Chapter 8 Fundamental Sampling Distributions and Data Distributions

吳明龍 副教授
成大資訊系/醫資所
minglong.wu@csie.ncku.edu.tw
辦公室:資訊系館 12 樓

8.1 Random Sampling

- Outcome of a statistical experiment
  - Numerical value: total value of a pair of dice tossed
  - Descriptive representation: blood types in blood test
- Sampling from distributions or populations
  - Sample mean and sample variance
- The use of high speed computer enhance the use of formal statistical inference with graphical techniques.

Random Sampling

- Definition 8.1: A population consists of the totality of the observations with which we are concerned.
  - Finite size: 600 students are classified according to blood type => a population of size 600
  - Infinite size: measuring the atmospheric pressure; some infinite populations are so large
- Each observation in a population is a value of a random variable *X* having some probability distribution *f*(*x*).
  - If one is inspecting items coming off an assembly line for defects, then each observation in population might be a value 0 or 1 of the binomial random variable X with probability distribution

    $$b(x;1,p) = p^x q^{1-x}, \qquad x = 0,1,$$

    where 0 indicates a nondefective item and 1 indicates a defective item.

Random Sampling

- Sometimes, it is impossible or impractical to observe the entire set of observations that make up the population.
- Definition 8.2: A sample is a subset of a population.
- Inference from the sample to the population are to be valid
  - Obtain representative samples
  - Bias: Erroneous inferences result from selecting convenient sampling members
  - Random sample: independent and at random

Random Sampling

- Definition 8.3: Let X_1, X_2 ,…, X_n be n independent random variables, each having the same probability distribution $f(x)$. We then define X_1, X_2, …, X_n to be a random sample of size n from the population $f(x)$ and write its joint probability distribution as

$$f(x_1, x_2, \cdots, x_n) = f(x_1) f(x_2) \cdots f(x_n).$$

  – If we assume the population of battery lives to be normal, the possible values of any random sample X_i, i = 1, 2,…, 8, will be precisely the same as those in the original population, and hence X_i has the same identical normal distribution as X.

8.2 Some Important Statistics

- Definition 8.4: Any function of the random variables constituting a random sample is called a statistic.
- Definition 8.5: If X_1, X_2 ,…, X_n represent a random sample of size n, then the sample mean is defined by the statistic

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

- Definition 8.6: If X_1, X_2 ,…, X_n represent a random sample of size n, then the sample variance is defined by the statistic

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}.$$

Some Important Statistics

- Example 8.2: A comparison of coffee prices at 4 randomly selected grocery stores in San Diego showed increases from the previous month of 12, 15, 17, and 20 cents for a 1-pound bag. Find the variance of this random sample of price increases.
  - Solution

$$\bar{x} = \frac{12+15+17+20}{4} = 16$$

$$s^2 = \frac{\sum_{i=1}^{4}(x_i - 16)^2}{4-1} = \frac{(12-16)^2 + (15-16)^2 + (17-16)^2 + (20-16)^2}{3}$$

$$= \frac{34}{3}$$

Some Important Statistics

- Theorem 8.1: If S^2 is the variance of a random sample of size n, we may write

$$S^2 = \frac{n\sum_{i=1}^{n} X_i^{\,2} - \left(\sum_{i=1}^{n} X_i\right)^2}{n(n-1)}.$$

  - Proof

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{\sum_{i=1}^{n}(X_i^2 - 2\overline{X}X_i + \overline{X}^2)}{n-1}$$

$$= \frac{\sum_{i=1}^{n} X_i^2 - 2\overline{X}\sum_{i=1}^{n} X_i + n\overline{X}^2}{n-1} = \frac{\sum_{i=1}^{n} X_i^2 - 2\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)\sum_{i=1}^{n} X_i + n\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)^2}{n-1}$$

Some Important Statistics

- (Definition 8.7): The <u>sample standard deviation</u>, denoted by *S*, is the positive square root of the sample variance.
- Example 8.3: Find the variance of the data 3, 4, 5, 6, 6, and 7, representing the number of trout caught by a random sample of 6 fishermen.
  - Solution

$$\sum_{i=1}^{6} x_i^2 = 171, \; \sum_{i=1}^{6} x_i = 31$$

$$s^2 = \frac{6\times 171 - 31^2}{6\times 5} = \frac{13}{6}$$

Skipped in 9th ed.

