Introduction

Frequency distribution

Numerical Summaries

Probability Plots

Summary

Chapter 6: Random sampling & Data Description

Learning objectives

1. Introduction to Statistics
2. Frequency distributions and histograms
3. Numerical distribution: measures of central tendency, measures of variation, measures of position.
4. Probability plots

WHAT IS DATA

Introduction

Frequency distribution

Numerical Summaries

Probability Plots

Summary

What is data

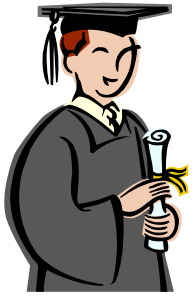
Consist of information coming from observations, counts, measurements, or responses.

Type of data

Qualitative Data

Quantitative data

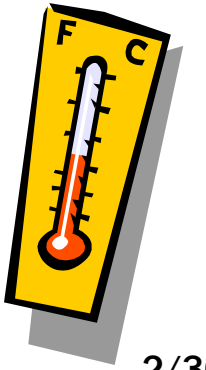
Major



Place of birth

Age

Temperature



WHAT IS STATISTICS

Introduction

Frequency distribution

Numerical Summaries

Probability Plots

Summary

What is Statistics

The science of collecting, organizing, analyzing, and interpreting data in order to make decisions.

Descriptive Statistics:
Involves organizing, summarizing, and displaying data.

e.g. Tables, charts, averages

Inferential Statistics
Involves using ***sample data*** to draw conclusions about a ***population.***

DATA COLLECTION: THE BASIC METHODS

Introduction

Frequency distribution

Numerical Summaries

Probability Plots

Summary

(1) Retrospective study using historical data

(2) Observational study

A researcher observes and measures characteristics of interest of part of a population.

(3) Experiment

A treatment is applied to part of a population and responses are observed.

DATA COLLECTION: THE BASIC METHODS

Introduction

Frequency distribution

Numerical Summaries

Probability Plots

Summary

(4) Simulation

(a) Uses a mathematical or physical model to reproduce the conditions of a situation or process.

(b) Often involves the use of computers.

(5) Survey

(a) An investigation of one or more characteristics of a population.

(b) Commonly done by interview, mail, or telephone.

Introduction

Frequency distribution

Numerical Summaries

Probability Plots

Summary

Frequency Distribution

(1) The **frequency distribution** is a summary table in which the data are arranged into numerically ordered class groupings.

(2) You must give attention to selecting the appropriate *number* of **class groupings** for the table, determining a suitable *width* of a class grouping, and establishing the *boundaries* of each class grouping to avoid overlapping.

(3) To determine the **width of a class interval,** you divide the **range** (Highest value–Lowest value) of the data by the number of class groupings desired.

FREQUENCY DISTRIBUTION

Introduction
Frequency distribution
Numerical Summaries
Probability Plots
Summary

Example

A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature

24, 35, 17, 21, 24, 37, 26, 46, 58, 30, 32, 13, 12, 38, 41, 43, 44, 27, 53, 27

Class	Frequency	Relative Frequency
[10, 20)	3	0.15
[20, 30)	6	0.30
[30, 40)	5	0.25
[40, 50)	4	0.20
[50, 60)	2	0.10
Total	20	1.00

Introduction

Frequency distribution

Numerical Summaries

Probability Plots

Summary

Histogram

(1) A graph of the data in a frequency distribution is called a **histogram.**

(2) The **class boundaries** (or **class midpoints**) are shown on the horizontal axis.

(3) The vertical axis is either **frequency, relative frequency,** or **percentage**.

(4) Bars of the appropriate heights are used to represent the number of observations within each class.

