Probability and statistics

Prof. Dr. Noman Islam

Introduction

- An experiment that can result in different outcomes, even though it is repeated in the same manner every time, is called a random experiment.
- The set of all possible outcomes of a random experiment is called the sample space of the experiment. The sample space is denoted as S.
- An event is a subset of the sample space of a random experiment.

- Two events, denoted as E1 and E2, such that E1 2 ∩ =∅ E are said to be mutually exclusive
- Probability is used to quantify the likelihood, or chance, that an outcome of a random experiment will occur

Probability

- Probability is a number that is assigned to each member of a collection of events from a random experiment that satisfies the following properties: If S is the sample space and E is any event in a random experiment,
- $P(S) = 1$
- $0 \leq P(E) \leq 1$

Conditional probability

- P(B/A)
- The conditional probability of an event B given an event A, denoted as P(B|A), is
- $P(B|A) = P(A \cap B)/ P(A)$

Bayesian theorem

- $P(A \mid B) = P(B \mid A) * P(A) / P(B)$

Random variable

- A random variable is a function that assigns a real number to each outcome in the sample space of a random experiment.
- A discrete random variable is a random variable with a finite (or countably infinite) range.
- A continuous random variable is a random variable with an interval (either finite or infinite) of real numbers for its range.

- Examples of continuous random variables: electrical current, length, pressure, temperature, time, voltage, weight
- Examples of discrete random variables: number of scratches on a surface, proportion of defective parts among 1000 tested, number of transmitted bits received in error

Example

- A voice communication system for a business contains 48 external lines.
- At a particular time, the system is observed, and some of the lines are being used.
- Let the random variable X denote the number of lines in use.
- Then X can assume any of the integer values 0 through 48.
- When the system is observed, if 10 lines are in use, x = 10.

- Random variables are so important in random experiments that sometimes we essentially ignore the original sample space of the experiment and focus on the probability distribution of the random variable

FIGURE 4-8 Continuous uniform probability density function.

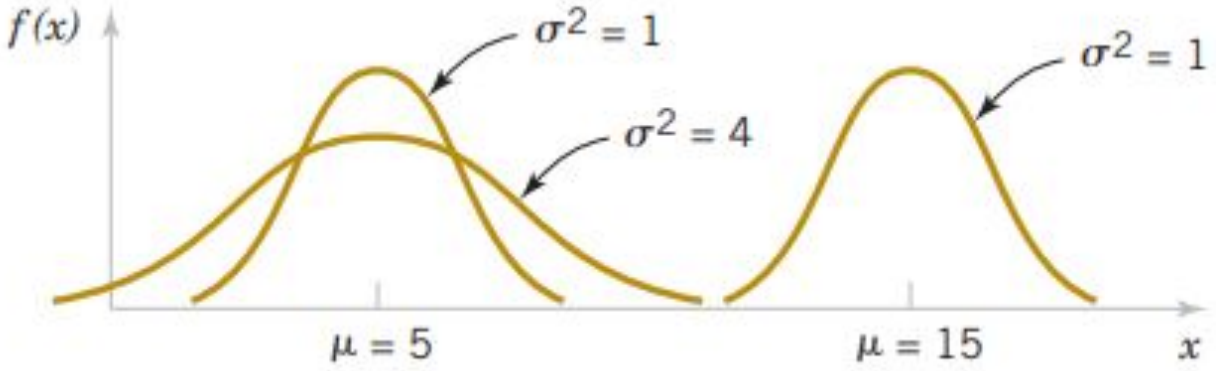


FIGURE 4-10 Normal probability density functions for selected values of the parameters μ and σ^2.

