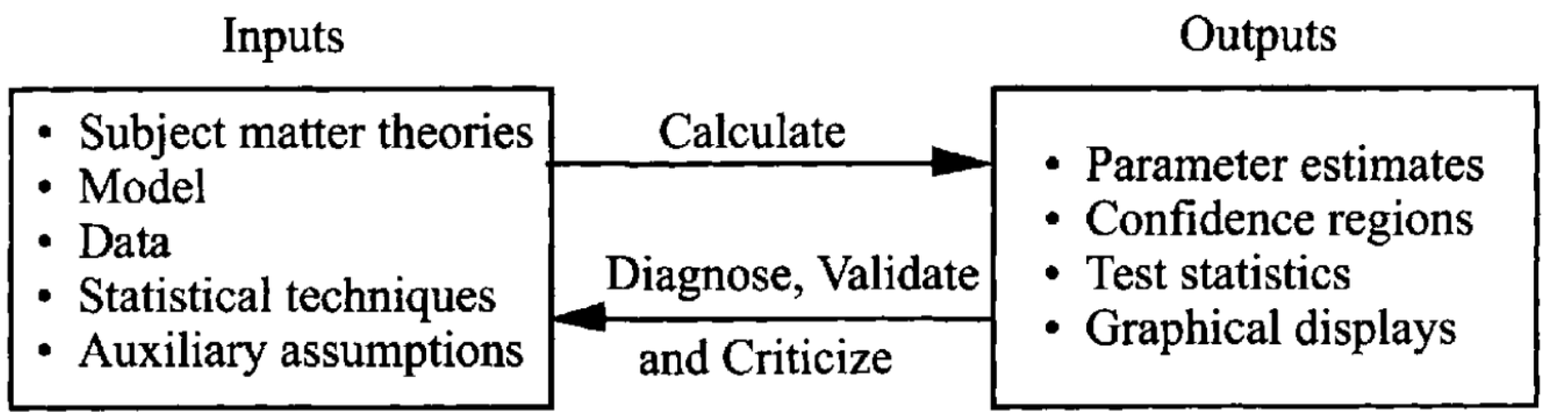


Figure 1.1 A schematic illustration of the iterative nature of the regression process.

1.4 Prerequisite I: Probability Distributions

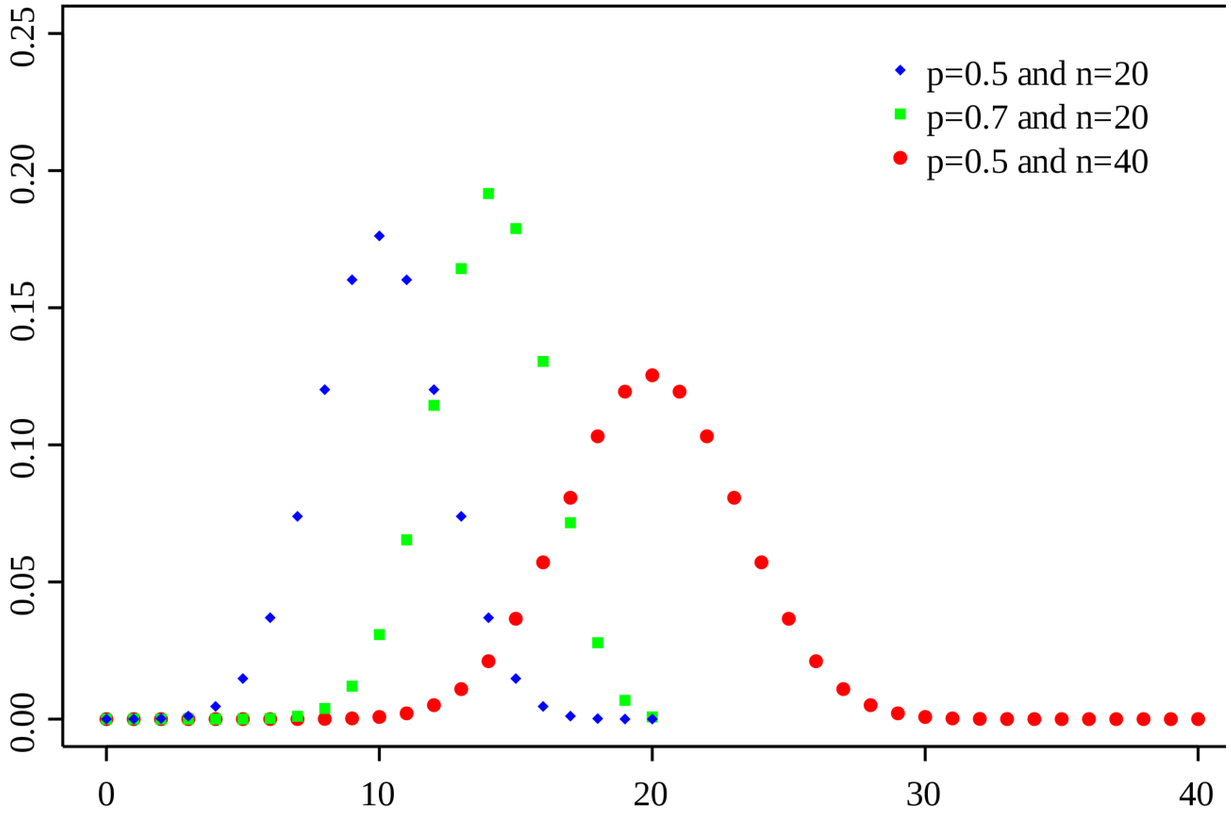
1.4 Prerequisite I: Probability Distributions

Bernoulli

$$X \sim \text{Bernoulli}(p) \quad \Leftrightarrow \quad P(X=1)=p, P(X=0)=1-p$$

Binomial

$$X \sim \text{Bin}(n,p) \quad \Leftrightarrow \quad P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}, k=0,1,\cdots,n$$