(8.3) Data Displays and Graphical Methods

- Motivation: Use creative displays to extract information about properties of a set.
  - The stem and leaf plots provide the viewer a look at symmetry of the data.
  - Normal probability plots and quantile plots are used to check normal distribution.
- Characterize statistical analysis as the process of drawing conclusion about system variability.
- Statistics provide single measures, whereas a graphical display adds additional information in terms of a picture.

Skipped in 9^{th} ed. p. 25

Box and Whisker Plot or Boxplot

- Box and whisker plot encloses the interquartile range of the data in a box that has median displayed within.
- Interquartile range: between the 75th percentile (upper quartile) and the 25th percentile (lower quartile).
- Boxplot provides the viewer information about outliers which represent "rare event".
- Example 8.3: Nicotine content was measured in a random sample of 40 cigarettes. The data is displayed right.
  - Mild outliers: 0.72, 0.85, and 2.55

1.09	1.92	2.31	1.79	2.28
1.74	1.47	1.97	0.85	1.24
1.58	2.03	1.70	2.17	2.55
2.11	1.86	1.90	1.68	1.51
1.64	0.72	1.69	1.85	1.82
1.79	2.46	1.88	2.08	1.67
1.37	1.93	1.40	1.64	2.09
1.75	1.63	2.37	1.75	1.69

Skipped
in 9th ed.
p. 25

Box and Whisker Plot or Boxplot

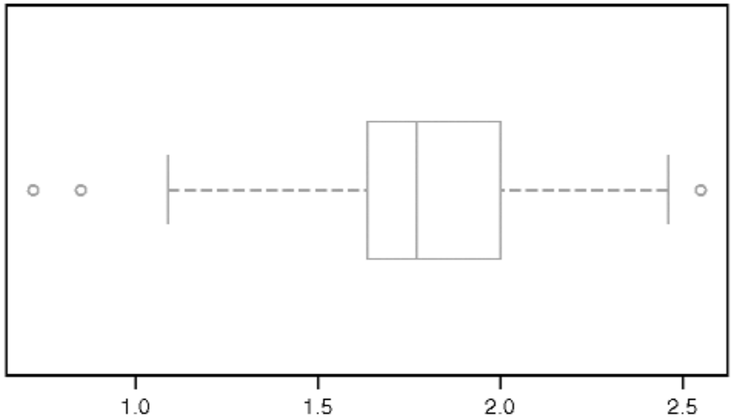


Figure 8.1: Box-and-whisker plot for nicotine data of Exercise 1.21.

```
Stem Leaf                     #
  25 5                        1
  24 6                        1
  23 17                       2
  22 8                        1
  21 17                       2
  20 389                      3
  19 0237                     4
  18 2568                     4
  17 045599                   6
  16 3447899                  7
  15 18                       2
  14 07                       2
  13 7                        1
  12 4                        1
  11
  10 9                        1
   9
   8 5                        1
   7 2                        1
     ----+----+----+----+
 Multiply Stem.Leaf by 10**-1
```

Figure 8.2: Stem-and-leaf plot for nicotine data.

Skipped in 9th ed. p. 26

Box and Whisker Plot or Boxplot

- Example 8.4: Consider the following data, consisting of 30 samples measuring the thickness of paint can ears. Figure 8.3 depicts a box and whisker plot for this asymmetric set of data.

Sample	Measurements					Sample	Measurements				
1	29	36	39	34	34	16	35	30	35	29	37
2	29	29	28	32	31	17	40	31	38	35	31
3	34	34	39	38	37	18	35	36	30	33	32
4	35	37	33	38	41	19	35	34	35	30	36
5	30	29	31	38	29	20	35	35	31	38	36
6	34	31	37	39	36	21	32	36	36	32	36
7	30	35	33	40	36	22	36	37	32	34	34
8	28	28	31	34	30	23	29	34	33	37	35
9	32	36	38	38	35	24	36	36	35	37	37
10	35	30	37	35	31	25	36	30	35	33	31
11	35	30	35	38	35	26	35	30	29	38	35
12	38	34	35	35	31	27	35	36	30	34	36
13	34	35	33	30	34	28	35	30	36	29	35
14	40	35	34	33	35	29	38	36	35	31	31
15	34	35	38	35	30	30	30	34	40	28	30