- A normal random variable with μ σ = = 0 1 2 and is called a standard normal random variable and is denoted as Z

- Mean
- Median
- Mode
- Covariance
- Correlation

Stem and Leaf plot

TABLE • 6-2 Compressive Strength (in psi) of 80 Aluminum-Lithium Alloy Specimens

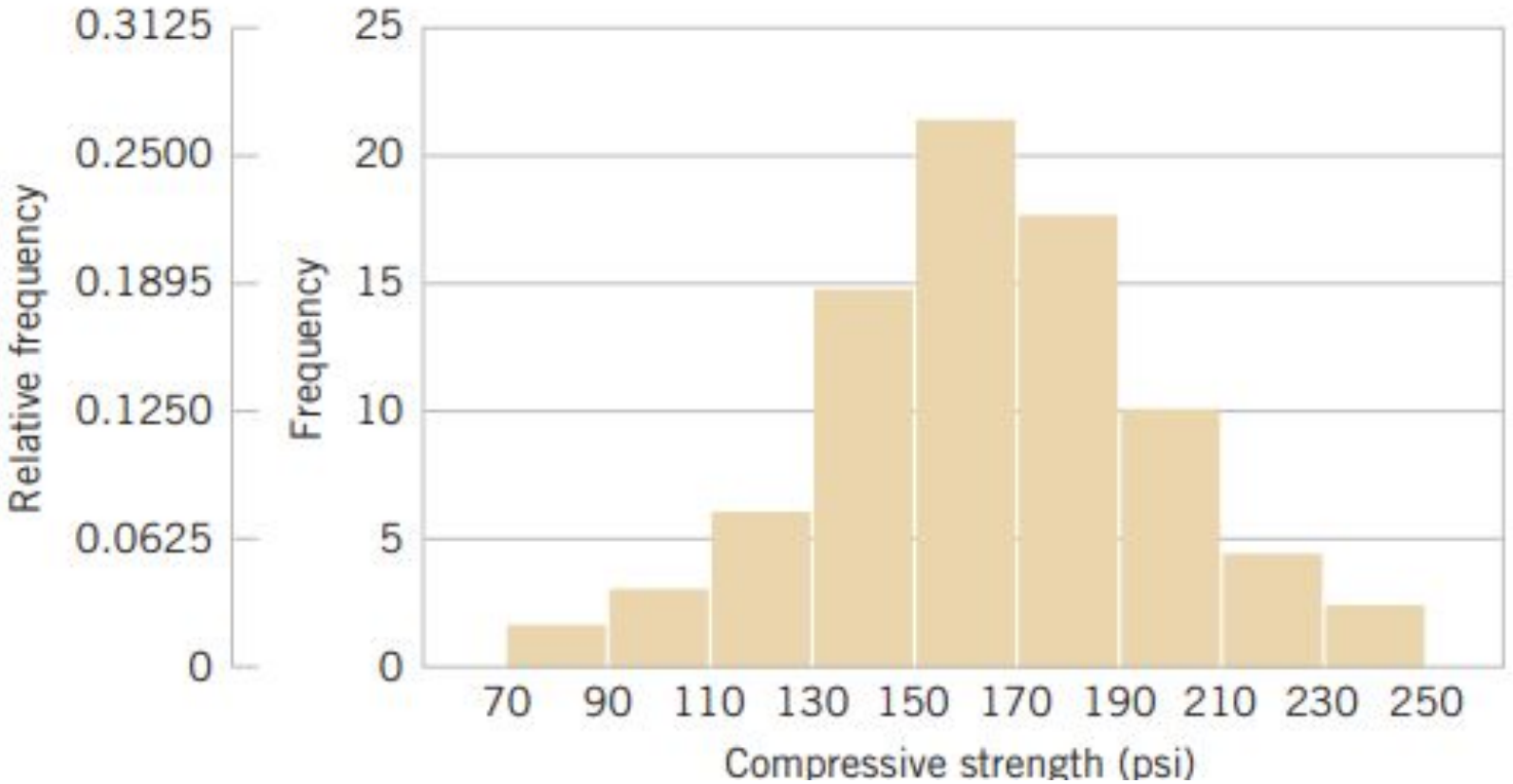
105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

