

# Contents

<b>1</b>	<b>Discrete random variable</b>	<b>4</b>
1.1	Discrete Probability Distributions . . . . .	4
1.2	Some Discrete Probability Distributions . . . . .	6
1.2.1	Discrete Uniform Random Variable . . . . .	6
1.2.2	Binomial Distribution . . . . .	6
1.3	Hypergeometric Distribution . . . . .	8
1.3.1	Poisson Distribution . . . . .	9
<b>2</b>	<b>Continuous random variable</b>	<b>10</b>
2.1	Probability density function . . . . .	10
2.2	Some Continuous Probability Distributions . . . . .	12
2.2.1	Continuous Uniform Distribution . . . . .	12
2.2.2	Normal Distribution . . . . .	12
2.2.3	Exponential Distribution . . . . .	15
<b>3</b>	<b>Fundamental Sampling Distributions</b>	<b>17</b>
3.1	Random sampling . . . . .	17
3.2	Some important statistics . . . . .	17
3.3	Sampling Distributions . . . . .	18
3.4	Sampling Distribution of Means and the Central Limit . . . . .	19
3.5	Sampling Distribution of the Difference between Two Means . . . . .	21
3.6	Sampling Distribution of $S^2$ . . . . .	23

3.7	t-Distribution . . . . .	25
3.8	F-Distribution . . . . .	28
3.8.1	The F-Distribution with Two Sample Variances . . . . .	28
3.8.2	Example . . . . .	29
3.9	Sampling Distribution of Proportions and the Central Limit . . . . .	29
3.10	Sampling Distribution of the Difference between Two Proportions . . . . .	30
<b>4</b>	<b>One and Two-Sample Estimation Problems</b>	<b>32</b>
4.1	One- and Two-Sample Estimation Problems . . . . .	32
4.1.1	Introduction . . . . .	32
4.1.2	Classical Methods of Estimation . . . . .	32
4.2	Single Sample: Estimating the Mean . . . . .	34
4.3	Standard Error of a Point Estimate . . . . .	37
4.4	Two Samples: Estimating the Difference between Two Means . . . . .	37
4.5	Paired Observations . . . . .	39
4.6	Single Sample: Estimating a Proportion . . . . .	41
4.7	Two Samples: Estimating the Difference between Two Proportions . . . . .	43
4.8	Single Sample: Estimating the Variance . . . . .	44
4.9	Two Samples: Estimating the Ratio of Two Variances . . . . .	45
<b>5</b>	<b>One and Two-Sample Tests of Hypotheses</b>	<b>47</b>
5.1	Single Sample: Tests Concerning a Single Mean . . . . .	49
5.2	Two Samples: Tests on Two Means . . . . .	53
5.3	One Sample: Test on a Single Proportion . . . . .	57
5.4	Two Samples: Tests on Two Proportions . . . . .	58
5.5	One- and Two-Sample Tests Concerning Variances . . . . .	60
<b>6</b>	<b>Chi square tests</b>	<b>63</b>
6.1	Goodness-of-Fit Test . . . . .	63
6.2	Test for Independence (Categorical Data) . . . . .	66
6.3	Test for Homogeneity . . . . .	68

<b>7</b>	<b>Simple linear regression and correlation</b>	<b>71</b>
7.1	Joint Probability Distributions . . . . .	71
7.2	Means and Variances of Linear Combinations of Random Variables . . . . .	75
7.3	Correlation . . . . .	78
7.3.1	Properties of $r$ . . . . .	79
7.4	Simple linear regression . . . . .	79
7.4.1	Least Squares and the Fitted Model . . . . .	80
7.4.2	Properties of the Least Squares Estimators . . . . .	82
7.4.3	Inferences Concerning the Regression Coefficients . . . . .	82

# Chapter 1

## Discrete random variable

### 1.1 Discrete Probability Distributions

**Definition 1** *The set of ordered pairs  $(x, f(x))$  is a probability function, probability mass function, or probability distribution of the discrete random variable  $X$  if, for each possible outcome  $x$ ,*

1.  $f(x) \geq 0$ ,
2.  $\sum_x f(x) = 1$ ,
3.  $P(X = x) = f(x)$ .

**Definition 2** *The cumulative distribution function  $F(x)$  of a discrete random variable  $X$  with probability distribution  $f(x)$  is*

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t), \text{ for } -\infty < x < \infty$$

**Definition 3 (Mean of a Random Variable)** *Let  $X$  be a random variable with probability distribution  $f(x)$ . The mean, or expected value, of  $X$  is*

$$\mu = E(X) = \sum_x x f(x)$$

**Example 4** *A lot containing 7 components is sampled by a quality inspector; the lot contains 4 good components and 3 defective components. A sample of 3 is taken by the inspector. Find*

the expected value of the number of good components in this sample.

**Example 5** Let  $X$  represent the number of good components in the sample. The probability

$$\text{distribution of } X \text{ is } f(x) = \frac{\binom{4}{x} \binom{3}{3-x}}{\binom{N}{n}}, \quad x = 0, 1, 2, 3.$$

Simple calculations yield  $f(0) = 1/35$ ,  $f(1) = 12/35$ ,  $f(2) = 18/35$ , and  $f(3) = 4/35$ . Therefore,

$$\mu = E(X) = (0)\frac{1}{35} + (1)\frac{12}{35} + (2)\frac{18}{35} + (3)\frac{4}{35} = 12/7 = 1.7$$

Thus, if a sample of size 3 is selected at random over and over again from a lot of 4 good components and 3 defective components, it will contain, on average, 1.7 good components.

**Theorem 6** Let  $X$  be a random variable with probability distribution  $f(x)$ . The expected value of the random variable  $g(X)$  is

$$\mu_{g(X)} = E[g(X)] = \sum_x g(x)f(x)$$

**Example 7** Suppose that the number of cars  $X$  that pass through a car wash between 4:00 P.M. and 5:00 P.M. on any sunny Friday has the following probability distribution:

$x$	4	5	6	7	8	9
$f(x)$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{1}{6}$

Let  $g(X) = 2X - 1$  represent the amount of money, in dollars, paid to the attendant by the manager. Find the attendant's expected earnings for this particular time period.

**Example 8** Let  $X$  be a random variable with probability distribution as follows:

$x$	0	1	2	3
$f(x)$	$\frac{1}{3}$	$\frac{1}{2}$	0	$\frac{1}{6}$

Find the expected value of  $Y = (X - 1)^2$ .

**Theorem 9 (Variance of Random Variable)** Let  $X$  be a random variable with probability distribution  $f(x)$  and mean  $\mu$ . The variance of  $X$  is

$$\sigma^2 = E[(X - \mu)^2] = \sum_x (x - \mu)^2 f(x)$$

The positive square root of the variance,  $\sigma$ , is called the standard deviation of  $X$ .

**Example 10** Calculate the variance of  $g(X) = 2X + 3$ , where  $X$  is a random variable with probability distribution

$x$	0	1	2	3
$f(x)$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{2}$	$\frac{1}{8}$

## 1.2 Some Discrete Probability Distributions

### 1.2.1 Discrete Uniform Random Variable

**Definition 11 (Discrete Uniform Random Variable)** A random variable  $X$  is called discrete uniform if it has a finite number of possible values, say  $x_1, x_2, \dots, x_n$ , and  $\Pr(X = x_i) = 1/n$  for all  $i$ .

### 1.2.2 Binomial Distribution

**Definition 12 (Bernoulli Process)** Strictly speaking, the Bernoulli process must possess the following properties:

1. The experiment consists of repeated trials.
2. Each trial results in an outcome that may be classified as a success or a failure.
3. The probability of success, denoted by  $p$ , remains constant from trial to trial.
4. The repeated trials are independent.

**Definition 13 (Binomial Distribution)** A Bernoulli trial can result in a success with probability  $p$  and a failure with probability  $q = 1 - p$ . Then the probability distribution of the binomial random variable  $X$ , the number of successes in  $n$  independent trials, is

$$\Pr(X = x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, \dots, n.$$

**Example 14** The probability that a certain kind of component will survive a shock test is  $3/4$ . Find the probability that exactly 2 of the next 4 components tested survive.

**Solution 15** Let  $X$  the number of components that will survive a shock test. Assuming that the tests are independent and  $p = 3/4$  for each of the 4 tests, then  $X$  is a binomial distribution  $Bin(4, 3/4)$ . Hence,

$$\Pr(X = 2) = \binom{4}{2} (3/4)^2 (1/4)^2 \approx 0.21$$

**Example 16** The probability that a patient recovers from a rare blood disease is 0.4. If 15 people are known to have contracted this disease, what is the probability that  
 (a) at least 10 survive,  
 (b) from 3 to 8 survive,  
 and (c) exactly 5 survive?