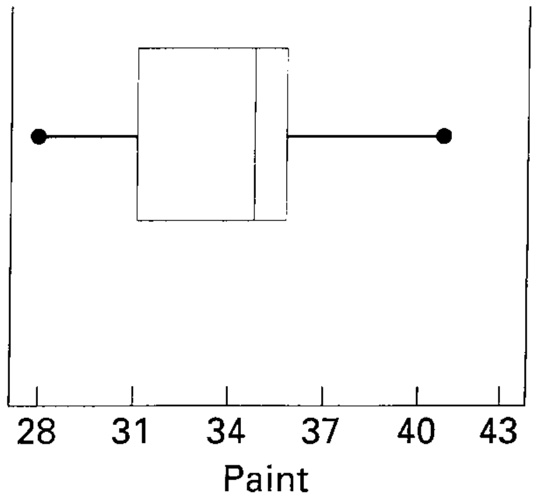


Figure 8.3 Box and whisker plot for thickness of paint can "ears."

Moved to 8.8
in 9th ed.

Quantile Plot

- Quantile plot
  - Compare samples of data
  - Draw distinctions
  - Depict cumulative distribution function
- Definition 8.6: A quantile of a sample, *q*(*f*), is a value for which a specified fraction *f* of the data values is less than or equal to *q*(*f*).
  - Sample median: *q*(0.5); 75th percentile: *q*(0.75); 25th percentile: *q*(0.25);
  - $f_i = \frac{i - \frac{3}{8}}{n + \frac{1}{4}}$, where i is the order of the observations when they are ranked from low to high.

If we denote the ranked observations as $y_{(1)} \le y_{(2)} \le \ldots \le y_{(n\text{-}1)} \le y_{(n)}$

then the quantile plot depicts a plot of $y_{(\mathrm{i})}$ against f_i

Moved to 8.8
in 9th ed.

Quantile Plot

- In Figure 8.15, quantile plot shows all observations.
  - Large clusters: slopes near zero
  - Sparse data: steeper slopes
- E.g.
  - Sparse data: 28-30
  - High density: 36-38

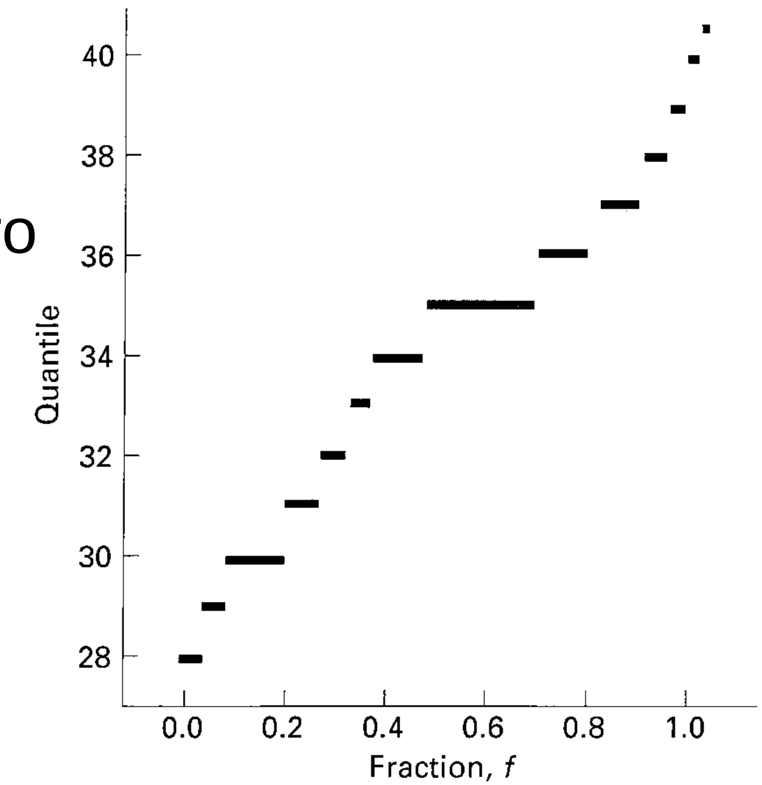


Figure 8.15 Quantile plot for paint data.