Stem	Leaf	Frequency
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

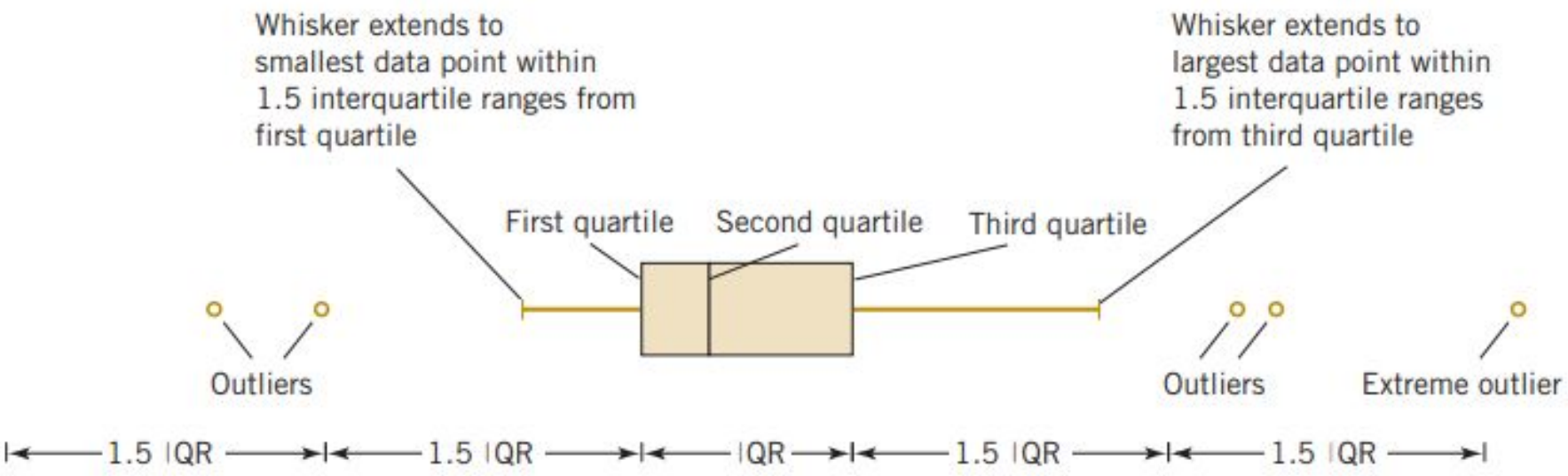
Stem: Tens and hundreds digits (psi); Leaf: Ones digits (psi).

FIGURE 6-4 Stem-and-leaf diagram for the compressive strength data in Table 6-2.