FREQUENCY DISTRIBUTION

Introduction

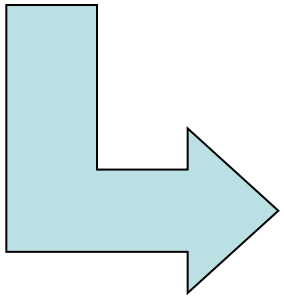
Frequency distribution

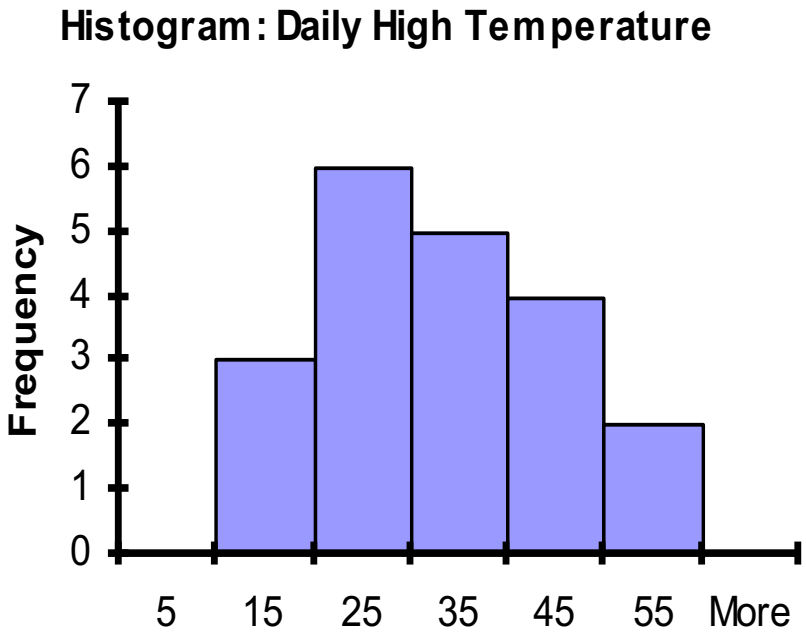
Numerical Summaries

Probability Plots

Summary

Class	Frequency	Relative Frequency
[10, 20)	3	0.15
[20, 30)	6	0.30
[30, 40)	5	0.25
[40, 50)	4	0.20
[50, 60)	2	0.10
Total	20	1.00





COMMON DISTRIBUTION SHAPES

Introduction

Frequency distribution

Numerical Summaries

Probability Plots

Summary

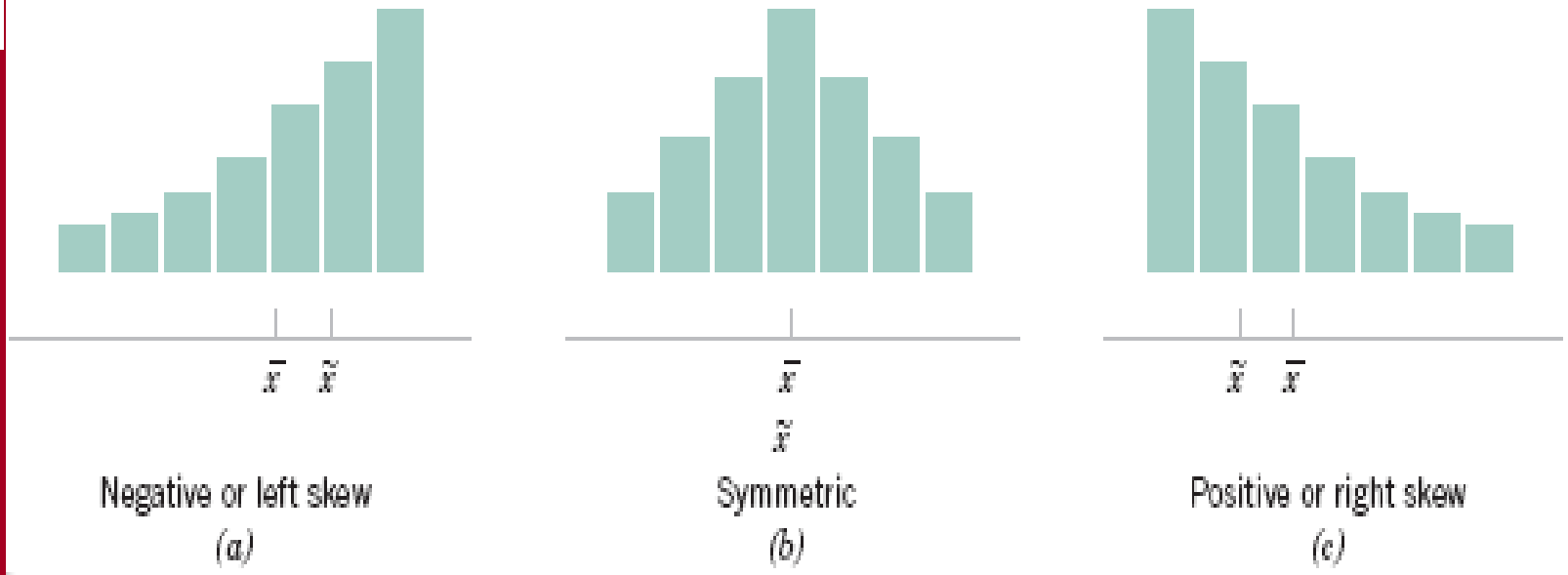


Figure 6-11 Histograms for symmetric and skewed distributions.