Moved to 8.8
in 9th ed.

Normal Quantile-Quantile Plot

- Approximation of quantile of normal distribution

$$q_{\mu,\sigma}(f) = \mu + \sigma\{4.91[f^{0.14} - (1-f)^{0.14}]\}$$

$$q_{0,1}(f) = 4.91[f^{0.14} - (1-f)^{0.14}]$$

- Definition 8.7: The normal quantile-quantile plot is a plot of

$y_{(i)}$ (ordered observations)

against $q_{0,1}(f_i)$, where $f_i = \dfrac{i - \frac{3}{8}}{n + \frac{1}{4}}$.

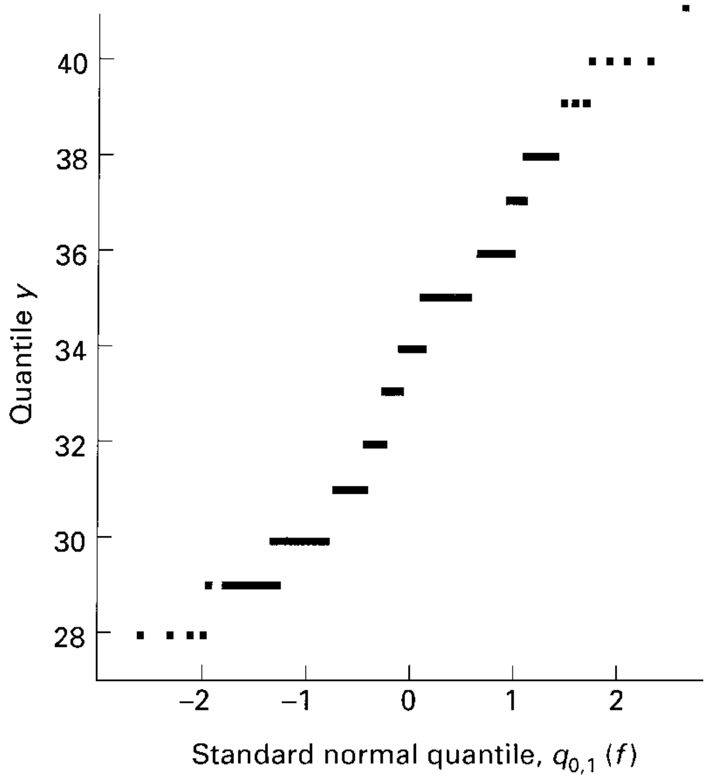
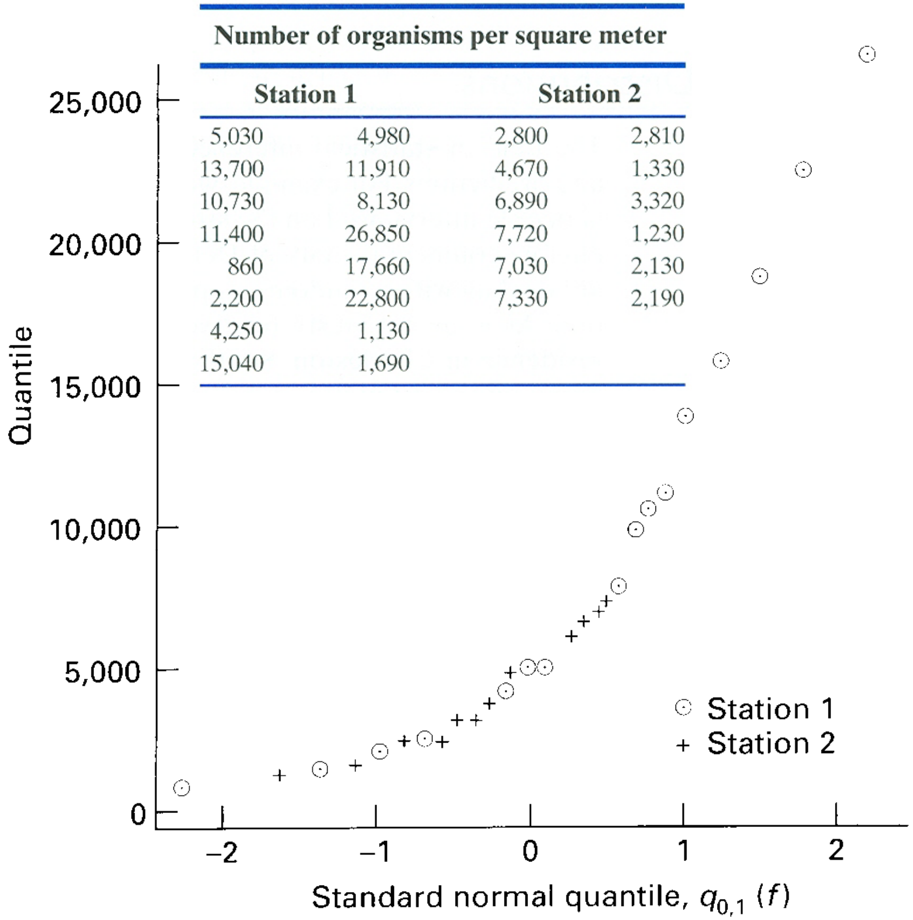