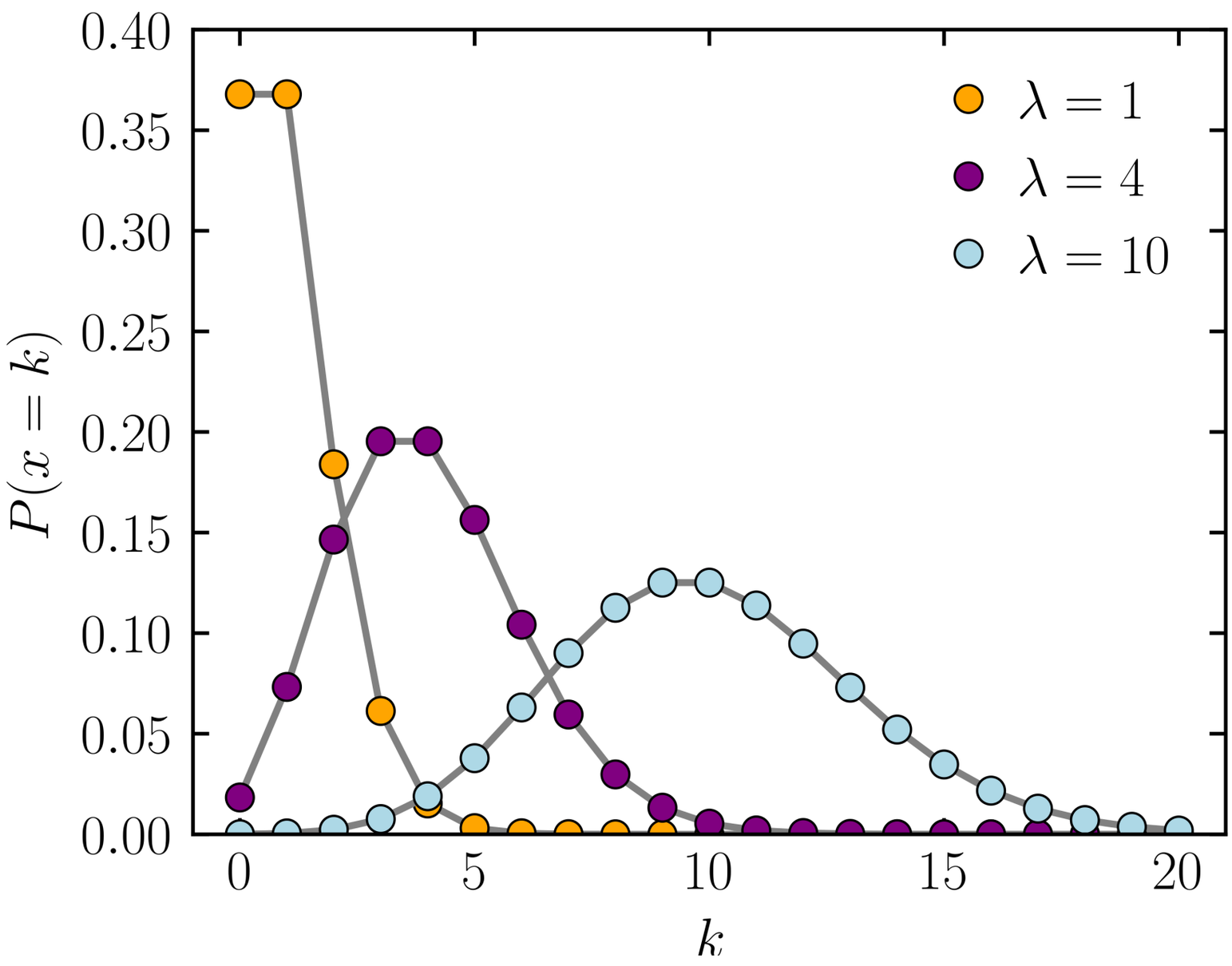


pmf of Binomial

1.4 Prerequisite I: Probability Distributions

Poisson

$$X \sim \text{Poisson}(\lambda) \quad \Leftrightarrow \quad P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \ k = 0, 1, \cdots$$