COMMON DISTRIBUTION SHAPES

Introduction

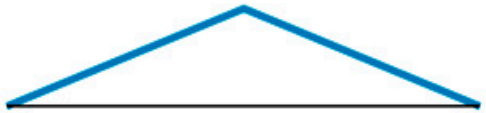
Frequency distribution

Numerical Summaries

Probability Plots

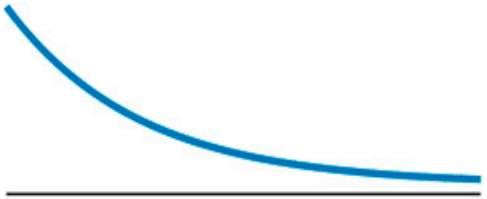
Summary

(a) Bell-shaped

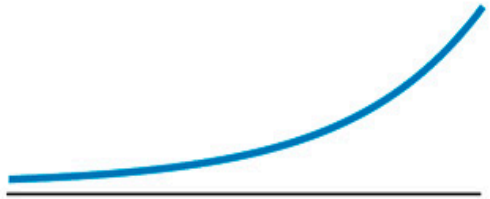


(b) Triangular

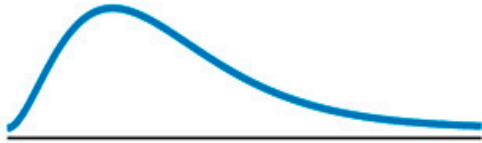
(c) Uniform (or rectangular)



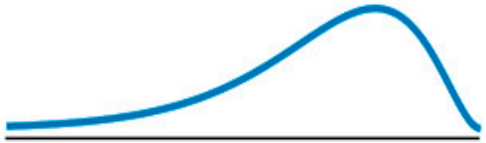
(d) Reverse J-shaped



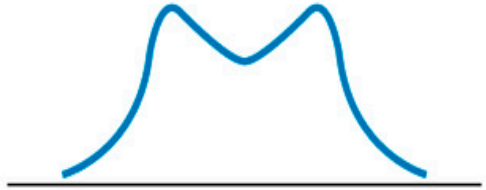
(e) J-shaped



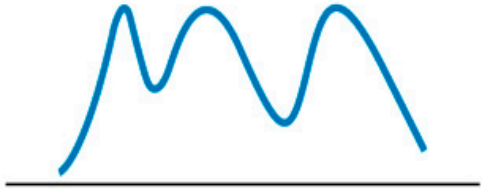
(f) Right skewed



(g) Left skewed



(h) Bimodal



(i) Multimodal

MEASURES OF CENTRAL TENDENCY

Introduction
Frequency distribution
Numerical Summaries
Probability Plots
Summary

Sample mean

If the n observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the sample mean is

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Example

Let's consider the weight of the eight observations collected from the prototype engine connectors: 12.6, 12.9, 13.4, 12.3, 13.6, 13.5, 12.6 and 13.1

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{12.6 + 12.9 + \ldots + 13.1}{8} = 13.0$$

MEASURES OF CENTRAL TENDENCY

Introduction
Frequency distribution
Numerical Summaries
Probability Plots
Summary

Sample median

(1) The value that lies in the middle of the data when the data set is ordered.

(2) Measures the center of an ordered data set by dividing it into two equal parts.

(3) If the data set has an

(a) even number of entries: median is the mean of the two middle data entries.

(b) odd number of entries: median is the middle data entry.

MEASURES OF CENTRAL TENDENCY

Introduction
Frequency distribution
Numerical Summaries
Probability Plots
Summary

Example

The prices (in dollars) for a sample of roundtrip flights from Chicago, Illinois to Cancun, Mexico are listed. Find the median of the flight prices.

872 432 397 427 388 782 397

First order the data.

388 397 397 427 432 782 872

The median price of the flights is $427.

MEASURES OF CENTRAL TENDENCY

Introduction
Frequency distribution
Numerical Summaries
Probability Plots
Summary

Sample mode

(1) The data entry that occurs with the greatest frequency.

(2) If no entry is repeated the data set has no mode.

(3) If two entries occur with the same greatest frequency, each entry is a mode (bimodal).

Example

At a political debate a sample of audience members was asked to name the political party to which they belong. Their responses are shown in the table. What is the mode of the responses?