Figure 8.5 Normal quantile-quantile plot for paint can data.

Moved to 8.8
in 9th ed.

Normal Quantile-Quantile Plot

- Example 8.12: Construct a normal quantile-quantile plot and draw conclusions regarding <u>whether or not</u> it is reasonable to assume that <u>the two samples are from the same $N(\mu, \sigma)$</u> distribution.
  - Solution
    - Far from a straight line
    - Station 1 reflect a few values in the lower tail of the distribution and several in the upper tail
    - Unlikely



Number of organisms per square meter			
Station 1		Station 2	
5,030	4,980	2,800	2,810
13,700	11,910	4,670	1,330
10,730	8,130	6,890	3,320
11,400	26,850	7,720	1,230
860	17,660	7,030	2,130
2,200	22,800	7,330	2,190
4,250	1,130		
15,040	1,690		

Figure 8.6 Normal quantile-quantile plot for density data of Example 8.5.

8.3 Sampling Distribution

- Statistical inference is concerned with generalizations and predictions.
  - Based on the opinions of several people interviewed on the street, that in a forthcoming election 60% of the eligible voters in the city of Detroit favor a certain candidate.
- Definition 8.5: The probability distribution of a statistic is called a sampling distribution.
  - E.g., the probability distribution of $\overline{X}$ is called the sampling distribution of the mean.
- The sampling distribution of a statistic depends on the distribution of the population, the size of the samples, and the method of choosing the samples.

8.4 Sampling Distribution of Means

- Suppose that a random sample of *n* observations is taken from a normal population with mean μ and variance σ^2.
- By the reproductive property of the normal distribution established in Theorem 7.11

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

$$\mu_{\overline{X}} = \frac{\mu + \mu + \ldots + \mu}{n} = \mu$$

$$\sigma_{\overline{X}}^2 = \frac{\sigma^2 + \sigma^2 + \ldots + \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

•Theorem 7.11: If X_1, X_2 ,…, X_n are independent random variables having normal distributions with means μ_1, μ_2 ,…, μ_n and variances σ_1^2, σ_2^2 ,…, σ_n^2, respectively, then the random variable

$$Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n$$

has a normal distribution with mean

$$\mu_Y = a_1\mu_1 + a_2\mu_2 + \ldots + a_n\mu_n$$

and variance

$$\sigma_Y^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \ldots + a_n^2\sigma_n^2.$$

Sampling Distribution of Means

- Theorem 8.2: (**Central Limit Theorem)** If $\overline{X}$ is the mean of a random sample of size *n* taken from a population with mean μ and finite variance σ^2, then the limiting form of the distribution of

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}},$$

as $n \to \infty$, is the standard normal distribution $n(z; 0, 1)$.

  - The normal approximation for $\overline{X}$ will generally be good if $n \geq 30$.
  - If $n < 30$, the approximation is good only if the population is not too different from a <u>normal distribution</u>.
  - If the population is known to be normal, the sampling distribution of $\overline{X}$ will follow a <u>normal distribution</u> exactly, no matter how small the size of the samples.