**Solution 17** (a)  $\Pr(X \geq 10) = 1 - \Pr(X < 10)$

$$= 1 - \sum_{x=0}^9 b(x; 15, 0.4) = 1 - 0.9662 = 0.0338$$

$$(b) \Pr(3 \leq X \leq 8) = \sum_{x=3}^8 b(x; 15, 0.4)$$

$$= \sum_{x=0}^8 b(x; 15, 0.4) - \sum_{x=0}^2 b(x; 15, 0.4)$$

$$= 0.9050 - 0.0271 = 0.8779$$

$$(c) \Pr(X = 5) = b(5; 15, 0.4) = \sum_{x=0}^5 b(x; 15, 0.4)$$

$$- \sum_{x=0}^4 b(x; 15, 0.4) = 0.4032 - 0.2173 = 0.1859$$

**Theorem 18** The mean and variance of the binomial distribution  $B(n, p)$  are

$$\mu = np \text{ and } \sigma^2 = npq.$$

### 1.3 Hypergeometric Distribution

**Definition 19 (Hypergeometric Distribution)** *The probability distribution of the hypergeometric random variable  $X$ , the number of successes in a random sample of size  $n$  selected from  $N$  items of which  $K$  are labeled success and  $N - K$  labeled failure, is*

$$\Pr(X = x) = \frac{\binom{K}{x} \binom{N - K}{n - x}}{\binom{N}{n}}$$

**Theorem 20** *The mean and variance of the hypergeometric distribution  $h(N, K, n)$  are*

$$\mu = n \frac{K}{N} \text{ and } \sigma^2 = n \frac{K}{N} \left(1 - n \frac{K}{N}\right) \frac{N - n}{N - 1}.$$

**Example 21** *Lots of 40 components each are deemed unacceptable if they contain 3 or more defectives. The procedure for sampling a lot is to select 5 components at random and to reject the lot if a defective is found. What is the probability that exactly 1 defective is found in the sample if there are 3 defectives in the entire lot?*

**Solution 22** *Using the hypergeometric distribution with  $n = 5$ ,  $N = 40$ ,  $k = 3$ , and  $x = 1$ , we find the probability of obtaining 1 defective to be*

$$h(1; 40, 5, 3) = \frac{\binom{3}{1} \binom{37}{4}}{\binom{40}{5}} = 0.3011.$$

**Theorem 23 (Approximation)** *If  $n$  is small compared to  $N$ , then a binomial distribution  $B(n, p = K/N)$  can be used to approximate the hypergeometric distribution  $h(N, K, n)$ .*

**Example 24** *A manufacturer of automobile tires reports that among a shipment of 5000 sent to a local distributor, 1000 are slightly blemished. If one purchases 10 of these tires at random from the distributor, what is the probability that exactly 3 are blemished?*

**Solution 25** *Since  $N = 5000$  is large relative to the sample size  $n = 10$ , we shall approximate the desired probability by using the binomial distribution. The probability of obtaining a*



blemished tire is 0.2. Therefore, the probability of obtaining exactly 3 blemished tires is

$$h(3; 5000, 10, 1000) \approx b(3; 10, 0.2) = 0.8791 - 0.6778 = 0.2013.$$

### 1.3.1 Poisson Distribution

**Definition 26** Let  $X$  the number of outcomes occurring during a given time interval.  $X$  is called a Poisson random variable when its probability distribution is given by

$$\Pr(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \dots,$$

where  $\lambda$  is the average number of outcomes.

**Example 27** During a laboratory experiment, the average number of radioactive particles passing through a counter in 1 millisecond is 4. What is the probability that 6 particles enter the counter in a given millisecond?

**Solution 28** Using the Poisson distribution with  $x = 6$  and  $\lambda = 4$  and referring to Table A.2, we have

$$p(6; 4) = \frac{e^{-4}4^6}{6!} = 0.1042.$$

**Theorem 29** Both the mean and the variance of the Poisson distribution  $P(\lambda)$  are  $\lambda$ .

**Theorem 30 (Approximation)** Let  $X$  be a binomial random variable with probability distribution  $B(n, p)$ . When  $n$  is large ( $n \rightarrow \infty$ ), and  $p$  small ( $p \rightarrow 0$ ), then the poisson distribution can be used to approximate the binomial distribution  $B(n, p)$  by taking  $\lambda = np$ .

**Example 31** In a certain industrial facility, accidents occur infrequently. It is known that the probability of an accident on any given day is 0.005 and accidents are independent of each other.

(a) What is the probability that in any given period of 400 days there will be an accident on one day?

(b) What is the probability that there are at most three days with an accident?

**Solution 32** Let  $X$  be a binomial random variable with  $n = 400$  and  $p = 0.005$ . Thus,  $np = 2$ . Using the Poisson approximation, (a)  $\Pr(X = 1) = e^{-2}2^1 = 0.271$  and (b)  $\Pr(X \leq 3) = e^{-2}2^x/x! = 0.857$ .

## Chapter 2

# Continuous random variable

### 2.1 Probability density function

**Definition 33** The function  $f(x)$  is a probability density function (pdf) for the continuous random variable  $X$ , defined over the set of real numbers, if

1.  $f(x) \geq 0$ , for all  $x \in R$ .

2. 
$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

3. 
$$\Pr(a \leq X \leq b) = \int_a^b f(x)dx.$$

**Example 34** Suppose that the error in the reaction temperature, in  $^{\circ}C$ , for a controlled laboratory experiment is a continuous random variable  $X$  having the probability density function

$$f(x) = \begin{cases} \frac{x^2}{3}, & -1 < x < 2 \\ 0, & \text{elsewhere} \end{cases}$$

(a) Verify that  $f(x)$  is a density function.

(b) Find  $\Pr(0 \leq X \leq 1)$ .

(c) Find  $\Pr(0 < X < 1)$ .

**Definition 35** The cumulative distribution function  $F(x)$  of a continuous random variable  $X$  with density function  $f(x)$  is

$$F(x) = \Pr(X \leq x) = \int_{-\infty}^x f(t)dt, \text{ for } -\infty < x < \infty.$$

**Example 36** For the density function of Example 2, find  $F(x)$ , and use it to evaluate  $\Pr(0 < X \leq 1)$ .

**Definition 37 (Mean of a Random Variable)** Let  $X$  be a random variable with probability distribution  $f(x)$ . The mean, or expected value, of  $X$  is

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

**Example 38** For the density function of Example 2, find  $E(X)$ .

**Theorem 39** Let  $X$  be a random variable with probability distribution  $f(x)$ . The expected value of the random variable  $g(X)$  is

$$\mu_{g(X)} = E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

**Theorem 40 (Variance of Random Variable)** Let  $X$  be a random variable with probability distribution  $f(x)$  and mean  $\mu$ . The variance of  $X$  is

$$\sigma^2 = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)$$

**Theorem 41** Let  $X$  a random variable. The variance of a random variable  $X$  is

$$\sigma^2 = E(X^2) - E(X)^2.$$

**Theorem 42** Let  $X$  a random variable. If  $a$  and  $b$  are constants, then  $E(aX + b) = aE(X) + b$ .

**Theorem 43** *The expected value of the sum or difference of two or more functions of a random variable  $X$  is the sum or difference of the expected values of the functions. That is,*

$$E[g(X) \pm h(X)] = E[g(X)] \pm E[h(X)].$$

## 2.2 Some Continuous Probability Distributions

### 2.2.1 Continuous Uniform Distribution

**Definition 44 (Uniform Distribution)** *The density function of the continuous uniform random variable  $X$  on the interval  $[a, b]$  is*

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0 & \text{elsewhere.} \end{cases}$$

**Example 45** *Suppose that a large conference room at a certain company can be reserved for no more than 4 hours. Both long and short conferences occur quite often. In fact, it can be assumed that the length  $X$  of a conference has a uniform distribution on the interval  $[0, 4]$ .*

- a) *What is the probability density function?*
- b) *What is the probability that any given conference lasts at least 3 hours?*

**Theorem 46** *The mean and variance of the uniform distribution are*

$$\mu = E(X) = \frac{a+b}{2} \text{ and } \sigma^2 = \frac{(b-a)^2}{12}$$

The proofs of the theorems are left to the reader.

### 2.2.2 Normal Distribution

**Definition 47 (Standard Normal Distribution)** *The density of the standard normal distribution  $Z$  is*

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \quad -\infty < x < \infty,$$

**Theorem 48** *The mean and variance of standard normal distribution are 0 and 1, respectively. We denote the standard normal distribution by  $N(0, 1)$ .*

**Example 49** *Given a standard normal distribution  $N(0, 1)$ , find the area under the curve that lies*

(a) *to the right of  $z = 1.84$*

(b) *between  $z = -1.97$  and  $z = 0.86$ .*

**Solution 50** (a) 0.0329 (b) 0.7807.

**Example 51** *Given a standard normal distribution  $N(0, 1)$ , find the value of  $k$  such that*

(a)  $\Pr(Z > k) = 0.3015$  and

(b)  $P(k < Z < -0.18) = 0.4197$ .

**Solution 52** (a)  $k = 0.52$  (b)  $k = -2.37$ .

**Definition 53 (Normal Distribution)** *The density of the normal random variable  $X$ , with mean  $\mu$  and variance  $\sigma^2$ , and denoted by  $N(\mu, \sigma)$ , is*

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

where  $\pi = 3.14159 \dots$  and  $e = 2.71828 \dots$

**Theorem 54** *If  $X$  is normal random variable  $N(\mu, \sigma)$ , then the random variable  $(X - \mu)/\sigma$  is a standard normal distribution  $Z$  with mean 0 and variance 1.*

**Example 55** *Given a random variable  $X$  having a normal distribution with  $\mu = 50$  and  $\sigma = 10$ , find the probability that  $X$  assumes a value between 45 and 62.*

**Solution 56** *Using Table A.3, we have*

$$\begin{aligned} \Pr(45 < X < 62) &= \Pr(-0.5 < Z < 1.2) = \Pr(Z < 1.2) - \Pr(Z < -0.5) \\ &= 0.8849 - 0.3085 = 0.5764. \end{aligned}$$

**Example 57** Given a normal distribution with  $\mu = 40$  and  $\sigma = 6$ , find the value of  $x$  that has

(a) 45% of the area to the left

(b) 14% of the area to the right.

**Solution 58** (a) We need to find a  $z$  value that leaves an area of 0.45 to the left. From Table A.3 we find  $\Pr(Z < -0.13) = 0.45$ , so the desired  $z$  value is  $-0.13$ . Hence,  $x = (6)(-0.13) + 40 = 39.22$ . (b) This time we require a  $z$  value that leaves 0.14 of the area to the right and hence an area of 0.86 to the left. Again, from Table A.3, we find  $P(Z < 1.08) = 0.86$ , so the desired  $z$  value is 1.08 and

$$x = (6)(1.08) + 40 = 46.48.$$

**Example 59** An electrical firm manufactures light bulbs that have a life, before burn-out, that is normally distributed with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a bulb burns between 778 and 834 hours.