pmf of Poisson

1.4 Prerequisite I: Probability Distributions

Normal

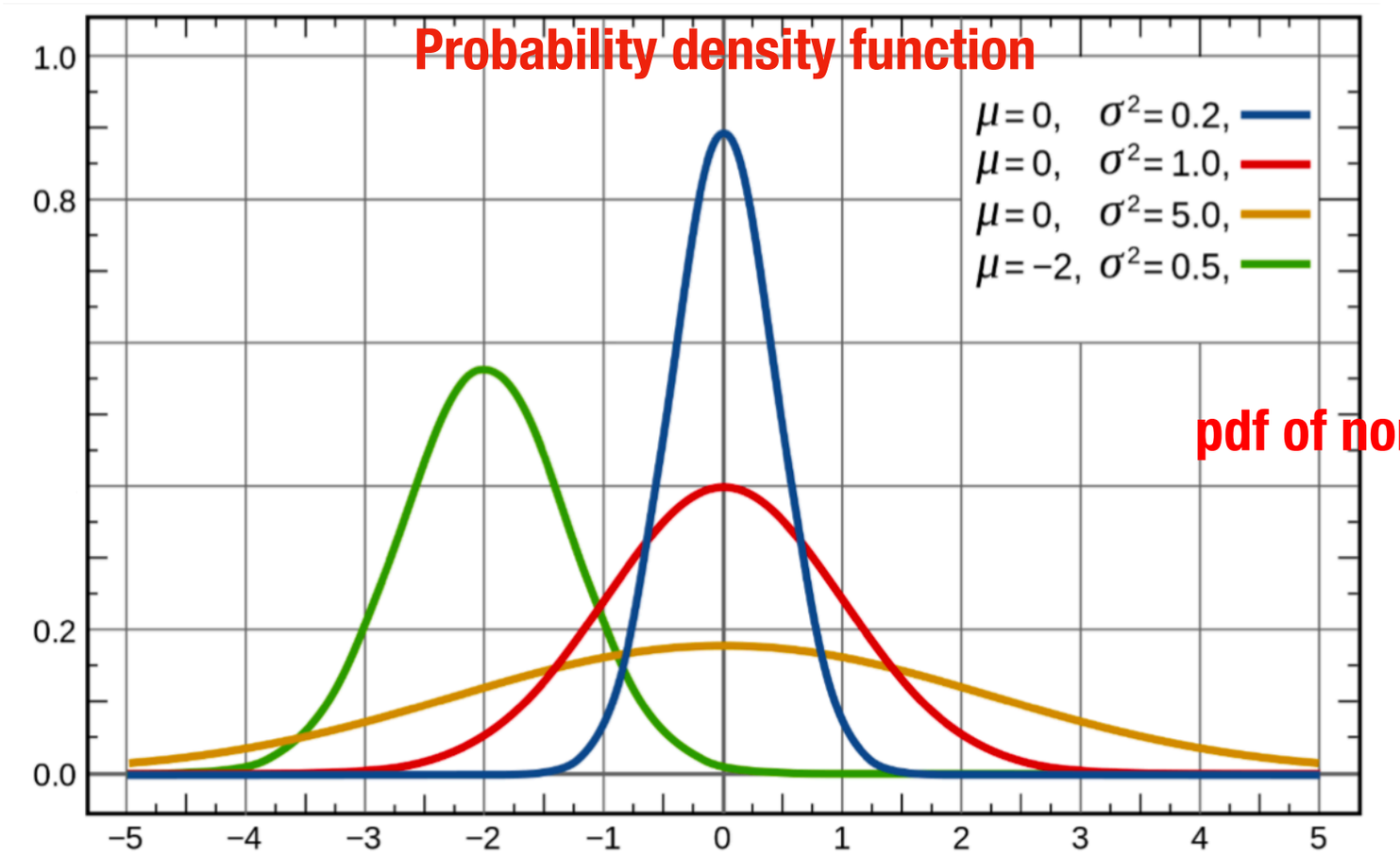
$N(\mu, \sigma^2)$

$$X \sim \text{Normal}(\mu, \sigma^2) \quad \Leftrightarrow \quad f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \forall x$$

$$X \sim N(\mu, \sigma^2)$$
$$\Rightarrow aX + b \sim N(a\mu + b, a^2\sigma^2)$$

The 3-σ rule:

$$P(|X-\mu| > 3\sigma) \approx 0.0026$$



Standard Normal

$N(0, 1)$

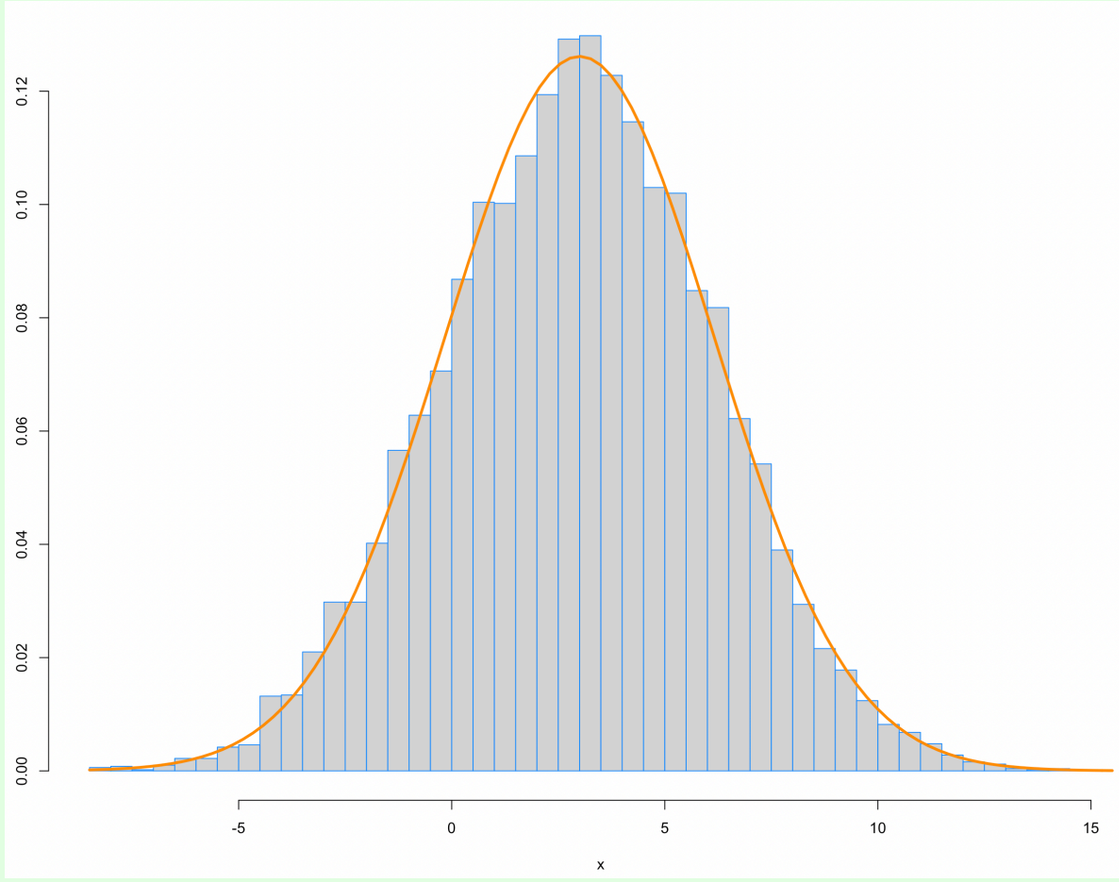
$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2};$$
$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} \, dy.$$

1.4 Prerequisite I: Probability Distributions

Empirical pdf by Histogram

If we obtain $10,000$ realizations of $N(3,10)$, how does the histogram look like?

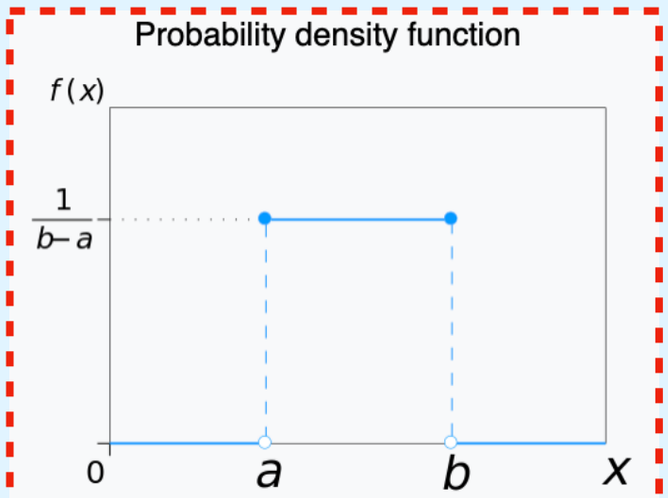
```
## normal distribution
n=10000
x=rnorm(n,3,sqrt(10))
hist(x, prob = TRUE, breaks = 50, main = "Histogram", border = "dodgerblue")
curve(dnorm(x, mean = 3, sd = sqrt(10)),
      col = "darkorange", add = TRUE, lwd = 3)
```