Sampling Distribution of Means

- Example 8.4: An electric firm manufactures light bulbs that have life mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.
  - Solution

The sampling distribution of $\overline{X}$ will be approximately normal,

with $\mu_{\overline{X}} = 800, \sigma_{\overline{X}} = 40/\sqrt{16} = 10$

$$\bar{x} = 775 \Rightarrow z = \frac{775 - 800}{10} = -2.5$$

$$\therefore P(\overline{X} < 775) = P(Z < -2.5) = 0.0062$$

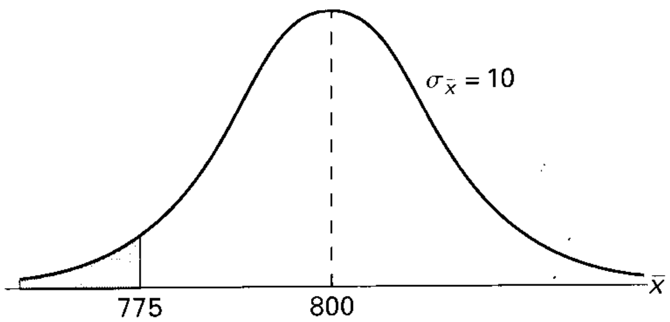


Figure 8.2(8.8): Area for Example 8.4(8.6)

Sampling Distribution of Means

- Case Study 8.1: A engineer conjectures that the population mean of a certain component parts is 5.0 millimeters. An experiment is conducted in which 100 parts produced by the process are selected randomly and the diameter measured on each. It is known that the population standard deviation $\sigma = 0.1$. The experiment indicates a sample average diameter $\bar{x}$ = 5.027 millimeters. Does this sample information appear to support or refute the engineer's conjecture?
  - Solution

$$P[|(\bar{X}-5)| \geq 0.027] = P[(\bar{X}-5) \geq 0.027] + P[(\bar{X}-5) \leq -0.027]$$

$$= 2P\left(\frac{\bar{X}-5}{0.1/\sqrt{100}} \geq 2.7\right)$$

$$= 2P(Z \geq 2.7) = 2 \times 0.0035 = 0.007.$$

$\therefore$ Strongly refutes the conjecture!

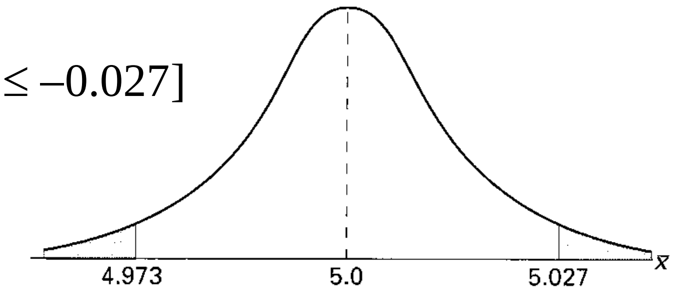


Figure 8.3(8.9): Area for Case Study 8.1(8.7)

Sampling Distribution of the Difference Between Two Averages

- Theorem 8.3: If independent sample of size n_1 and n_2 are drawn at random from two populations, discrete or continuous, with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, then the sampling distribution of the differences of means, $\overline{X}_1 - \overline{X}_2$, is approximately normally distributed with mean and variance given by

$$\mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2 \text{ and } \sigma^2_{\overline{X}_1 - \overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

$$\text{Hence} \qquad Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2 / n_1) + (\sigma_2^2 / n_2)}}$$

is approximately a standard normal variable.