**Solution 60** The  $z$  values corresponding to  $x_1 = 778$  and  $x_2 = 834$  are

$$z_1 = \frac{778 - 800}{40} = -0.55 \text{ and } z_2 = \frac{834 - 800}{40} = 0.85.$$

Hence,

$$\begin{aligned} \Pr(778 < X < 834) &= P(-0.55 < Z < 0.85) \\ &= P(Z < 0.85) - P(Z < -0.55) \\ &= 0.8023 - 0.2912 = 0.5111. \end{aligned}$$

**Example 61** A certain machine makes electrical resistors having a mean resistance of 40 ohms and a standard deviation of 2 ohms. Assuming that the resistance follows a normal distribution and can be measured to any degree of accuracy, what percentage of resistors will have a resistance exceeding 43 ohms?

**Solution 62** A percentage is found by multiplying the relative frequency by 100%. Since the relative frequency for an interval is equal to the probability of a value falling in the interval, we must find the area to the right of  $x = 43$ . This can be done by transforming  $x = 43$  to the

corresponding  $z$  value, obtaining the area to the left of  $z$  from Table A.3, and then subtracting this area from 1. We find

$$z = \frac{43 - 40}{2} = 1.5.$$

Therefore,  $\Pr(X > 43) = \Pr(Z > 1.5) = 1 - \Pr(Z < 1.5) = 1 - 0.9332 = 0.0668$ . Hence, 6.68% of the resistors will have a resistance exceeding 43 ohms.

### 2.2.3 Exponential Distribution

The exponential random variable is used when we are interested in the time of the first arrival or the time between arrival.

**Definition 63** The continuous random variable  $X$  has an exponential distribution, with parameter  $\lambda$ , if its density function is given by  $f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0 & \text{elsewhere} \end{cases}$  where  $\lambda > 0$ .

**Theorem 64** The mean and variance of the exponential distribution are  $\mu = 1/\lambda$  and  $\sigma^2 = 1/\lambda^2$ .

If  $X$  is the time of arrival of the first customer and if the average time is 30 minutes, then  $\lambda = 1/30$ .

**Example 65** Suppose that a system contains a certain type of component whose time, in years, to failure is given by  $T$ . The random variable  $T$  is modeled nicely by the exponential distribution with mean time to failure is 5.

(a) If one component is installed, what is the probability that it is still functioning at the end of 8 years?

(b) If 5 of these components are installed in different systems, what is the probability that at least 2 are still functioning at the end of 8 years? (Hint: use the binomial distribution).

**Solution 66** (a) The probability that a given component is still functioning after 8 years is given by

$$\Pr(T > 8) = \frac{1}{5} \int_8^{\infty} e^{-t/5} dt = e^{-8/5} \approx 0.2.$$

(b) Let  $X$  represent the number of components functioning after 8 years.  $X$  is binomial distribution  $\text{Bin}(8, 0.2)$ . Then we have

$$\begin{aligned}\Pr(X \geq 2) &= \sum_{x=2}^5 \Pr(X = x) = 1 - \left( \sum_{x=0}^1 \Pr(X = x) \right) \\ &= 1 - 0.7373 = 0.2627.\end{aligned}$$



## Chapter 3

# Fundamental Sampling Distributions

### 3.1 Random sampling

**Definition 67** *A population consists of the totality of the observations with which we are concerned*

**Definition 68** *A sample is a subset of a population.*

In the field of statistical inference, statisticians are interested in arriving at conclusions concerning a population when it is impossible or impractical to observe the entire set of observations that make up the population. Therefore, we must depend on a subset of observations from the population to help us make **inferences** concerning that same population.

**Definition 69** *A sample is a subset of a population.*

To eliminate any possibility of **bias** in the sampling procedure, it is desirable to choose a random sample in the sense that the observations are made independently and at random.

### 3.2 Some important statistics

**Definition 70** *Any function of the random variables constituting a random sample is called a statistic.*

- Sample mean:  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
- Sample median:  $\tilde{X} = \begin{cases} x(n+1)/2, & \text{if } n \text{ is odd,} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{if } n \text{ is even.} \end{cases}$
- Sample variance:  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

The computed value of  $S^2$  for a given sample is denoted by  $s^2$ .

**Theorem 71** *If  $S^2$  is the variance of a random sample of size  $n$ , we may write*

$$S^2 = \frac{1}{n-1} \left[ \sum_{i=1}^n X_i^2 - n\bar{X}^2 \right]$$

- Sample standard deviation:  $S = \sqrt{S^2}$

### 3.3 Sampling Distributions

Let us consider a soft-drink machine designed to dispense, on average, 240 milliliters per drink. A company official who computes the mean of 40 drinks obtains  $\bar{x} = 236$  milliliters. On the basis of this value, she decides that the machine is still dispensing drinks with an average content of  $\mu = 240$  milliliters. The 40 drinks represent a sample from the infinite population of possible drinks that will be dispensed by this machine. The company official made the decision that the soft-drink machine dispenses drinks with an average content of 240 milliliters, even though the sample mean was 236 milliliters, because he knows from sampling theory that, if  $\mu = 240$  milliliters, such a sample value could easily occur. In fact, if she ran similar tests, say every hour, she would expect the values of the statistic  $\bar{x}$  to fluctuate above and below  $\mu = 240$  milliliters. Only when the value of  $\bar{x}$  is **substantially** different from 240 milliliters will the company official initiate action to adjust the machine.

Since a statistic is a random variable that depends only on the observed sample, it must have a probability distribution.

**Definition 72** *The probability distribution of a statistic is called a sampling distribution.*

### 3.4 Sampling Distribution of Means and the Central Limit

**Theorem 73** *If  $X_1, X_2, \dots, X_n$  are independent (?) random variables having normal distributions with means  $\mu_1, \mu_2, \dots, \mu_n$  and variances  $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ , respectively, then the random variable  $Y = a_1X_1 + a_2X_2 + \dots + a_nX_n$  has a normal distribution with mean*

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

*and variance*

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

Suppose that a random sample of  $n$  observations is taken from a normal population with mean  $\mu$  and variance  $\sigma^2$ . Each observation  $X_i, i = 1, 2, \dots, n$ , of the random sample will then have the same normal distribution. Hence, from Theorem 7, we conclude that

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

has a normal distribution with mean

$$\mu_{\bar{X}} = \frac{1}{n} \{\mu + \mu + \dots + \mu\} = \sum_{i=1}^n \mu = \mu$$

and variance

$$\sigma_{\bar{X}}^2 = \frac{1}{n^2} \{\sigma^2 + \sigma^2 + \dots + \sigma^2\} = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}.$$

Hence, we have

**Corollary 74** *If  $X_1, X_2, \dots, X_n$  are independent random variables having normal distributions with means  $\mu$  and variances  $\sigma$ , then the sample mean  $\bar{X}$  is normally distributed with mean equal to  $\mu$  and standard deviation equal to  $\sigma/\sqrt{n}$ . Consequently the random variable*

$$Z = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}}$$

*is a standard normal distribution.*

**Theorem 75 (Central Limit Theorem)** If  $\bar{X}$  is the mean of a random sample of size  $n$  taken from a population with mean  $\mu$  and finite variance  $\sigma^2$ , then the limiting form of the distribution of

$$Z = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}}$$

as  $n \rightarrow \infty$ , is the standard normal distribution  $N(0, 1)$ .

The normal approximation for  $\bar{X}$  will generally be good if  $n \geq 30$ .

**Example 76** An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with 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 77** Here  $\mu = 800$ ,  $\sigma = 40$  and  $n = 16$ . The random variable  $\bar{X}$  is normally distributed with mean  $\mu_{\bar{X}} = \mu = 800$  and standard standard deviation  $\sigma_{\bar{X}} = \sigma_X/\sqrt{n} = 10$ . Then  $(\bar{X} - 800)/10$  is a standard normal distribution  $N(0, 1)$ . Hence,

$$\begin{aligned}\Pr(\bar{X} < 775) &= P((\bar{X} - 800)/10 < (775 - 800)/10) \\ &= P(Z < -2.5) = 0.0062.\end{aligned}$$

**Example 78** Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 5 minutes. In a given week, a bus transported passengers 40 times. What is the probability that the average transport time was more than 30 minutes?

**Solution 79** In this case,  $\mu = 28$  and  $\sigma = 5$ . We need to calculate the probability  $\Pr(\bar{X} > 30)$  with  $n = 40$ . Hence,

$$\begin{aligned}\Pr(\bar{X} > 30) &= \Pr\left(\frac{\bar{X} - 28}{5/\sqrt{40}} \geq \frac{30 - 28}{5/\sqrt{40}}\right) = \Pr(Z \geq 2.53) \\ &= 1 - \Pr(Z \leq 2.53) = 1 - 0.9925 = 0.0075.\end{aligned}$$

There is only a slight chance that the average time of one bus trip will exceed 30 minutes.

### 3.5 Sampling Distribution of the Difference between Two Means

A scientist or engineer may be interested in a comparative experiment in which two manufacturing methods, 1 and 2, are to be compared. The basis for that comparison is  $\mu_1 - \mu_2$ , the difference in the population means. Suppose that we have two populations, the first with mean  $\mu_1$  and variance  $\sigma_1^2$ , and the second with mean  $\mu_2$  and variance  $\sigma_2^2$ . Let the statistic  $\bar{X}_1$  represent the mean of a random sample of size  $n_1$  selected from the first population, and the statistic  $\bar{X}_2$  represent the mean of a random sample of size  $n_2$  selected from the second population, independent of the sample from the first population. What can we say about the sampling distribution of the difference  $\bar{X}_1 - \bar{X}_2$  for repeated samples of size  $n_1$  and  $n_2$ ? According to Theorem 8, the variables  $\bar{X}_1$  and  $\bar{X}_2$  are both approximately normally distributed with means  $\mu_1$  and  $\mu_2$  and variances  $\sigma_1^2/n_1$  and  $\sigma_2^2/n_2$ , respectively. This approximation improves as  $n_1$  and  $n_2$  increase. We can conclude that  $\bar{X}_1 - \bar{X}_2$  is approximately normally distributed with mean

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2$$

and variance

$$\sigma_{\bar{X}_1 - \bar{X}_2}^2 = \sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2 = \sigma_1^2/n_1 + \sigma_2^2/n_2$$

The Central Limit Theorem can be easily extended to the two-sample, two-population case.