1.4 Prerequisite I: Probability Distributions

Uniform Distribution

$$X \sim \text{Unif}(a, b) \quad \Leftrightarrow \quad f(x) = \frac{1}{b-a}, \forall x \in (a, b)$$



pdf of uniform

Exponential distribution

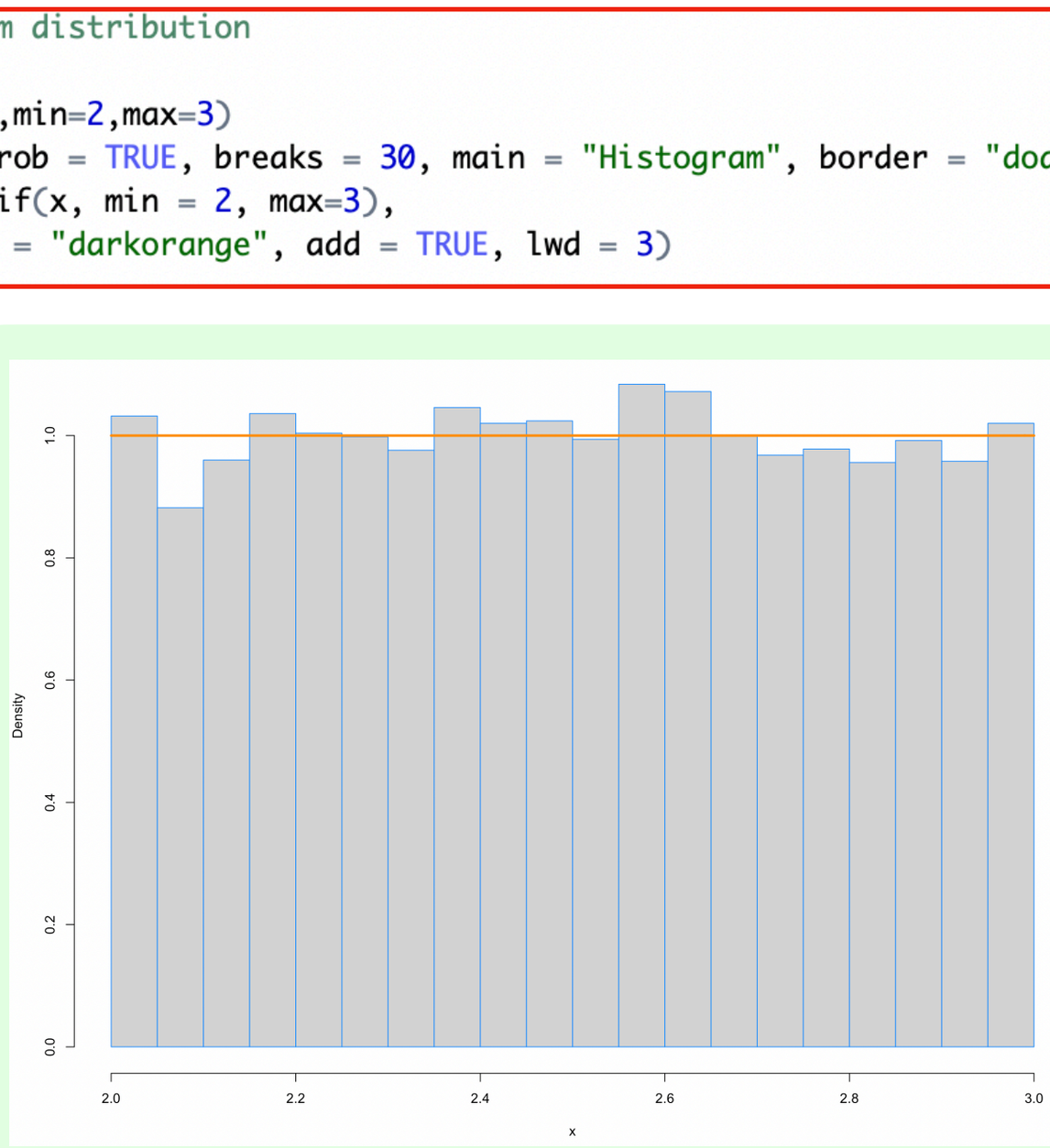
$$X \sim \text{Exp}(\lambda) \quad \Leftrightarrow \quad f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

1.4 Prerequisite I: Probability Distributions

Empirical pdf by Histogram

If we obtain $10,000$ realizations of $\text{Unif}(2,3)$, how does the histogram look like?

```
## uniform distribution
n=10000
x=runif(n,min=2,max=3)
hist(x, prob = TRUE, breaks = 30, main = "Histogram", border = "dodgerblue")
curve(dunif(x, min = 2, max=3),
      col = "darkorange", add = TRUE, lwd = 3)
```



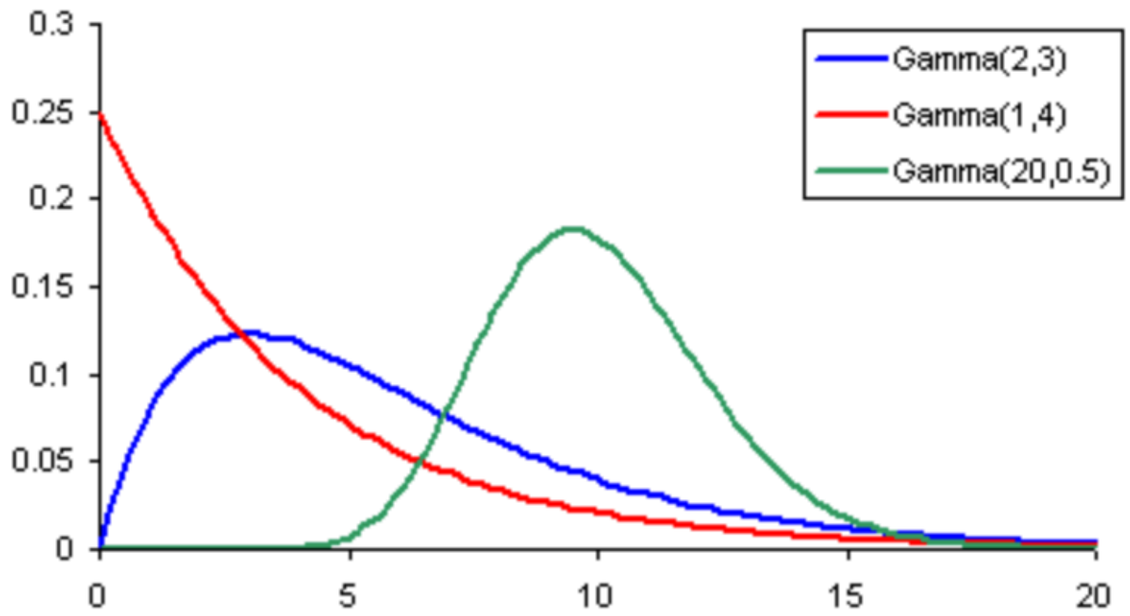
1.4 Prerequisite I: Probability Distributions