•Theorem 7.11: If X_1, X_2 ,…, X_n are independent random variables having normal distributions with means μ_1, μ_2 ,…, μ_n and variances σ_1^2, σ_2^2 ,…, σ_n^2, respectively, then the random variable

$$Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n$$

has a normal distribution with mean

$$\mu_Y = a_1\mu_1 + a_2\mu_2 + \ldots + a_n\mu_n$$

and variance

$$\sigma_Y^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \ldots + a_n^2\sigma_n^2.$$

Sampling Distribution of the Difference Between Two Averages

- Case Study 8.2: Two independent experiments are being run in which two different types of paints are compared. Eighteen specimens are painted using type *A* and the drying time in hours is recorded on each. The same is done with type *B*. The population standard deviations are both known to be 1.0. Assuming that the mean drying time is equal for the two types of paint, find $P(\overline{X}_A - \overline{X}_B > 1.0)$, where $\overline{X}_A$ and $\overline{X}_B$ are average drying times for samples of size $n_A = n_B = 18$.
  - Solution

$$\mu_{\overline{X}_A - \overline{X}_B} = \mu_A - \mu_B = 0 \text{ and } \sigma^2_{\overline{X}_A - \overline{X}_B} = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B} = \frac{1}{18} + \frac{1}{18} = \frac{1}{9}.$$

$$z = \frac{(\overline{X}_A - \overline{X}_B) - (\mu_A - \mu_B)}{\sqrt{(\sigma_A^2 / n_A) + (\sigma_B^2 / n_B)}} = \frac{1-0}{\sqrt{1/9}} = 3.0$$

$$P(Z > 3.0) = 1 - P(Z < 3.0) = 1 - 0.9987 = 0.0013$$

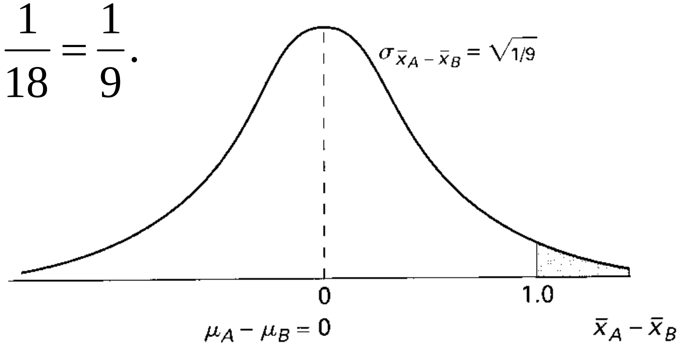


Figure 8.5(8.10): Area for Case Study 8.2(8.10)

Sampling Distribution of the Difference Between Two Averages

- Example 8.6: The television picture tubes of manufacturer *A* have a mean lifetime of 6.5 yeas and a standard deviation of 0.9 year, while those of manufacturer *B* have a mean lifetime of 6.0 years and a standard deviation of 0.8 year. What is the probability that a random sample of 36 tubes from manufacturer *A* will have a mean lifetime that is at least 1 year more than the mean lifetime of a sample of 49 tubes from manufacturer *B*?
  - Solution

$$\mu_{\bar{X}_1-\bar{X}_2} = 6.5 - 6.0 = 0.5 \text{ and } \sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{0.9^2}{36} + \frac{0.8^2}{49}} = 0.189.$$

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2 / n_1) + (\sigma_2^2 / n_2)}} = \frac{1 - 0.5}{0.189} = 2.65$$

$$P(\bar{X}_1 - \bar{X}_2 \geq 1.0) = P(Z > 2.65)$$

$$= 1 - P(Z < 2.65) = 1 - 0.9960 = 0.004$$

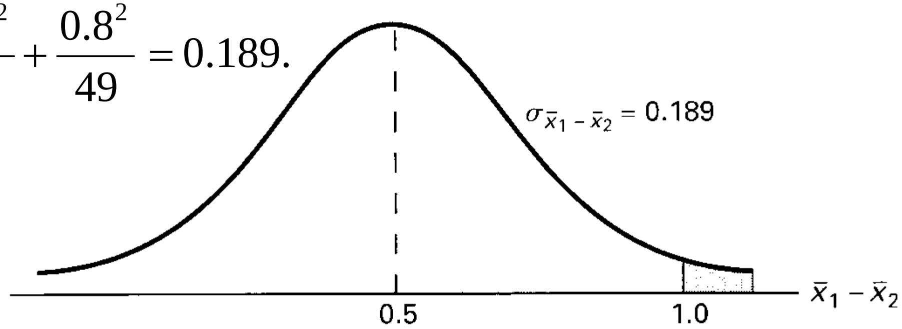


Figure 8.6(8.11): Area for Example 8.6(8.9)