**Theorem 80** *If independent samples of size  $n_1$  and  $n_2$  are drawn at random from two populations, discrete or continuous, with means  $\mu_1$  and  $\mu_2$  and variances  $\sigma_1^2$  and  $\sigma_2^2$ , respectively, then the sampling distribution of the differences of means,  $\bar{X}_1 - \bar{X}_2$ , is approximately normally distributed with mean and variance given by*

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 \text{ and } \sigma_{\bar{X}_1 - \bar{X}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Hence,

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

is approximately a standard normal variable.

If both  $n_1$  and  $n_2$  are greater than or equal to 30, the normal approximation for the distribution of  $\bar{X}_1 - \bar{X}_2$  is good. Two independent experiments are run in which two different types of paint are compared.

**Example 81** Eighteen specimens are painted using type A, and the drying time, in hours, is recorded for 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  $P(\bar{X}_A - \bar{X}_B > 1.0)$ , where  $\bar{X}_A$  and  $\bar{X}_B$  are average drying times for samples of size  $n_A = n_B = 18$ .

**Solution 82** From the sampling distribution of  $\bar{X}_A - \bar{X}_B$ , we know that the distribution is approximately normal with mean  $\mu_{\bar{X}_A - \bar{X}_B} = \mu_A - \mu_B = 0$  and variance  $\sigma_{\bar{X}_A - \bar{X}_B}^2 = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B} = 1/9$ . Corresponding to the value  $\bar{X}_A - \bar{X}_B = 1.0$ , we have

$$z = \frac{1 - (\mu_A - \mu_B)}{\sqrt{1/9}} = \frac{1 - 0}{\sqrt{1/9}} = 3$$

so

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

**Example 83** The television picture tubes of manufacturer A have a mean lifetime of 6.5 years 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 84** We are given the following information:

Population 1	Population 2
$\mu_1 = 6.5$	$\mu_2 = 6.0$
$\sigma_1 = 0.9$	$\sigma_2 = 0.8$
$n_1 = 36$	$n_2 = 49$

If we use, the sampling distribution of  $\bar{X}_1 - \bar{X}_2$  will be approximately normal and will have a

mean and standard deviation

$$\mu_{\bar{X}_1 - \bar{X}_2} = 6.5 - 6.0 \text{ and } \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{0.81}{36} + \frac{0.64}{49}} = 0.189$$

Hence,

$$\begin{aligned} \Pr(\bar{X}_1 - \bar{X}_2 \geq 1.0) &= P(Z > 2.65) = 1 - P(Z < 2.65) \\ &= 1 - 0.9960 = 0.0040. \end{aligned}$$

### 3.6 Sampling Distribution of $S^2$

**Theorem 85** If  $X_1, X_2, \dots, X_n$  an independent random sample that have the same standard normal distribution then  $X = \sum_{i=1}^n X_i^2$  is chi-squared distribution, with  $\nu = n$  degrees of freedom.

**Theorem 86** The mean and variance of the chi-squared distribution  $\chi^2$  with  $\nu$  degrees of freedom are  $\mu = \nu$  and  $\sigma^2 = 2\nu$ .

Table A.5 gives values of  $\chi_\alpha^2$  for various values of  $\alpha$  and  $\nu$ . Hence, the  $\chi^2$  value with 7 degrees of freedom, leaving an area of 0.05 to the right, is  $\chi_{0.05}^2 = 14.067$ . Owing to lack of symmetry, we must also use the tables to find  $\chi_{0.95}^2 = 2.167$  for  $\nu = 7$ .

**Example 87** For a chi-squared distribution, find

- (a)  $\chi_{0.025}^2$  when  $\nu = 15$ ;
- (b)  $\chi_{0.01}^2$  when  $\nu = 7$ ;
- (c)  $\chi_{0.05}^2$  when  $\nu = 24$ .

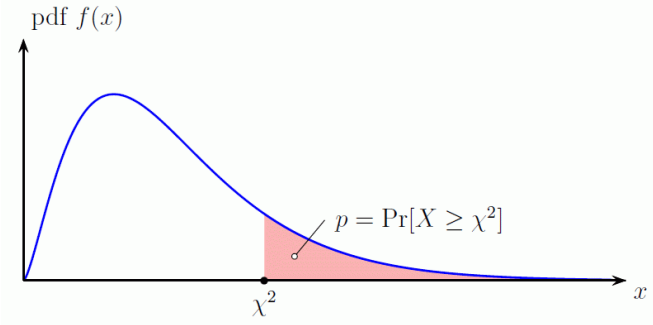
**Solution 88** (a) 27.488. (b) 18.475. (c) 36.415

For a chi-squared distribution  $X$ , find  $\chi_\alpha^2$  such that

- (a)  $P(X > \chi_\alpha^2) = 0.99$  when  $\nu = 4$ ;
- (b)  $P(X > \chi_\alpha^2) = 0.025$  when  $\nu = 19$ ;
- (c)  $P(37.652 < X < \chi_\alpha^2) = 0.045$  when  $\nu = 25$ .

**Solution 89** (a)  $\chi_{\alpha}^2 = \chi_{0.99}^2 = 0.297$ . (b)  $\chi_{\alpha}^2 = \chi_{0.025}^2 = 32.852$ . (c)  $\chi_{0.05}^2 = 37.652$ . Therefore,  $\alpha = 0.05 - 0.045 = 0.005$ . Hence,  $\chi_{\alpha}^2 = \chi_{0.005}^2 = 46.928$ .





**Theorem 90** *If  $S^2$  is the variance of a random sample of size  $n$  taken from a normal population having the variance  $\sigma^2$ , then the statistic*

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2}$$

*has a chi-squared distribution with  $\nu = n - 1$  degrees of freedom.*

### 3.7 t-Distribution

**Theorem 91** *Let  $Z$  be a standard normal random variable and  $V$  a chi-squared random variable with  $\nu$  degrees of freedom. If  $Z$  and  $\nu$  are independent, then the distribution of the random variable  $T$ , where*

$$T = \frac{Z}{\sqrt{V/\nu}}$$

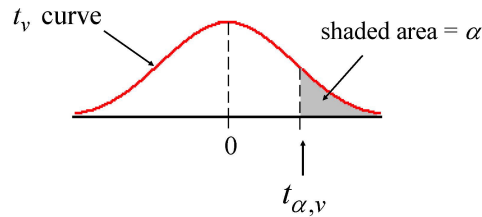
*This is known as the t-distribution with  $\nu$  degrees of freedom.*

**Corollary 92** *Let  $X_1, X_2, \dots, X_n$  be independent random variables that are all normal with mean  $\mu$  and standard deviation  $\sigma$ . Let*

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

*Then the random variable  $T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$  has a t-distribution with  $\nu = n - 1$  degrees of freedom.*

### Pictorial Definition of $t_{\alpha, \nu}$



**Example 93** The  $t$ -value with  $\nu = 14$  degrees of freedom that leaves an area of 0.025 to the left, and therefore an area of 0.975 to the right, is

$$t_{0.975} = -t_{0.025} = -2.145$$

**Example 94** Find  $\Pr(-t_{0.025} < T < t_{0.05})$ .

**Solution 95** Since  $t_{0.05}$  leaves an area of 0.05 to the right, and  $-t_{0.025}$  leaves an area of 0.025 to the left, we find a total area of  $1 - 0.05 - 0.025 = 0.925$  between  $-t_{0.025}$  and  $t_{0.05}$ . Hence  $\Pr(-t_{0.025} < T < t_{0.05}) = 0.925$ .

**Example 96** Find  $k$  such that  $\Pr(k < T < -1.761) = 0.045$  for a random sample of size 15 selected from a normal distribution with  $T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$ .

**Solution 97** From Table A.4 we note that 1.761 corresponds to  $t_{0.05}$  when  $\nu = 14$ . Therefore,  $-t_{0.05} = -1.761$ . Since  $k$  in the original probability statement is to the left of  $-t_{0.05} = -1.761$ , let  $k = -t_{\alpha}$ . Then, by using figure, we have

$$0.045 = 0.05 - \alpha, \text{ or } \alpha = 0.005.$$

Hence, from Table A.4 with  $\nu = 14$ ,

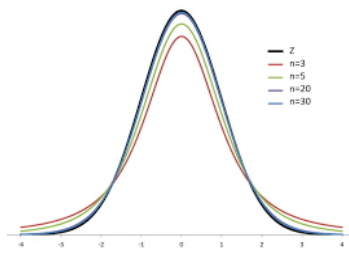


Figure 3-1: t-distribution

$k = -t_{0.005} = -2.977$  and  $\Pr(-2.977 < T < -1.761) = 0.045$ .

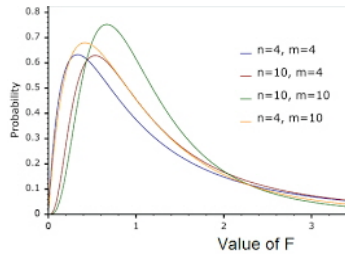


Figure 3-2: F-distribution

### 3.8 F-Distribution

The statistic  $F$  is defined to be the ratio of two independent chi-squared random variables, each divided by its number of degrees of freedom.

**Theorem 98** *The random variable*

$$F = \frac{U/\nu_1}{V/\nu_2}$$

where  $U$  and  $V$  are independent random variables having chi-squared distributions with  $\nu_1$  and  $\nu_2$  degrees of freedom, respectively, is the **F-distribution** with  $\nu_1$  and  $\nu_2$  degrees of freedom (d.f.).

**Theorem 99** *Writing  $f_\alpha(\nu_1, \nu_2)$  for  $f_\alpha$  with  $\nu_1$  and  $\nu_2$  degrees of freedom, we have*

$$f_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{f_\alpha(\nu_2, \nu_1)}$$

Thus, the  $f$ -value with 6 and 10 degrees of freedom, leaving an area of 0.95 to the right, is  $f_{0.95}(6, 10) = \frac{1}{f_{0.05}(10, 6)} = \frac{1}{4.06} = 0.246$ .