Gamma Distribution

$$X \sim \Gamma(\alpha, \lambda) \quad \Leftrightarrow \quad f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$ **pdf of Gamma**

$\Gamma(\alpha)$, called the gamma function, is defined by $\Gamma(\alpha) = \int_0^\infty e^{-y} y^{\alpha-1} dy$

For integer value n, $\Gamma(n) = (n-1)!$

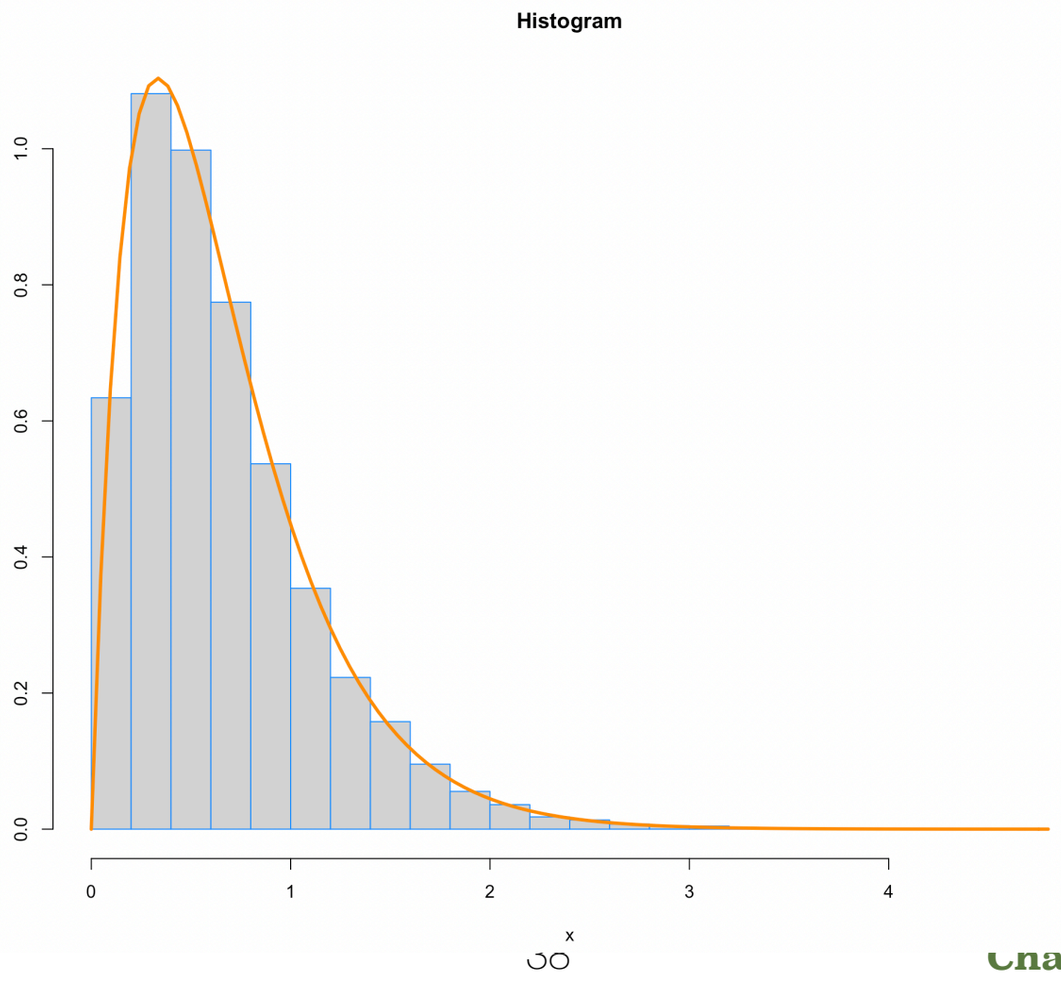


1.4 Prerequisite I: Probability Distributions

Empirical pdf by Histogram

If we obtain 10,000 realizations of Gamma$(2, 3)$, how does the histogram look like?

```
## Gamma distribution
n=10000
x=rgamma(n,shape=2,scale=1/3)
hist(x, prob = TRUE, breaks = 30, main = "Histogram", border = "dodgerblue")
curve(dgamma(x, shape = 2, scale=1/3),
      col = "darkorange", add = TRUE, lwd = 3)
```