MEASURES OF CENTRAL TENDENCY

Introduction
Frequency distribution
Numerical Summaries
Probability Plots
Summary

Political Party	Frequency, f
Democrat	34
Republican	56
Other	21
Did not respond	9

The mode is Republican (the response occurring with the greatest frequency). In this sample there were more Republicans than people of any other single affiliation.

Introduction

Frequency distribution

Numerical Summaries

Probability Plots

Summary

Sample Variance and sample standard deviation

(1) If $x_1, x_2, \ldots, x_n$, is a sample of n observations, the sample variance is

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{n-1}$$

(2) The sample standard deviation, s, is the positive square root of the sample variance.

Introduction

Frequency distribution

Numerical Summaries

Probability Plots

Summary

Computing formula for σ^2

$$s^2 = \frac{n}{n-1}\left(\frac{\sum_{i=1}^{n} x_i^2}{n} - (\bar{x})^2\right)$$

Let us return to the Example in page of this slide.

Sample Variance

$$s^2 = \frac{8}{8-1}\left(\frac{12.6^2 + \ldots + 13.1^2}{8} - 13^2\right) = 0.2286$$

Sample standard deviation

$$s = \sqrt{0.2286} = 0.48$$

Introduction

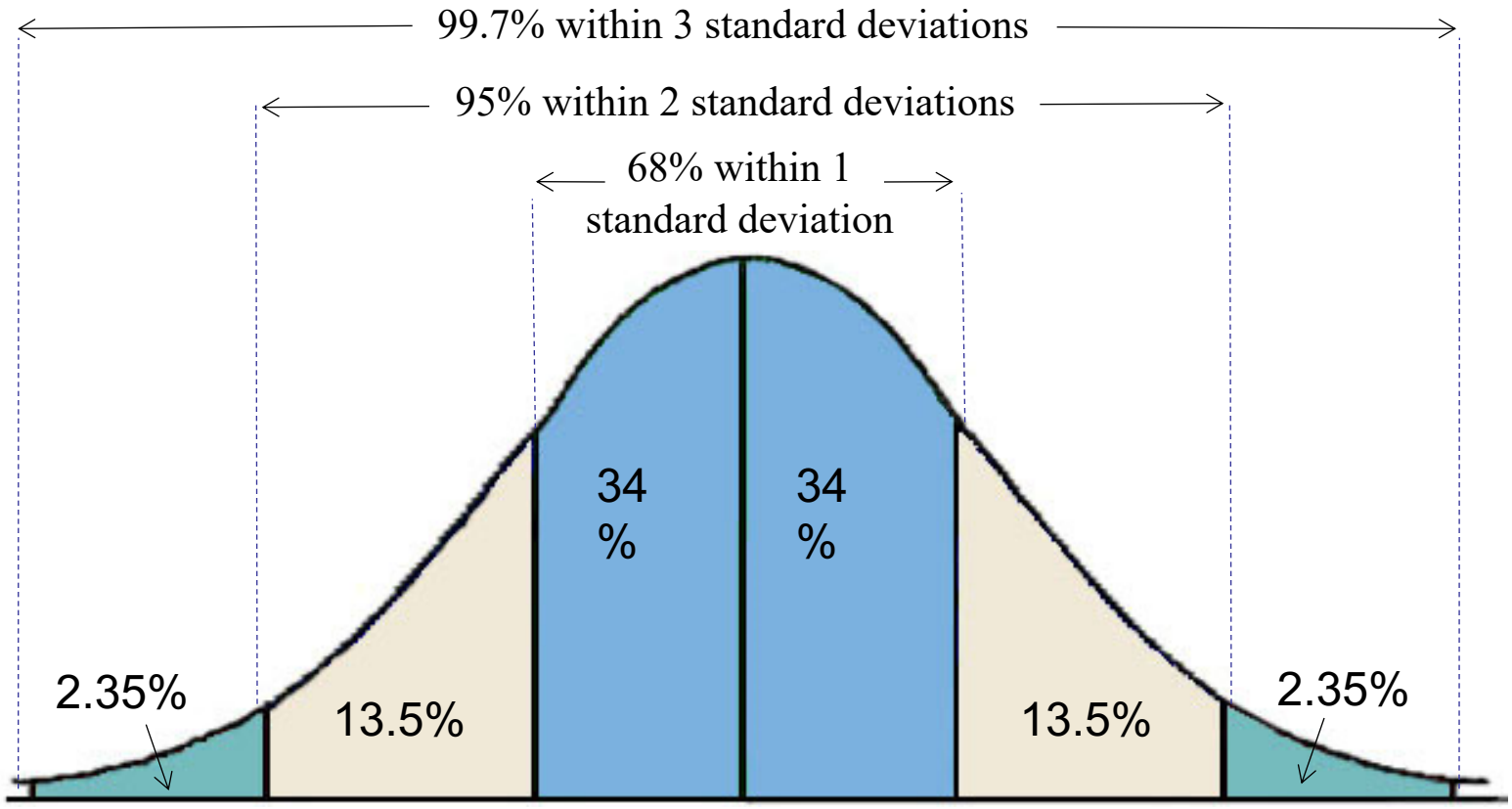
Frequency distribution

Numerical Summaries

Probability Plots

Summary

Interpreting standard deviation: For data with a bell-shaped distribution