#### 3.8.1 The F-Distribution with Two Sample Variances

Suppose that random samples of size  $n_1$  and  $n_2$  are selected from two normal populations with variances  $\sigma_1^2$  and  $\sigma_2^2$ , respectively. From Theorem 16, we know that

$$\chi_1^2 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \text{ and } \chi_2^2 = \frac{(n_2 - 1)S_2^2}{\sigma_2^2}$$

are random variables having chi-squared distributions with  $\nu_1 = n_1 - 1$  and  $\nu_2 = n_2 - 1$  degrees of freedom. Furthermore, since the samples are selected at random, we are dealing with independent random variables. Then, using Theorem 24 with  $\chi_1^2 = U$  and  $\chi_2^2 = V$ , we obtain the following result.

**Theorem 100** *If  $S_1^2$  and  $S_2^2$  are the variances of independent random samples of size  $n_1$  and  $n_2$  taken from normal populations with variances  $\sigma_1^2$  and  $\sigma_2^2$ , respectively, then*

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

*has an  $F$ -distribution with  $\nu_1 = n_1 - 1$  and  $\nu_2 = n_2 - 1$  degrees of freedom.*

### 3.8.2 Example

For an  $F$ -distribution, find

- (a)  $f_{0.05}$  with  $\nu_1 = 7$  and  $\nu_2 = 15$ ;
- (b)  $f_{0.05}$  with  $\nu_1 = 15$  and  $\nu_2 = 7$ ;
- (c)  $f_{0.01}$  with  $\nu_1 = 24$  and  $\nu_2 = 19$ ;
- (d)  $f_{0.95}$  with  $\nu_1 = 19$  and  $\nu_2 = 24$ ;
- (e)  $f_{0.99}$  with  $\nu_1 = 28$  and  $\nu_2 = 12$ .

**Solution 101** (a) 2.71. (b) 3.51. (c) 2.92. (d)  $1/2.11 = 0.47$ . (e)  $1/2.90 = 0.34$ .

## 3.9 Sampling Distribution of Proportions and the Central Limit

In many situations the use of the sample proportion is easier and more reliable because, unlike the mean, the proportion does not depend on the population variance, which is usually an unknown quantity. We will represent the sample proportion by  $\hat{P}$  and the population proportion by  $p$ . Construction of the sampling distribution of the sample proportion is done in a manner similar to that of the mean. One has  $\hat{P} = X/n$  where  $X$  is a number of success for a sample of size  $n$ . It is clear that  $X$  is a binomial distribution  $Bin(n, p)$ . Its mean  $\mu_X = np$  and its variance  $\sigma_X^2 = np(1 - p)$ . Consequently:

**Theorem 102** The mean  $\mu_{\hat{p}}$  of the sample distribution  $\hat{P}$  is equal to the true population proportion  $p$ , and its variance  $\sigma_{\hat{p}}^2$  is equal to  $p(1-p)/n$ .

**Theorem 103 (Theorem Central Limit)** If  $np \geq 5$  and  $n(1-p) \geq 5$ , then the random variable  $\hat{P}$  is approximation a normal distribution with mean  $\mu_{\hat{p}} = p$  and standard deviation (or standard error)  $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$ . Hence

$$Z = \frac{\hat{P} - p}{\sqrt{p(1-p)/n}}$$

is approxiamately a standard normal distribution.

**Example 104** In the mid seventies, according to a report by the National Center for Health Statistics, 19.4 percent of the adult U.S. male population was obese. What is the probability that in a simple random sample of size 150 from this population fewer than 15 percent will be obese?

**Solution 105** Here  $n = 150$ ,  $p = 0.194$ . Since  $np \geq 5$  and  $n(1-p) \geq 5$ , hence

$$Z = \frac{\hat{P} - 0.194}{\sqrt{0.194(1-0.194)/150}} = \frac{\hat{P} - 0.194}{0.032}$$

is approxiamately a standard normal distribution.

$$\Pr(\hat{P} \leq 0.15) = \Pr\left(\frac{\hat{P}-0.194}{0.03} \leq \frac{0.15-0.194}{0.03}\right) \simeq \Pr(Z \leq -1.37) = 0.0853.$$

### 3.10 Sampling Distribution of the Difference between Two Proportions

In some applications there are two actual physical dichotomous populations so that  $p_1$  denotes the population success proportion for population one and  $p_2$  denotes the population success proportion for population two. The sampling distribution of the difference between two sample proportions is constructed in a manner similar to the difference between two means. Independent random samples of size  $n_1$  and  $n_2$  are drawn from two populations of dichotomous variables where the proportions of observations with the character of interest in the two populations are  $p_1$  and  $p_2$ , respectively.

**Theorem 106** The mean  $\mu_{\hat{p}_1 - \hat{p}_2}$  of the sample distribution of the difference between two sample proportions  $\hat{P}_1 - \hat{P}_2$  is equal to the difference  $p_1 - p_2$  between the true population proportions, and its variance  $\sigma_{\hat{p}_1 - \hat{p}_2}^2$  will be equal to  $p_1(1 - p_1)/n_1 + p_1(1 - p_2)/n_2$ .

**Theorem 107** If  $n_1p_1 \geq 5$ ,  $n_1(1 - p_1) \geq 5$ ,  $n_2p_2 \geq 5$ ,  $n_2(1 - p_2)$ , then the random variable  $\hat{P}_1 - \hat{P}_2$  is approximation a normal distribution with mean  $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$  and standard deviation (or standard error)  $\sigma_{\hat{p}} = \sqrt{p_1(1 - p_1)/n_1 + p_1(1 - p_2)/n_2}$ . Hence

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

is approxiamately a standard normal distribution.

**Example 108** Suppose that there are two large high schools, each with more than 2000 students, in a certain town. At School 1, 70% of students did their homework last night. Only 50% of the students at School 2 did their homework last night. The counselor at School 1 takes a sample random sample of 100 students and records the proportion that did homework. School 2's counselor takes a sample random sample of 200 students and records the proportion that did homework. Find the probability of getting a difference in sample proportion  $\hat{P}_1 - \hat{P}_2$  of 0.10 or less from the two surveys.

**Solution 109** Here  $p_1 = 0.7$ ,  $p_2 = 0.5$ ,  $n_1 = 100$  and  $n_2 = 200$ . It is clear that  $n_1p_1 \geq 5$ ,  $n_1(1 - p_1) \geq 5$ ,  $n_2p_2 \geq 5$ ,  $n_2(1 - p_2)$ . Also  $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2 = 0.2$  and  $\sigma_{\hat{p}} = \sqrt{p_1(1 - p_1)/n_1 + p_1(1 - p_2)/n_2} = 0.058$ . Hence,

$$Z = \frac{\hat{P}_1 - \hat{P}_2 - p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} = \frac{\hat{P}_1 - \hat{P}_2 - 0.2}{0.058}$$

is approximately a standard normal.

$$\Pr(\hat{P}_1 - \hat{P}_2 \leq 0.10) = \Pr\left(\frac{\hat{P}_1 - \hat{P}_2 - 0.2}{0.058} \leq \frac{0.10 - 0.2}{0.058}\right) \simeq \Pr(Z \leq -1.72) = 0.0427.$$

## Chapter 4

# One and Two-Sample Estimation Problems

### 4.1 One- and Two-Sample Estimation Problems

#### 4.1.1 Introduction

In previous chapters, we emphasized sampling properties of the sample mean and variance. The purpose of these presentations is to build a foundation that allows us to draw conclusions about the population parameters from experimental data.

#### 4.1.2 Classical Methods of Estimation

A point estimate of some population parameter  $\theta$  is a single value  $\hat{\theta}$  of a statistic  $\hat{\Theta}$ . For example, the value  $\bar{x}$  of the statistic  $\bar{X}$ , computed from a sample of size  $n$ , is a point estimate of the population parameter  $\mu$ . Similarly,  $\hat{p} = x/n$  is a point estimate of the true proportion  $p$  for a binomial experiment.

An estimator is not expected to estimate the population parameter without error. We do not expect  $\bar{X}$  to estimate  $\mu$  exactly, but we certainly hope that it is not far off.



## Unbiased Estimator

What are the desirable properties of a “good” decision function that would influence us to choose one estimator rather than another? Let  $\hat{\Theta}$  be an estimator whose value  $\hat{\theta}$  is a point estimate of some unknown population parameter  $\theta$ . Certainly, we would like the sampling distribution of  $\hat{\Theta}$  to have a mean equal to the parameter estimated. An estimator possessing this property is said to be unbiased.

**Definition 110** *A statistic  $\hat{\Theta}$  is said to be an unbiased estimator of the parameter  $\theta$  if  $\mu_{\hat{\Theta}} = E(\hat{\Theta}) = \theta$ .*

**Example 111** *Show that  $S^2$  is an unbiased estimator of the parameter  $\sigma^2$ . Hint:  $(X_i - \bar{X}) = (X_i - \mu) - (\bar{X} - \mu)$ .*

## Variance of a Point Estimator

If  $\hat{\Theta}_1$  and  $\hat{\Theta}_2$  are two unbiased estimators of the same population parameter  $\theta$ , we want to choose the estimator whose sampling distribution has the smaller variance.

Hence, if  $\sigma_{\hat{\Theta}_1}^2 < \sigma_{\hat{\Theta}_2}^2$ , we say that  $\hat{\Theta}_1$  is a more efficient estimator of  $\theta$  than  $\hat{\Theta}_2$ .

**Definition 112** *If we consider all possible unbiased estimators of some parameter  $\theta$ , the one with the smallest variance is called the most efficient estimator of  $\theta$ .*

## Interval Estimation

Even the most efficient unbiased estimator is unlikely to estimate the population parameter exactly. There is no reason we should expect a **point estimate** from a given sample to be exactly equal to the population parameter it is supposed to estimate. There are many situations in which it is preferable to determine an interval within which we would expect to find the value of the parameter. Such an interval is called an interval estimate. An interval estimate of a population parameter  $\theta$  is an interval of the form  $\hat{\theta}_L < \theta < \hat{\theta}_U$ , where  $\hat{\theta}_L$  and  $\hat{\theta}_U$  depend on the value of the statistic  $\hat{\Theta}$  for a particular sample and also on the sampling distribution of  $\hat{\Theta}$ .