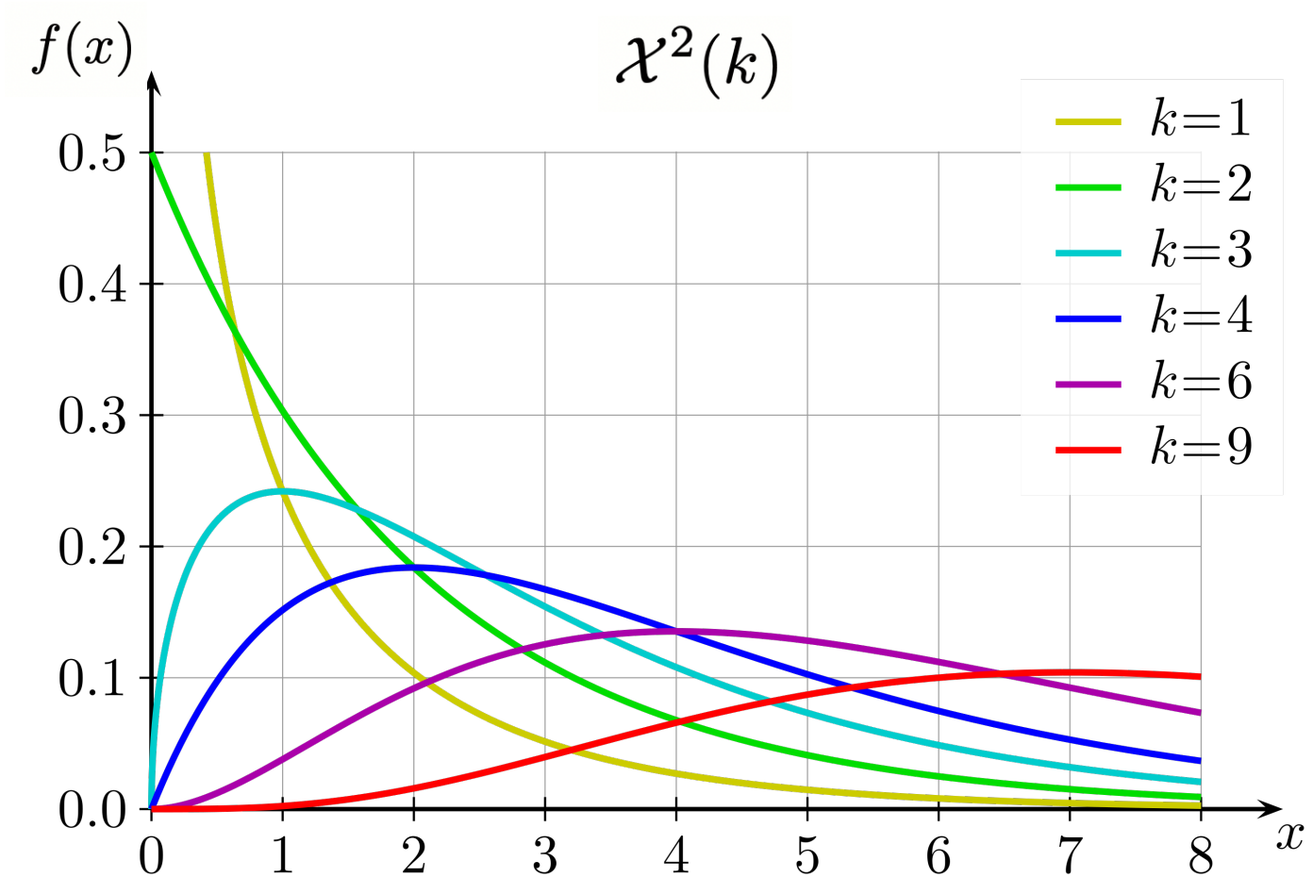
1.4 Prerequisite I: Probability Distributions

Chi-squared distribution

called the degrees of freedom

$$X \sim \mathcal{X}^2(k) \quad \Leftrightarrow \quad \Gamma(\frac{k}{2}, \frac{1}{2})$$

If $Z_1, \cdots, Z_k$ are i.i.d. standard normal, then the sum of squares $\sum_{i=1}^{k} Z_i^2 \sim \mathcal{X}^2(k)$



pdf of Chi-squared

1.4 Prerequisite I: Probability Distributions

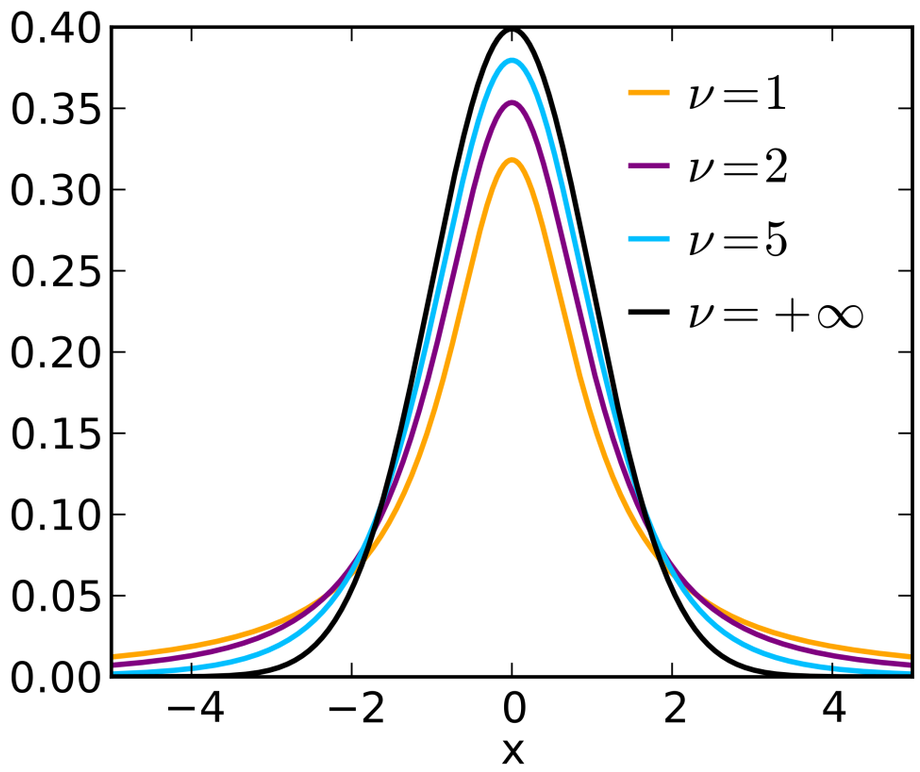
t-distribution

$$X \sim t_\nu \quad \Leftrightarrow \quad f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad \forall x$$

pdf

called the degrees of freedom

If $X_1, \cdots, X_n$ are i.i.d. $N(\mu, \sigma^2)$, then $\dfrac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$ where $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$, and $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$



pdf of t-distribution

1.4 Prerequisite I: Probability Distributions

F-distribution

$$X \sim F(d_1, d_2) \quad \Leftrightarrow \quad f(x) = \frac{1}{B(d_1/2, d_2/2)} \left(\frac{d_1}{d_2}\right)^{d_1/2} x^{d_1/2-1} \left(1 + \frac{d_1}{d_2} x\right)^{-(d_1+d_2)/2}, \quad \forall x \geq 0$$

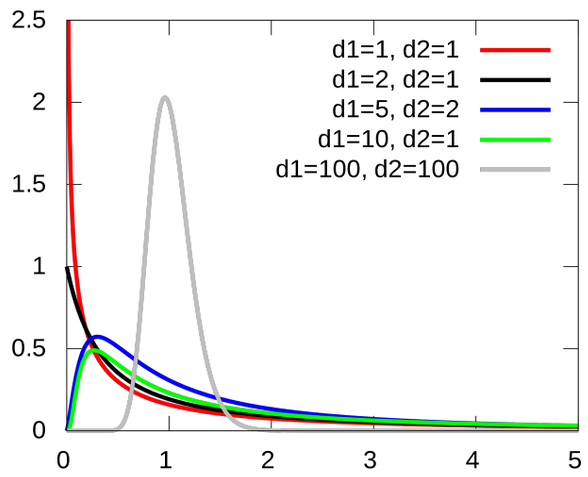
called the degrees of freedom

pdf

$$B(a, b), \text{ called Beta function, is defined by } B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx$$