MEASURES OF VARIATION

Introduction

Frequency distribution

Numerical Summaries

Probability Plots

Summary

Sample range

- The difference between the maximum and minimum data entries in the set.

- The data must be quantitative.

- If the n observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the sample range is

$$r = max(x_i) - min(x_i)$$

MEASURES OF POSITION: QUARTILES

Introduction
Frequency distribution
Numerical Summaries
Probability Plots
Summary

(1) **Fractiles** are numbers that partition (divide) an ordered data set into equal parts.

(2) **Quartiles** approximately divide an ordered data set into four equal parts.

(a) **First quartile, *Q*1**: About one quarter of the data fall on or below $Q1$.

(b) **Second quartile, *Q*2**: About one half of the data fall on or below $Q2$ (median).

(c) **Third quartile, *Q*3**: About three quarters of the data fall on or below $Q3$.

MEASURES OF POSITION: QUARTILES

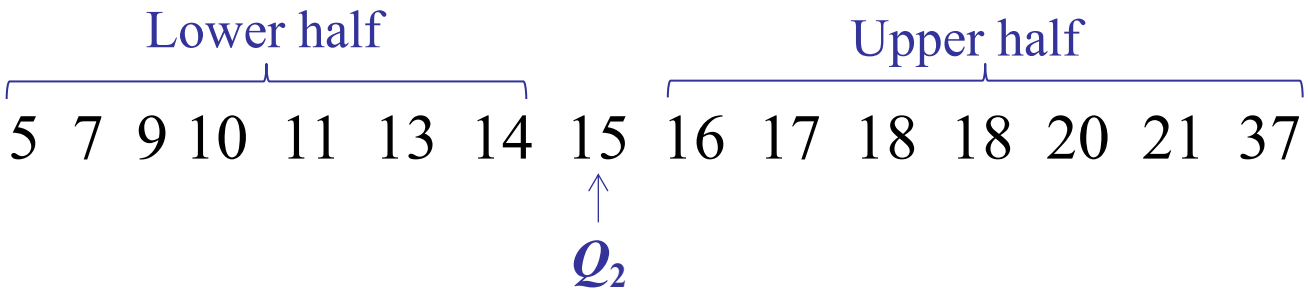
Introduction
Frequency distribution
Numerical Summaries
Probability Plots
Summary

Example

The test scores of 15 employees enrolled in a CPR training course are listed. Find the first, second, and third quartiles of the test scores.

13 9 18 15 14 21 7 10 11 20 5 18 37 16 17

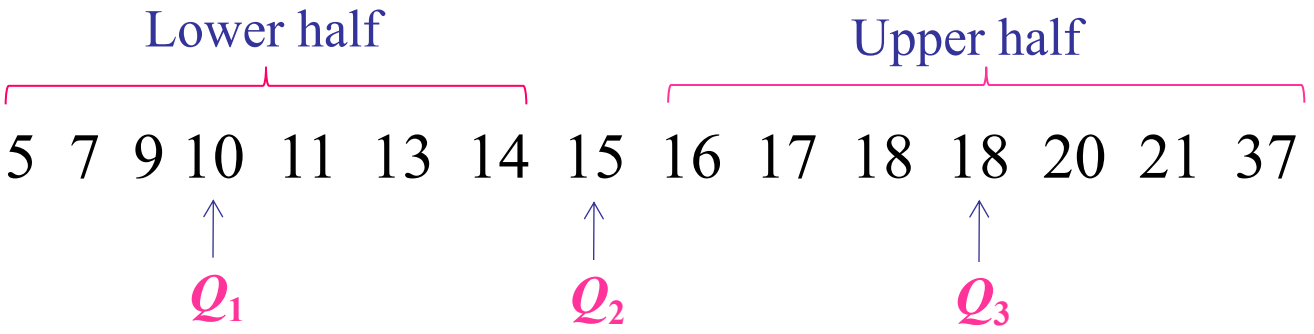
Q_2 divides the data set into two halves.