## 4.2 Single Sample: Estimating the Mean

The sampling distribution of  $\bar{X}$  is centered at  $\mu$ , and in most applications the variance is smaller than that of any other estimators of  $\mu$ . Thus, the sample mean  $\bar{x}$  will be used as a point estimate for the population mean  $\mu$ .

Let us now consider the interval estimate of  $\mu$ . If our sample is selected from a normal population or, failing this, if  $n$  is sufficiently large, we can establish a confidence interval for  $\mu$  by considering the sampling distribution of  $\bar{X}$ .

**Definition 113 (Confidence Interval on  $\mu$ ,  $\sigma^2$  Known)** *If  $\bar{x}$  is the mean of a random sample of size  $n$  from a population with known variance  $\sigma^2$ , a  $100(1 - \alpha)\%$  confidence interval for  $\mu$  is given by*

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}},$$

where  $z_{\alpha/2}$  is the  $z$ -value leaving an area of  $\alpha/2$  to the right.

**Example 114** *The average zinc concentration recovered from a sample of measurements taken in 36 different locations in a river is found to be 2.6 grams per milliliter. Find the 95% and 99% confidence intervals for the mean zinc concentration in the river. Assume that the population standard deviation is 0.3 gram per milliliter.*

**Solution 115** *The point estimate of  $\mu$  is  $\bar{x} = 2.6$ . The  $z$ -value leaving an area of 0.025 to the right, and therefore an area of 0.975 to the left, is  $z_{0.025} = 1.96$  (Table A.3). Hence, the 95% confidence interval is*

$$2.6 - (1.96) \left( \frac{0.3}{\sqrt{36}} \right) < \mu < 2.6 + (1.96) \left( \frac{0.3}{\sqrt{36}} \right)$$

which reduces to  $2.50 < \mu < 2.70$ . To find a 99% confidence interval, we find the  $z$ -value leaving an area of 0.005 to the right and 0.995 to the left. From Table A.3 again,  $z_{0.005} = 2.575$ , and the 99% confidence interval is

$$2.6 - (2.575) \left( \frac{0.3}{\sqrt{36}} \right) < \mu < 2.6 + (2.575) \left( \frac{0.3}{\sqrt{36}} \right)$$

or simply

$$2.47 < \mu < 2.73.$$

The error in estimating  $\mu$  by  $\bar{x}$  is the absolute value of the difference between  $\mu$  and  $\bar{x}$ , and we can be  $100(1 - \alpha)\%$  confident that this difference will not exceed  $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ .

**Theorem 116** *If  $\bar{x}$  is used as an estimate of  $\mu$ , we can be  $100(1 - \alpha)\%$  confident that the error will not exceed  $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ .*

**Theorem 117** *If  $\bar{x}$  is used as an estimate of  $\mu$ , we can be  $100(1 - \alpha)\%$  confident that the error will not exceed a specified amount  $e$  when the sample size is*

$$n = \left( \frac{z_{\alpha/2} \sigma}{e} \right)^2$$

**Example 118** *How large a sample is required if we want to be 95% confident that our estimate of  $\mu$  in Example 5 is off by less than 0.05?*

**Solution 119** *The population standard deviation is  $\sigma = 0.3$ . Then,*

$$n = \left[ \frac{(1.96)(0.3)}{0.05} \right]^2 = 138.3.$$

*Therefore, we can be 95% confident that a random sample of size 139 will provide an estimate  $\bar{x}$  differing from  $\mu$  by an amount less than 0.05.*

The reader should recall learning in Chapter 3 that if we have a random sample from a normal distribution, then the random variable

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a Student  $t$ -distribution with  $n - 1$  degrees of freedom. Here  $S$  is the sample standard deviation. In this situation, with  $\sigma$  unknown,  $T$  can be used to construct a confidence interval on  $\mu$ .

**Definition 120 (Confidence Interval on  $\mu$ ,  $\sigma^2$  unknown)** If  $\bar{x}$  and  $s$  are the mean and standard deviation of a random sample of size  $n$  from a normal population with unknown variance  $\sigma^2$ , a  $100(1 - \alpha)\%$  confidence interval for  $\mu$  is

$$\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}},$$

where  $t_{\alpha/2}$  is the  $t$ -value with  $\nu = n - 1$  degrees of freedom, leaving an area of  $\alpha/2$  to the right.

**Example 121** The contents of seven similar containers of sulfuric acid are 9.8, 10.2, 10.4, 9.8, 10.0, 10.2, 9.6 liters. Find a 95% confidence interval for the mean contents of all such containers, assuming an approximately normal distribution.

**Solution 122** The sample mean and standard deviation for the given data are

$$\bar{x} = 10.0 \text{ and } s = 0.283.$$

Using Table A.4, we find  $t_{0.025} = 2.447$  for  $\nu = 6$  degrees of freedom. Hence, the 95% confidence interval for  $\mu$  is

$$10.0 - (2.447) \left( \frac{0.283}{\sqrt{7}} \right) < \mu < 10.0 + (2.447) \left( \frac{0.283}{\sqrt{7}} \right)$$

which reduces to  $9.74 < \mu < 10.26$ .

## Concept of a Large-Sample Confidence Interval

Often statisticians recommend that even when normality cannot be assumed,  $\sigma$  is unknown, and  $n \geq 30$ ,  $s$  can replace  $\sigma$  and the confidence interval

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

may be used. This is often referred to as a large-sample confidence interval.

**Example 123** Scholastic Aptitude Test (SAT) mathematics scores of a random sample of 500 high school seniors in the state of Texas are collected, and the sample mean and standard

deviation are found to be 501 and 112, respectively. Find a 99% confidence interval on the mean SAT mathematics score for seniors in the state of Texas.

### 4.3 Standard Error of a Point Estimate

We indicated earlier that a measure of the quality of an unbiased estimator is its variance. The variance of  $\bar{X}$  is

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

Thus, the standard deviation of  $\bar{X}$ , or standard error of  $\bar{X}$ , is  $\sigma/\sqrt{n}$ . Simply put, the standard error of an estimator is its standard deviation. For  $\bar{X}$ , the computed confidence limit

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \text{ is written } \bar{x} \pm z_{\alpha/2} \text{ s.e.}(\bar{x})$$

In the case where  $\sigma$  is unknown and sampling is from a normal distribution,  $s$  replaces  $\sigma$  and the estimated standard error  $s/\sqrt{n}$  is involved. Thus, the confidence limits on  $\mu$  are limit

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \text{ is written } \bar{x} \pm t_{\alpha/2} \text{ s.e.}(\bar{x})$$

### 4.4 Two Samples: Estimating the Difference between Two Means

**Theorem 124** *Confidence Interval for  $\mu_1 - \mu_2$ ,  $\sigma_1^2$  and  $\sigma_2^2$  known*

If  $\bar{x}_1$  and  $\bar{x}_2$  are means of independent random samples of sizes  $n_1$  and  $n_2$  from populations with known variances  $\sigma_1^2$  and  $\sigma_2^2$ , respectively, a  $100(1-\alpha)\%$  confidence interval for  $\mu_1 - \mu_2$  is given by

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

where  $z_{\alpha/2}$  is the  $z$ -value leaving an area of  $\alpha/2$  to the right.

**Example 125** *A study was conducted in which two types of engines, A and B, were compared. Gas mileage, in miles per gallon, was measured. Fifty experiments were conducted using engine type A and 75 experiments were done with engine type B. The gasoline used and other conditions were held constant. The average gas mileage was 36 miles per gallon for engine A and 42 miles*

per gallon for engine B. Find a 96% confidence interval on  $\mu_B - \mu_A$ , where  $\mu_A$  and  $\mu_B$  are population mean gas mileages for engines A and B, respectively. Assume that the population standard deviations are 6 and 8 for engines A and B, respectively.

**Solution 126** The point estimate of  $\mu_B - \mu_A$  is  $\bar{x}_B - \bar{x}_A = 42 - 36 = 6$ . Using  $\alpha = 0.04$ , we find  $z_{0.02} = 2.05$  from Table A.3. Hence, with substitution in the formula above, the 96% confidence interval is

$$6 - 2.05\sqrt{\frac{64}{75} + \frac{36}{50}} < \mu_B - \mu_A < 6 + 2.05\sqrt{\frac{64}{75} + \frac{36}{50}}$$

or simply  $3.43 < \mu_B - \mu_A < 8.57$ .

## Variations Unknown but Equal

Consider the case where  $\sigma_1^2$  and  $\sigma_2^2$  are unknown and  $\sigma_1^2 = \sigma_2^2 (= \sigma^2)$ . A point estimate of the unknown common variance  $\sigma^2$  can be obtained by pooling the sample variances. Denoting the pooled estimator by  $S_p^2$ , we have the following.

**Definition 127 (of Variance)**  $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1+n_2-1)}$

**Theorem 128 Confidence Interval for  $\mu_1 - \mu_2$ ,  $\sigma_1^2 = \sigma_2^2$  but Both Unknown**

If  $\bar{x}_1$  and  $\bar{x}_2$  are means of independent random samples of sizes  $n_1$  and  $n_2$ , respectively, from approximately normal populations with unknown but equal variances, a  $100(1 - \alpha)\%$  confidence interval for  $\mu_1 - \mu_2$  is given by

$$(\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2}} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2}} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where  $s_p$  is the pooled estimate of the population standard deviation and  $t_{\alpha/2}$  is the  $t$ -value with  $\nu = n_1 + n_2 - 2$  degrees of freedom, leaving an area of  $\alpha/2$  to the right.