If $X_1 \sim \mathcal{X}^2(d_1)$ and $X_2 \sim \mathcal{X}^2(d_2)$ are independent, then $\dfrac{X_1/d_1}{X_2/d_2} \sim F(d_1, d_2)$

If $X_1 \sim \Gamma(\alpha_1, \lambda_1)$ and $X_2 \sim \Gamma(\alpha_2, \lambda_2)$ are independent, then $\dfrac{\lambda_1 X_1/\alpha_1}{\lambda_2 X_2/\alpha_2} \sim F(2\alpha_1, 2\alpha_2)$



pdf of F-distribution

1.4 Prerequisite I: Probability Distributions

Central Limit Theorem

Theorem If $X_1, \cdots, X_n$ are i.i.d. random variables, then as n becomes large (≥ 30),

$$n^{1/2}. \frac{\bar{X} - \mathbb{E}X}{(\text{var}X)^{1/2}} \text{ is approximately } N(0,1)$$

where $\bar{X} = (X_1 + \cdots + X_n)/n$ is the sample mean.

```
#### Simulation for CLT
set.seed(2021)
n=200    ## sample size
x=runif(n,min=0,max=1)   ## generate 100 Binomial random varialbes with Bin(10, 0.3)
meanx=(mean(x)-0.5)/sqrt(1/(12*n))
```

```
> meanx
[1] -0.4768709
```

So we get a number -0.47687, why do we say that the sample mean follows a distribution?

1.4 Prerequisite I: Probability Distributions

Central Limit Theorem

Class Discussions

How do we numerically verify that $n^{1/2} \cdot \frac{\bar{X} - \mathbb{E}X}{(\mathrm{var}(X))^{1/2}}$ is approximately $N(0,1)$?

Random Variable $\Longleftrightarrow$ Realization

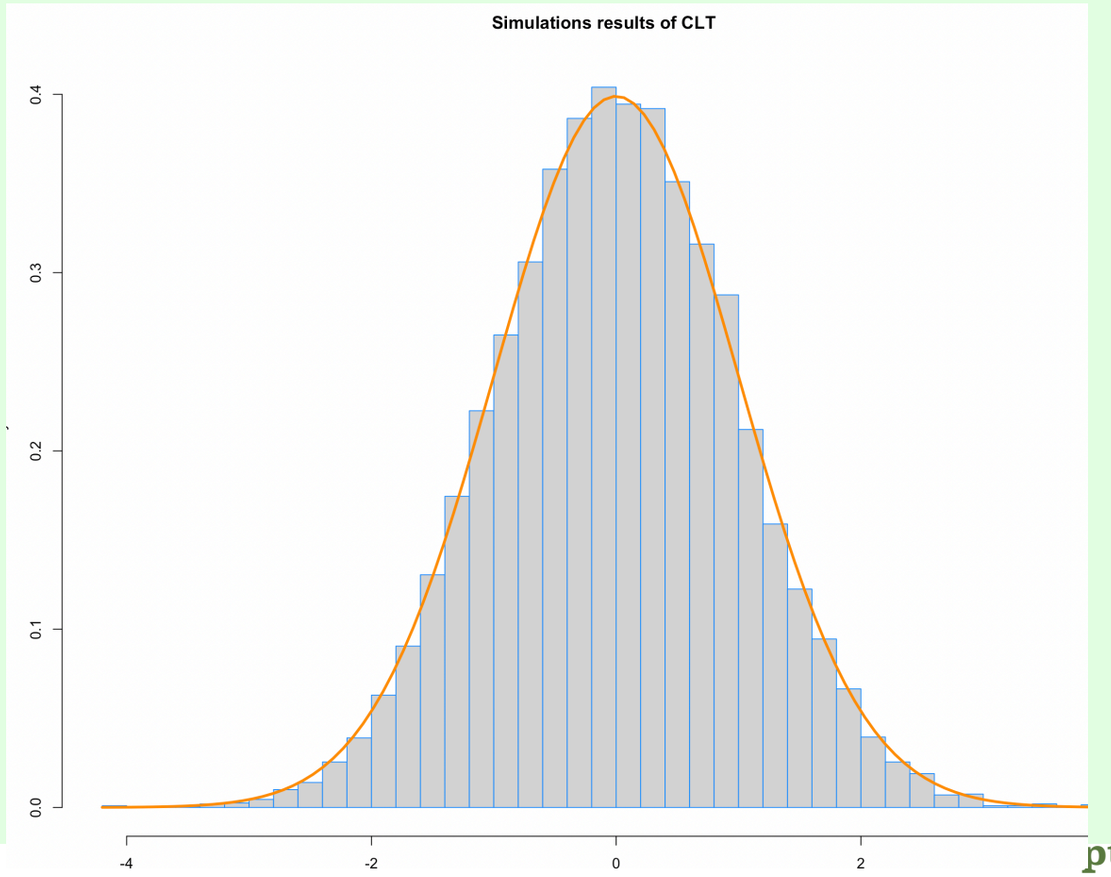
1.4 Prerequisite I: Probability Distributions

Central Limit Theorem

Repeat the simulation for $10,000$ times, obtaining $10,000$ realizations of the random variable $n^{1/2} \cdot \dfrac{\bar{X} - \mathbb{E}X}{(\text{var } X)^{1/2}}$

```r
sim_repeat=10000   ##  repeat the simulation for 10000 times to obtain 10000 realizations, then use histogram to check the empirical distribution
normalized_meanx=rep(0,sim_repeat)
for (i in 1:sim_repeat) {
  x=runif(n,min=0,max=1)
  normalized_meanx[i]=(mean(x)-0.5)/sqrt(1/(12*n))
}

#### draw the histogram
hist(normalized_meanx, prob = TRUE, breaks = 50, main = "Simulations results of CLT", border = "dodgerblue")
curve(dnorm(x, mean = 0, sd = sqrt(1)),
      col = "darkorange", add = TRUE, lwd = 3)
```



1.5 Prerequisite II: Confidence Interval and t-Test

1.5 Prerequisite II: Confidence Interval and t-Test

Let θ be a parameter (or "thing") that we want to estimate. Let $\hat{\theta}$ be an estimate of θ ("estimate of thing").

Typically, $\hat{\theta}$ will follow a normal distribution, either exactly because of the normality of the observations , or approximately due to the effect of the Central Limit Theorem.