MEASURES OF POSITION: QUARTILES

Introduction
Frequency distribution
Numerical Summaries
Probability Plots
Summary

The first and third quartiles are the medians of the lower and upper halves of the data set.



About one fourth of the employees scored 10 or less, about one half scored 15 or less; and about three fourths scored 18 or less.

Introduction

Frequency distribution

Numerical Summaries

Probability Plots

Summary

Probability Plots

(1) **Probability plotting** is a graphical method for determining whether sample data conform to a hypothesized distribution based on a subjective visual examination of the data.

(2) Probability plotting typically uses special graph paper, known as **probability paper,** that has been designed for the hypothesized distribution. Probability paper is widely available for the normal, lognormal, Weibull, and various chi-square and gamma distributions.

PROBABILITY PLOTS

Introduction
Frequency distribution
Numerical Summaries
Probability Plots
Summary

Example

Ten observations on the effective service life in minutes of batteries used in a portable personal computer are as follows:

176, 191, 214, 220, 205, 192, 201, 190, 183, 185.

We hypothesize that battery life is adequately modeled by a normal distribution. To use probability plotting to investigate this hypothesis, first arrange the observations in ascending order and calculate their cumulative frequencies (j-0.5)/10 as shown in Table 6-6.

PROBABILITY PLOTS

Introduction
Frequency distribution
Numerical Summaries
Probability Plots
Summary

Table 6-6 Calculation for Constracting a Normal Probability Plot

j	$x_{(j)}$	$(j - 0.5)/10$	z_j
1	176	0.05	−1.64
2	183	0.15	−1.04
3	185	0.25	−0.67
4	190	0.35	−0.39
5	191	0.45	−0.13
6	192	0.55	0.13
7	201	0.65	0.39
8	205	0.75	0.67
9	214	0.85	1.04
10	220	0.95	1.64