**Example 129** Two independent sampling stations, station 1 and station 2, were chosen for a study on pollution. For 12 monthly samples collected at station 1, the species diversity index had a mean value  $\bar{x}_1 = 3.11$  and a standard deviation  $s_1 = 0.771$ , while 10 monthly samples collected at the station 2 had a mean index value  $\bar{x}_2 = 2.04$  and a standard deviation  $s_2 = 0.448$ . Find a

90% confidence interval for the difference between the population means for the two locations, assuming that the populations are approximately normally distributed with equal variances.

**Solution 130** Let  $\mu_1$  and  $\mu_2$  represent the population means, respectively, for the species diversity indices at the downstream and upstream stations. We wish to find a 90% confidence interval for  $\mu_1 - \mu_2$ . Our point estimate of  $\mu_1 - \mu_2$  is

$$\bar{x}_1 - \bar{x}_2 = 3.11 - 2.04 = 1.07.$$

The pooled estimate,  $s_p^2$ , of the common variance,  $\sigma^2$ , is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 1)} = \frac{(11)(0.7712) + (9)(0.4482)}{12 + 10 - 2} = 0.417.$$

Taking the square root, we obtain  $s_p = 0.646$ . Using  $\alpha = 0.1$ , we find in Table A.4 that  $t_{0.05} = 1.725$  for  $\nu = n_1 + n_2 - 2 = 20$  degrees of freedom. Therefore, the 90% confidence interval for  $\mu_1 - \mu_2$  is

$$\begin{aligned} 1.07 + 1.725(0.646)\sqrt{\frac{1}{12} + \frac{1}{10}} &< \mu_1 - \mu_2 \\ &< 1.07 + 1.725(0.646)\sqrt{\frac{1}{12} + \frac{1}{10}} \end{aligned}$$

which simplifies to  $0.593 < \mu_1 - \mu_2 < 1.547$ .

## 4.5 Paired Observations

Now we shall consider estimation procedures for the difference of two means when the samples are not independent and the variances of the two populations are not necessarily equal. The situation considered here deals with a very special experimental condition, namely that of paired observations. For example, if we run a test on a new diet using 15 individuals, the weights before and after going on the diet form the information for our two samples. The two populations are “before” and “after,” and the experimental unit is the individual. Obviously, the observations in a pair have something in common. To determine if the diet is effective, we consider the differences  $d_1, d_2, \dots, d_n$  in the paired observations. These differences are the values

of a random sample  $D_1, D_2, \dots, D_n$  from a population of differences that we shall assume to be normally distributed with mean  $\mu_D = \mu_1 - \mu_2$  and variance  $\sigma_D^2$ . We estimate  $\sigma_D^2$  by  $\sigma_d^2$ , the variance of the differences that constitute our sample. The point estimator of  $\mu_D$  is given by  $\bar{D}$ .

**Theorem 131 Confidence Interval for  $\mu_D = \mu_1 - \mu_2$ , for Paired Observations**

If  $\bar{d}$  and  $s_d$  are the mean and standard deviation, respectively, of the normally distributed differences of  $n$  random pairs of measurements, a  $100(1 - \alpha)\%$  confidence interval for  $\mu_D = \mu_1 - \mu_2$  is

$$\bar{d} - t_{\alpha/2} \frac{s_d}{\sqrt{n}} < \mu < \bar{d} + t_{\alpha/2} \frac{s_d}{\sqrt{n}},$$

where  $t_{\alpha/2}$  is the  $t$ -value with  $\nu = n - 1$  degrees of freedom, leaving an area of  $\alpha/2$  to the right.

**Example 132** A study published in *Chemosphere* reported the levels of the dioxin TCDD of 10 Massachusetts Vietnam veterans who were possibly exposed to Agent Orange. The TCDD levels in plasma and in fat tissue are listed in Table 1. Find a 95% confidence interval for  $\mu_1 - \mu_2$ , where  $\mu_1$  and  $\mu_2$  represent the true mean TCDD levels in plasma and in fat tissue, respectively. Assume the distribution of the differences to be approximately normal.

Veteran	TCDD levels in Plasma	TCDD levels in Fat Tissue	$d_i$
1	2.5	4.9	-2.4
2	3.1	5.9	-2.8
3	2.1	4.4	-2.3
4	3.5	6.9	-3.4
5	3.1	7.0	-3.9
6	1.8	4.2	-2.4
7	6.0	10.0	-4.0
8	3.0	5.5	-2.5
9	36.0	41.0	-5.0
10	4.7	4.4	0.3

**Solution 133** The point estimate of  $\mu_D$  is  $\bar{d} = -2.84$ . The standard deviation,  $s_d$ , of the sample differences is 1.42. Using  $\alpha = 0.05$ , we find in Table A.4 that  $t_{0.025} = 2.262$  for  $\nu =$



$n - 1 = 9$  degrees of freedom. Therefore, the 95% confidence interval is

$$-2.84 - (2.262) \left( \frac{1.42}{\sqrt{10}} \right) < \mu_D < -2.84 + (2.262) \left( \frac{1.42}{\sqrt{10}} \right)$$

or simply  $-3.85 < \mu_D < -1.82$ .

## 4.6 Single Sample: Estimating a Proportion

A point estimator of the proportion  $p$  in a binomial experiment is given by the

statistic  $\hat{P} = X/n$ , where  $X$  represents the number of successes in  $n$  trials. Therefore,

**Definition 134** the sample proportion  $\hat{p} = x/n$  will be used as the point estimate of the

parameter  $p$ .

**Theorem 135 (Large-Sample Confidence Intervals for  $p$ )** If  $\hat{p}$  is the proportion of successes in a random sample of size  $n$  and  $\hat{q} = 1 - \hat{p}$ , an approximate  $100(1 - \alpha)\%$  confidence interval, for the binomial parameter  $p$  is given by

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

where  $z_{\alpha/2}$  is the  $z$ -value leaving an area of  $\alpha/2$  to the right.

**Example 136** In a random sample of  $n = 500$  families owning television sets in the city of Hamilton, Canada, it is found that  $x = 340$  subscribe to HBO. Find a 95% confidence interval for the actual proportion of families with television sets in this city that subscribe to HBO.

**Solution 137** The point estimate of  $p$  is  $\hat{p} = 340/500 = 0.68$ . Using Table A.3, we find that  $z_{0.025} = 1.96$ . Therefore, the 95% confidence interval for  $p$  is

$$0.68 - 1.96 \sqrt{\frac{(0.68)(0.32)}{500}} < p < 0.68 + 1.96 \sqrt{\frac{(0.68)(0.32)}{500}}$$

which simplifies to  $0.6391 < p < 0.7209$ .

**Theorem 138** If  $\hat{p}$  is used as an estimate of  $p$ , we can be  $100(1 - \alpha)\%$  confident that the error will not exceed  $z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$ .

## Choice of Sample Size

Let us now determine how large a sample is necessary to ensure that the error in estimating  $p$  will be less than a specified amount  $e$ . By Theorem 23, we must choose  $n$  such that  $z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} = e$ .

**Theorem 139** If  $\hat{p}$  is used as an estimate of  $p$ , we can be  $100(1 - \alpha)\%$  confident that the error will be less than a specified amount  $e$  when the sample size is approximately

$$n = \frac{z_{\alpha/2}^2 \hat{p}\hat{q}}{e^2}$$

**Example 140** How large a sample is required if we want to be 95% confident that our estimate of  $p$  in Example 21 is within 0.02 of the true value?

**Solution 141** Let us treat the 500 families as a preliminary sample, providing an estimate  $\hat{p} = 0.68$ . Then,

$$n = \frac{(1.96)^2(0.68)(0.32)}{0.02^2} = 2089.8 \approx 2090$$

Occasionally, it will be impractical to obtain an estimate of  $p$  to be used for determining the sample size for a specified degree of confidence. If this happens, we use the following theorem.

**Theorem 142** If  $\hat{p}$  is used as an estimate of  $p$ , we can be  $100(1 - \alpha)\%$  confident that the error will not exceed than a specified amount  $e$  when the sample size is approximately

$$n = \frac{z_{\alpha/2}^2}{4e^2}$$

**Example 143** How large a sample is required if we want to be at least 95% confident that our estimate of  $p$  in Example 21 is within 0.02 of the true value?

**Solution 144** Let assume that no preliminary sample has been taken to provide an estimate of  $p$ . Consequently, we can be at least 95% confident that our sample proportion will not differ from the true proportion by more than 0.02 if we choose a sample of size

$$n = \frac{(1.96)^2}{4(0.02)^2} = 2401$$

Comparing the results of Examples 28 and 29, we see that information concerning  $p$ , provided by a preliminary sample or from experience, enables us to choose a smaller sample while maintaining our required degree of accuracy.

## 4.7 Two Samples: Estimating the Difference between Two Proportions

Consider the problem where we wish to estimate the difference between two binomial parameters  $p_1$  and  $p_2$ . For example,  $p_1$  might be the proportion of smokers with lung cancer and  $p_2$  the proportion of nonsmokers with lung cancer, and the problem is to estimate the difference between these two proportions.

### **Theorem 145** *Large-Sample Confidence Interval for $p_1 - p_2$*

If  $\hat{p}_1$  and  $\hat{p}_2$  are the proportions of successes in random samples of sizes  $n_1$  and  $n_2$ , respectively,  $\hat{q}_1 = 1 - \hat{p}_1$ , and  $\hat{q}_2 = 1 - \hat{p}_2$ , an approximate  $100(1 - \alpha)\%$  confidence interval for the difference of two binomial parameters,  $p_1 - p_2$ , is given by

$$(\hat{p}_1 - \hat{p}_2) - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

**Example 146** *A certain change in a process for manufacturing component parts is being considered. Samples are taken under both the existing and the new process so as to determine if the new process results in an improvement. If 75 of 1500 items from the existing process are found to be defective and 80 of 2000 items from the new process are found to be defective, find a 90% confidence interval for the true difference in the proportion of defectives between the existing and the new process.*

**Solution 147** *Let  $p_1$  and  $p_2$  be the true proportions of defectives for the existing and new processes, respectively. Hence,  $\hat{p}_1 = 75/1500 = 0.05$  and  $\hat{p}_2 = 80/2000 = 0.04$ , and the point estimate of  $p_1 - p_2$  is*

$$\hat{p}_1 - \hat{p}_2 = 0.05 - 0.04 = 0.01$$

Using Table A.3, we find  $z_{0.05} = 1.645$ . Therefore, substituting into the formula, with

$$1.645\sqrt{\frac{(0.05)(0.95)}{1500} + \frac{(0.04)(0.96)}{2000}} = 0.0117,$$

we find the 90% confidence interval to be  $-0.0017 < p_1 - p_2 < 0.0217$ .