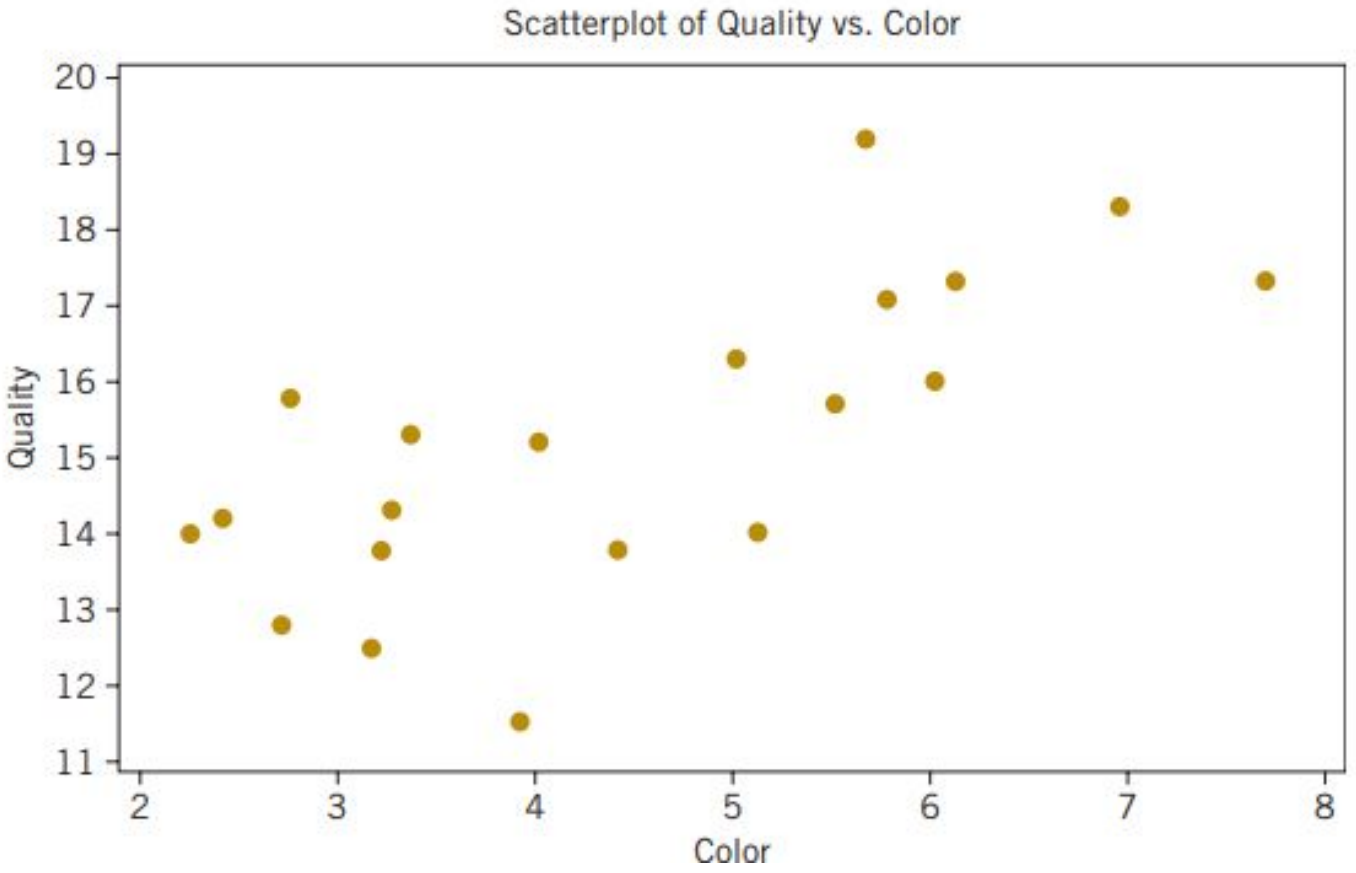
Histogram


0.3125
0.2500
0.1895
0.1250
0.0625
0
Relative frequency
25
20
15
10
5
0
Frequency
70
90
110
130
150
170
190
210
230
250
Compressive strength (psi)

Box plot


Whisker extends to smallest data point within 1.5 interquartile ranges from first quartile
Whisker extends to largest data point within 1.5 interquartile ranges from third quartile
First quartile
Second quartile
Third quartile
Outliers
Outliers
Extreme outlier
1.5 IQR
1.5 IQR
IQR
1.5 IQR
1.5 IQR

Scatter plot



Statistical inference

- Statistical methods are used to make decisions and draw conclusions about populations. This aspect of statistics is generally called statistical inference

Mean square error

- The mean squared error of an estimator θ

Empirical models

- Many problems in engineering and the sciences involve a study or analysis of the relationship between two or more variables.
- The collection of statistical tools that are used to model and explore relationships between variables that are related in a nondeterministic manner is called regression analysis.