Example

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ is exactly normal if } X_i \text{ are normal}$$

or approximately normal if n is large, i.e., $n \geq 30$ by CLT

$\bar{X}$ is an estimate of the expectation $\mu = \mathbb{E}X_1$

Let $\sigma_{\hat{\theta}}$ be the true standard deviation of $\hat{\theta}$ and let $\text{se}(\hat{\theta})$ be the estimated standard deviation of $\hat{\theta}$ ("standard error of estimate of thing"), based on ν degrees of freedom

Example

$$\text{se}(\bar{X}) = \frac{S}{\sqrt{n}}, \quad \text{where } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

the degrees of freedom is $\nu = n - 1$.

1.5 Prerequisite II: Confidence Interval and t-Test

Oftentimes, $\dfrac{\hat{\theta}-\theta}{\text{se}(\hat{\theta})}$ follows a t-distribution with degrees of freedom $\nu = n-1$.

$$\frac{\text{estimate of thing} - \text{thing}}{\text{standard error of estimate of thing}} \sim t_\nu$$

Example

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

Confidence interval based on t-distribution

$$P(T_{n-1} \geq t_{n-1,\alpha/2}) = \alpha/2$$

Percentile of t-distribution

The $(1-\alpha)100\%$ confidence interval of thing is

$$\left\{\begin{matrix}\text{Estimate}\\ \text{of thing}\end{matrix}\right\} \pm \left\{\begin{matrix}\text{A } t \text{ percentage point}\\ \text{leaving } \alpha/2 \text{ in the}\\ \text{upper tail, based on}\\ \nu \text{ degrees of freedom}\end{matrix}\right\} \left\{\begin{matrix}\text{Standard error}\\ \text{of estimate}\\ \text{of thing}\end{matrix}\right\}.$$

Example

$$\left[\bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}, \quad \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right]$$

1.5 Prerequisite II: Confidence Interval and t-Test

To test $\theta = \theta_0$, where θ_0 is some specified value of θ that is presumed to be valid.

We evaluate the statistics:

$$t = \frac{\hat{\theta} - \theta_0}{\text{se}(\hat{\theta})}$$

or

$$t = \frac{\left\{\begin{matrix}\text{Estimate}\\ \text{of thing}\end{matrix}\right\} - \left\{\begin{matrix}\text{Postulated or test}\\ \text{value of thing}\end{matrix}\right\}}{\left\{\begin{matrix}\text{Standard error of}\\ \text{estimate of thing}\end{matrix}\right\}}.$$

one-sided test **two-sided test**

The observed value of t or $|t|$ is then compared with the critical value $t_{\nu,\alpha}$ or $t_{\nu,\alpha/2}$

Example

$$H_0 : \theta = \theta_0 \quad v.s. \quad H_1 : \theta \neq \theta_0$$

Reject H_0 if $|t| > t_{\nu,\alpha/2}$ at the significance level α.

$\alpha = 0.05$ or $\alpha = 0.01$

1.6 Prerequisite III: Elements of Matrix Algebra

1.6 Prerequisite III: Elements of Matrix Algebra

bold-face lower-case letter denote a vector

$$\mathbf{a} = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \qquad \mathbf{a}' = (-1, 0, 1)$$

column vector **row vector**

A $p \times q$ matrix $\mathbf{M}$ is a rectangular array of numbers containing p rows and q columns

$$\mathbf{M} = \begin{bmatrix} m_{11} & m_{12} & \dots & m_{1q} \\ m_{21} & m_{22} & \dots & m_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ m_{p1} & m_{p2} & \dots & m_{pq} \end{bmatrix}.$$

bold-face upper-case letter denote a matrix

1.6 Prerequisite III: Elements of Matrix Algebra

Sum and Difference of Matrices

The sum (or differences) of two matrices is the matrix each of whose elements is the sum (or difference) of the corresponding elements of the matrices added (or subtracted). ***For example***

$$\begin{bmatrix} 1 & 3 \\ 1 & 0 \\ 1 & 2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 7 & 5 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 1+0 & 3+0 \\ 1+7 & 0+5 \\ 1+2 & 2+1 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 8 & 5 \\ 3 & 3 \end{bmatrix}$$

Transpose

$$\mathbf{M} = \begin{pmatrix} m_{11} & m_{12} & \cdots & m_{1q} \\ m_{21} & m_{22} & \cdots & m_{2q} \\ \vdots & \vdots & \vdots & \vdots \\ m_{p1} & m_{p2} & \cdots & m_{pq} \end{pmatrix} \quad \rightarrow \quad \mathbf{M}' = \begin{pmatrix} m_{11} & m_{21} & \cdots & m_{p1} \\ m_{12} & m_{22} & \cdots & m_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ m_{1q} & m_{2q} & \cdots & m_{pq} \end{pmatrix}$$

Symmetry

A matrix $\mathbf{M}$ is said to be *symmetric* if $\mathbf{M}' = \mathbf{M}$.

1.6 Prerequisite III: Elements of Matrix Algebra

Product of Matrices

Suppose we have two matrices, $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$, then $\mathbf{C} = \mathbf{AB} \in \mathbb{R}^{m\times p}$

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{np} \end{pmatrix} \longrightarrow \mathbf{C} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mp} \end{pmatrix}$$

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} = \sum_{k=1}^{n} a_{ik}b_{kj}, \qquad \text{for } i = 1, ..., m \text{ and } j = 1, ..., p.$$

Identity matrix

A square matrix with 1's on the diagonal and 0's elsewhere

$$\mathbf{I}_n = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix},$$

1.6 Prerequisite III: Elements of Matrix Algebra

Orthogonality

Two vectors $\mathbf{a} = (a_1, \cdots, a_n)$ and $\mathbf{b} = (b_1, \cdots, b_n)$ are said orthogonal if $\mathbf{a}'\mathbf{b} = \sum_{i=1}^{n} a_i b_i = 0$

Inverse matrix

The inverse $\mathbf{M}^{-1}$ of a square matrix $\mathbf{M}$ is the unique matrix such that

$$\mathbf{M}^{-1}\mathbf{M} = \mathbf{I} = \mathbf{M}\mathbf{M}^{-1}$$

Determinant

$$|A| = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc.$$

$$|A| = \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a\begin{vmatrix} e & f \\ h & i \end{vmatrix} - b\begin{vmatrix} d & f \\ g & i \end{vmatrix} + c\begin{vmatrix} d & e \\ g & h \end{vmatrix}$$
$$= aei + bfg + cdh - ceg - bdi - afh.$$