Introduction

Frequency distribution

Numerical Summaries

Probability Plots

Summary

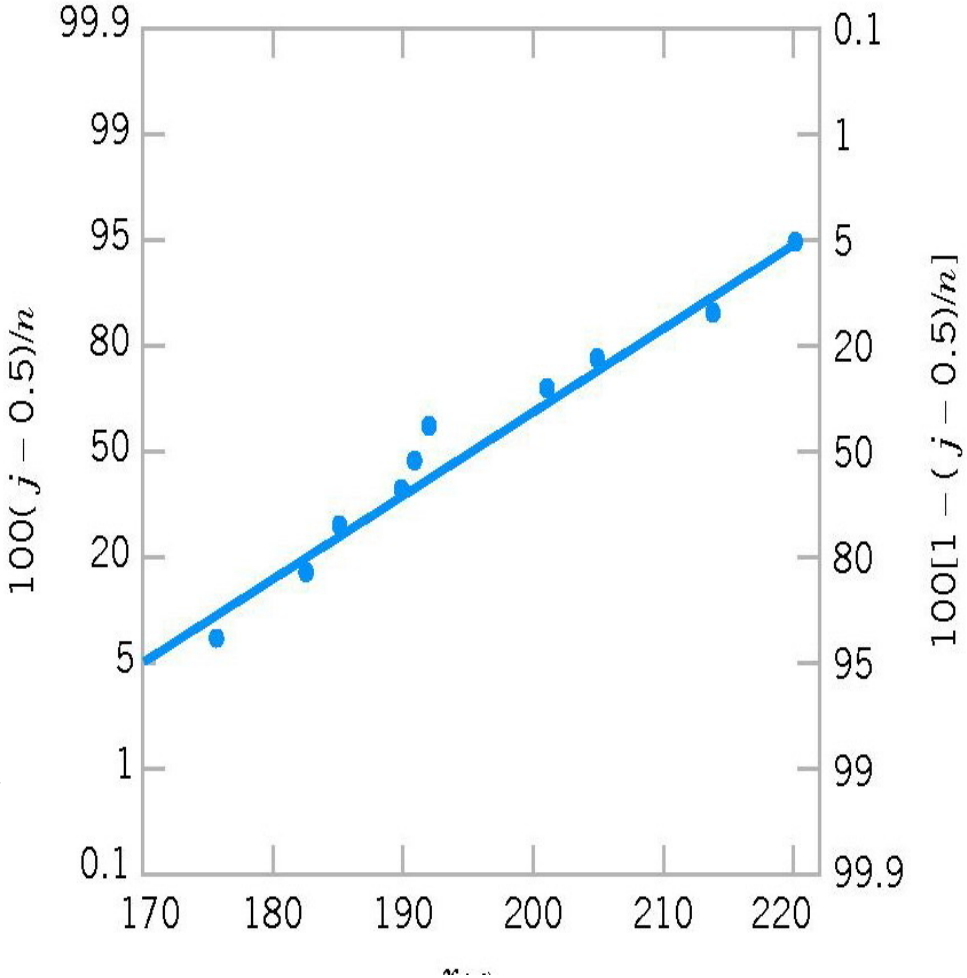


Figure 6-19 Normal probability plot for battery life.

PROBABILITY PLOTS

Introduction
Frequency distribution
Numerical Summaries
Probability Plots
Summary

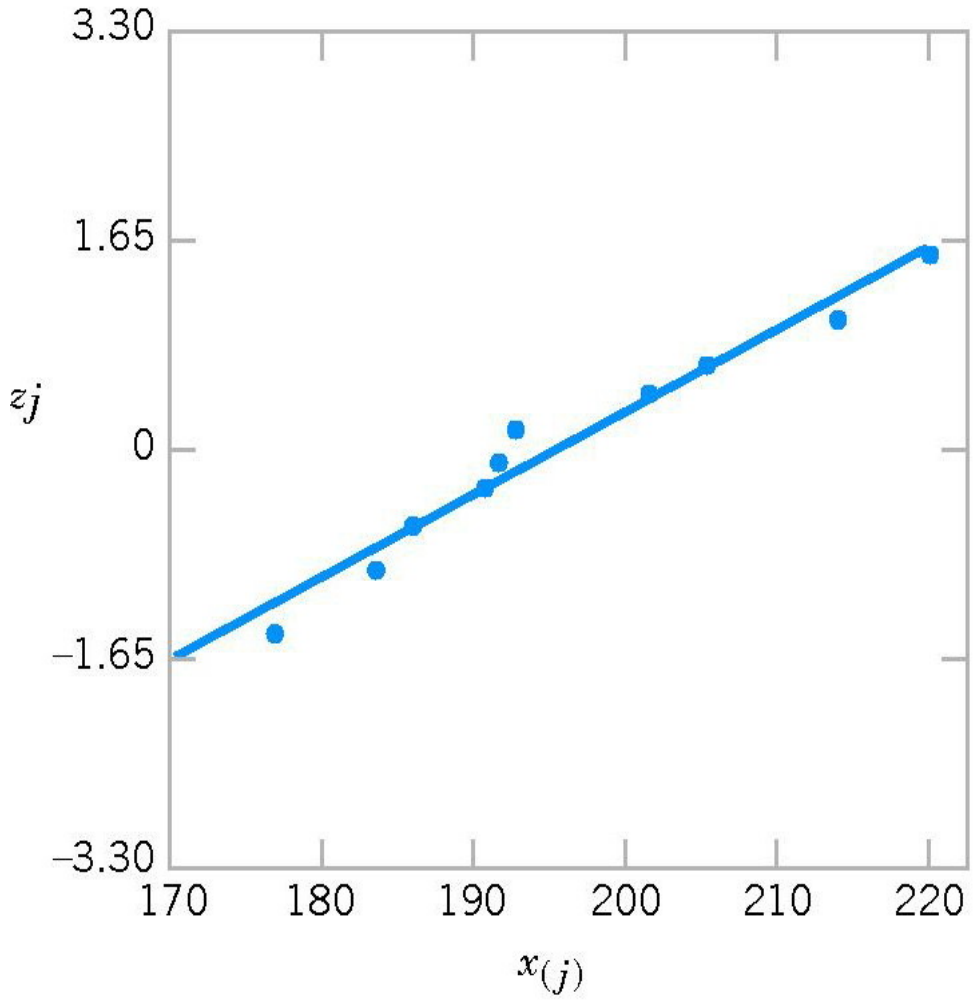


Figure 6-20
Normal probability plot obtained from standardized normal scores.