Simple linear regression

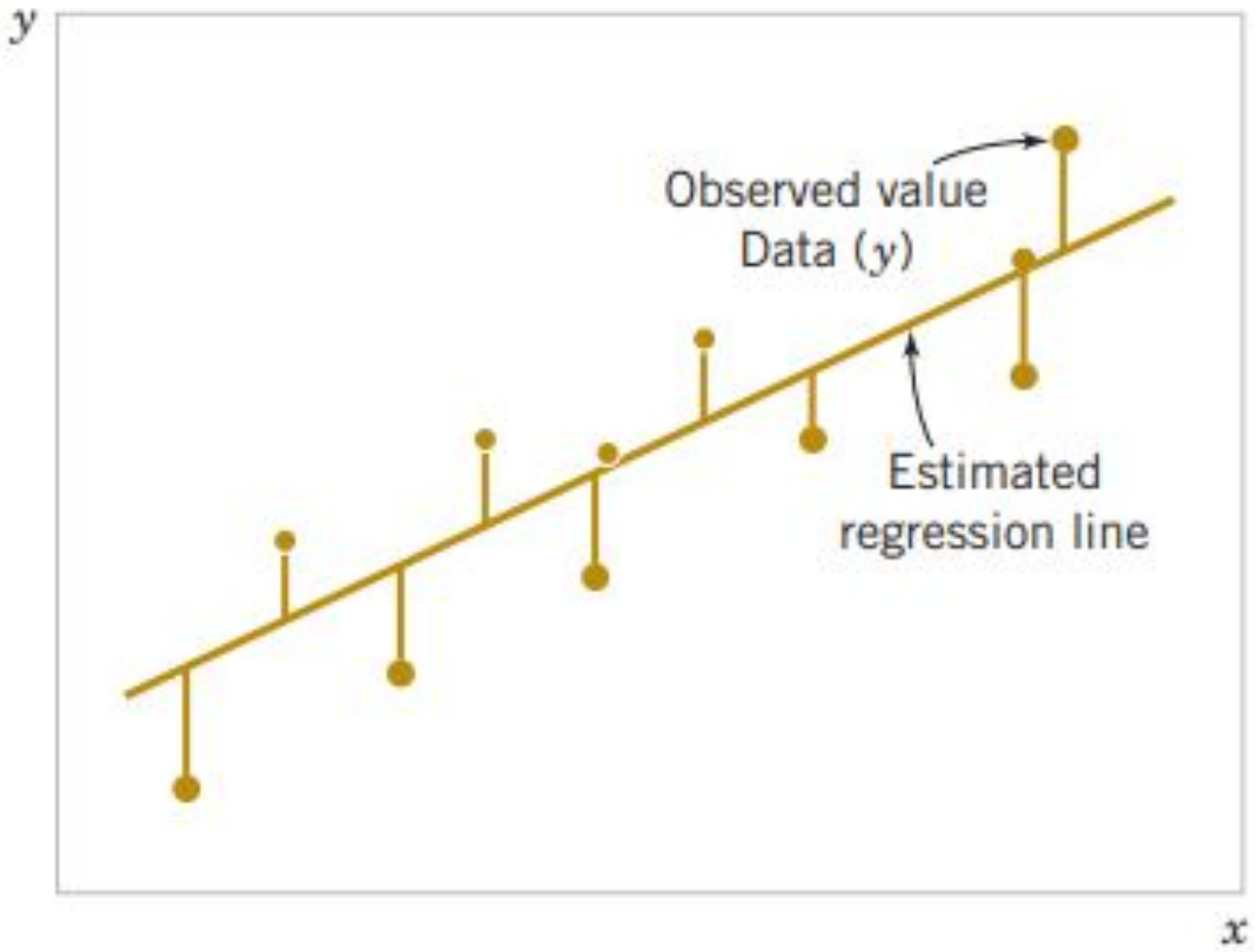
- $Y = \beta_0 + \beta_1 x + \varepsilon$
- Suppose that we have n pairs of observations The estimates of β_0 and β_1 should result in a line that is (in some sense) a "best fit" to the data.
- The German scientist Karl Gauss (1777–1855) proposed estimating the parameters β0 and β1 to minimize the sum of the squares of the vertical deviations.
- We call this criterion for estimating the regression coefficients the method of least squares.

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{11-7}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \tag{11-8}$$

where $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ and $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$.


y
Observed value
Data (y)
Estimated
regression line
x

Logistic regression

- Linear regression often works very well when the response variable is quantitative.
- We now consider the situation in which the response variable takes on only two possible values, 0 and 1.

$$E(Y) = \frac{1}{1 + \exp\left[-(\beta_0 + \beta_1 x)\right]}$$

Multiple linear regression

- $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Polynomial regression

$$Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \epsilon$$