Introduction
Frequency distribution
Numerical Summaries
Probability Plots
Summary

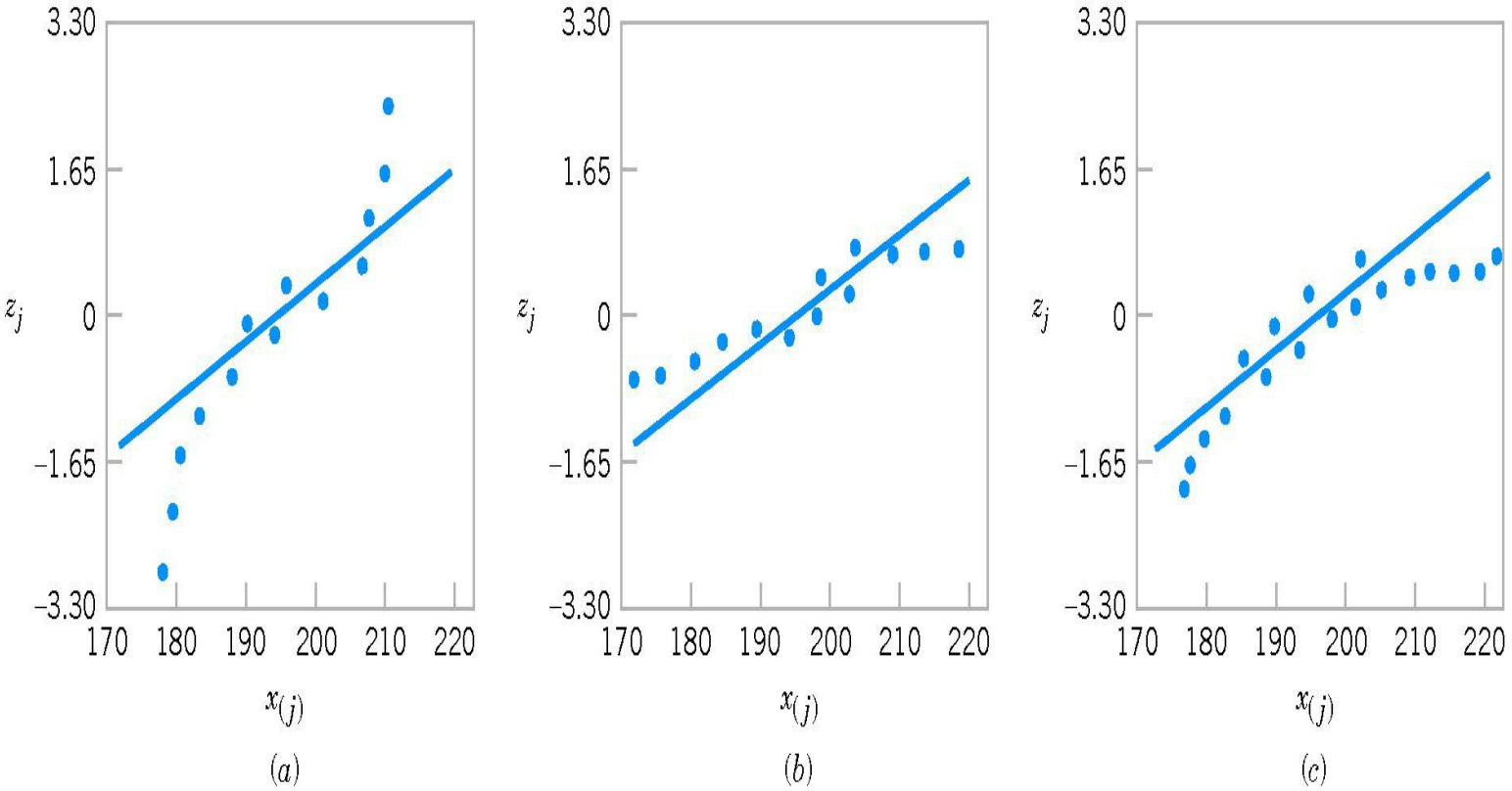


Figure 6-21 Normal probability plots indicating a non-normal distribution. (a) Light-tailed distribution. (b) Heavy-tailed distribution. (c) A distribution with positive (or right) skew.

SUMMARY

Introduction

Frequency distribution

Numerical Summaries

Probability Plots

Summary

In this chapter, we studied

1. Introduction to Statistics
2. Frequency distributions and histograms
3. Numerical distribution: measures of central tendency, measures of variation, measures of position.
4. Probability plots

Homework: Read slides of the next lecture.