## 4.8 Single Sample: Estimating the Variance

If a sample of size  $n$  is drawn from a normal population with variance  $\sigma^2$  and the sample variance  $s^2$  is computed, we obtain a value of the statistic  $S^2$ . This computed sample variance is used as a point estimate of  $\sigma^2$ . Hence, the statistic  $S^2$  is called an estimator of  $\sigma^2$ . An interval estimate of  $\sigma^2$  can be established by using the statistic

$$X = \frac{(n-1)S^2}{\sigma^2}$$

the statistic  $X$  has a chi-squared distribution with  $n - 1$  degrees of freedom when samples are chosen from a normal population.

**Theorem 148 (Confidence Interval for  $\sigma^2$ )** *If  $s^2$  is the variance of a random sample of size  $n$  from a normal population, a  $100(1 - \alpha)\%$  confidence interval for  $\sigma^2$  is*

$$\frac{(n-1)s^2}{\chi_{\alpha/2}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2}$$

where  $\chi_{\alpha/2}^2$  and  $\chi_{1-\alpha/2}^2$  are  $\chi^2$ -values with  $\nu = n - 1$  degrees of freedom, leaving areas of  $\alpha/2$  and  $1 - \alpha/2$ , respectively, to the right.

An approximate  $100(1 - \alpha)\%$  confidence interval for  $\sigma$  is obtained by taking the square root of each endpoint of the interval for  $\sigma^2$ .

**Example 149** *The following are the weights, in decagrams, of 10 packages of grass seed distributed by a certain company: 46.4, 46.1, 45.8, 47.0, 46.1, 45.9, 45.8, 46.9, 45.2, 46.0. Find a 95% confidence interval for the variance of the weights of all such packages of grass seed distributed by this company, assuming a normal population.*

**Solution 150** First we find  $s^2 = 0.286$ . To obtain a 95% confidence interval, we choose  $\alpha = 0.05$ . Then, using Table A.5 with  $\nu = 9$  degrees of freedom, we find  $\chi_{0.025}^2 = 19.023$  and  $\chi_{0.975}^2 = 2.700$ . Therefore, the 95% confidence interval for  $\sigma^2$  is

$$\frac{(9)(0.286)}{19.023} < \sigma^2 < \frac{(9)(0.286)}{2.700}$$

or simply  $0.135 < \sigma^2 < 0.953$ .

## 4.9 Two Samples: Estimating the Ratio of Two Variances

A point estimate of the ratio of two population variances  $\sigma_1^2/\sigma_2^2$  is given by the ratios  $s_1^2/s_2^2$  of the sample variances. Hence, the statistic  $S_1^2/S_2^2$  is called an estimator of  $\sigma_1^2/\sigma_2^2$ . If  $\sigma_1^2$  and  $\sigma_2^2$  are the variances of normal populations, we can establish an interval estimate of  $\sigma_1^2/\sigma_2^2$  by using the statistic

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

According to Theorem 25 of chapter 3, the random variable  $F$  has an  $F$ -distribution with  $\nu_1 = n_1 - 1$  and  $\nu_2 = n_2 - 1$  degrees of freedom.

**Theorem 151 (Confidence Interval for  $\sigma_1^2/\sigma_2^2$ )** If  $s_1^2$  and  $s_2^2$  are the variances of independent samples of sizes  $n_1$  and  $n_2$ , respectively, from normal populations, then a  $100(1 - \alpha)\%$  confidence interval for  $\sigma_1^2/\sigma_2^2$  is

$$\frac{s_1^2}{s_2^2} \frac{1}{f_{\alpha/2}(\nu_1, \nu_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} f_{\alpha/2}(\nu_2, \nu_1)$$

where  $f_{\alpha/2}(\nu_1, \nu_2)$  is an  $f$ -value with  $\nu_1 = n_1 - 1$  and  $\nu_2 = n_2 - 1$  degrees of freedom, leaving an area of  $\alpha/2$  to the right, and  $f_{\alpha/2}(\nu_2, \nu_1)$  is a similar  $f$ -value with  $\nu_2 = n_2 - 1$  and  $\nu_1 = n_1 - 1$  degrees of freedom.

an approximate  $100(1 - \alpha)\%$  confidence interval for  $\sigma_1/\sigma_2$  is obtained by taking the square root of each endpoint of the interval for  $\sigma_1^2/\sigma_2^2$ .

**Example 152** A study was conducted to estimate the difference in the amounts of the chemical orthophosphorus measured at two different stations. Fifteen samples were collected from station

1, and 12 samples were obtained from station 2. The 15 samples from station 1 had an average orthophosphorus content of 3.84 milligrams per liter and a standard deviation of 3.07 milligrams per liter, while the 12 samples from station 2 had an average content of 1.49 milligrams per liter and a standard deviation of 0.80 milligram per liter. Determine a 98% confidence interval for  $\sigma_1^2/\sigma_2^2$  and for  $\sigma_1/\sigma_2$ , where  $\sigma_1^2$  and  $\sigma_2^2$  are the variances of the populations of orthophosphorus contents at station 1 and station 2, respectively.

**Solution 153** We have  $n = 15$ ,  $n_2 = 12$ ,  $s_1 = 3.07$ , and  $s_2 = 0.80$ . For a 98% confidence interval,  $\alpha = 0.02$ . Interpolating in Table A.6, we find  $f_{0.01}(14, 11) \approx 4.30$  and  $f_{0.01}(11, 14) \approx 3.87$ . Therefore, the 98% confidence interval for  $\sigma_1^2/\sigma_2^2$  is

$$\left(\frac{3.07^2}{0.80^2}\right) \left(\frac{1}{4.30}\right) < \frac{\sigma_1^2}{\sigma_2^2} < \left(\frac{3.07^2}{0.80^2}\right) (3.87),$$

which simplifies to  $3.425 < \frac{\sigma_1^2}{\sigma_2^2} < 56.991$ . Taking square roots of the confidence limits, we find that a 98% confidence interval for  $\sigma_1/\sigma_2$  is

$$1.851 < \frac{\sigma_1}{\sigma_2} < 7.549.$$

Since this interval does not allow for the possibility of  $\sigma_1/\sigma_2$  being equal to 1, we were correct in assuming that  $\sigma_1 \neq \sigma_2$  (and  $\sigma_1^2 \neq \sigma_2^2$ ).

## Chapter 5

# One and Two-Sample Tests of Hypotheses

Let consider a medical researcher who should decide on the basis of experimental evidence whether coffee drinking increases the risk of cancer in humans or a sociologist who might wish to collect appropriate data to enable him or her to decide whether a person's blood type and eye color are independent variables. In each of these cases, the scientist postulates or conjectures something about a system. In addition, each must make use of experimental data and make a decision based on the data. In each case, the conjecture can be put in the form of a statistical hypothesis. Procedures that lead to the acceptance or rejection of statistical hypotheses such as these comprise a major area of statistical inference. First, let us define precisely what we mean by a statistical hypothesis.

**Definition 154** *A **statistical hypothesis** is an assertion or conjecture concerning one or more populations.*

The truth or falsity of a statistical hypothesis is never known with absolute certainty unless we examine the entire population. This, of course, would be impractical in most situations. Instead, we take a random sample from the population of interest and use the data contained in this sample to provide evidence that either supports or does not support the hypothesis. Evidence from the sample that is inconsistent with the stated hypothesis leads to a rejection of

the hypothesis.

## The Role of Probability in Hypothesis Testing

It should be made clear to the reader that the decision procedure must include the probability of a wrong conclusion. For example, suppose that the hypothesis postulated by the engineer is that the fraction defective  $p$  in a certain process is 0.10. The experiment is to observe a random sample of the product in question. Suppose that 100 items are tested and 20 items are found defective. If, indeed,  $p = 0.10$ , the probability of obtaining 20 or more defectives is approximately 0.002. With the resulting small risk of a wrong conclusion, it would seem safe to reject the hypothesis that  $p = 0.10$ . As a result, the reader must be accustomed to understanding that rejection of a hypothesis implies that the sample evidence refutes it. Put another way, rejection means that there is a small probability of obtaining the sample information observed when, in fact, the hypothesis is true.

The formal statement of a hypothesis is often influenced by the structure of the probability of a wrong conclusion. If the scientist is interested in strongly supporting a contention, he or she hopes to arrive at the contention in the form of rejection of a hypothesis. If the medical researcher wishes to show strong evidence in favor of the contention that coffee drinking increases the risk of cancer, the hypothesis tested should be of the form “there is no increase in cancer risk produced by drinking coffee.” As a result, the contention is reached via a rejection. Similarly, to support the claim that one kind of gauge is more accurate than another, the engineer tests the hypothesis that there is no difference in the accuracy of the two kinds of gauges.

## The Null and Alternative Hypotheses

The alternative hypothesis  $H_1$  usually represents the question to be answered or the theory to be tested, and thus its specification is crucial. The null hypothesis  $H_0$  nullifies or opposes  $H_1$  and is often the logical complement to  $H_1$ . Our conclusions will be:

fail to reject  $H_0$  because of insufficient evidence in the data.

reject  $H_0$  in favor of  $H_1$  because of sufficient evidence in the data or

fail to reject  $H_0$  because of insufficient evidence in the data.

## The Probability of a Type I Error



The decision procedure could lead to either of two wrong conclusions.

**Definition 155** *Rejection of the null hypothesis when it is true is called a **type I error**.*

The probability of committing a type I error, also called the level of significance, is denoted by the Greek letter  $\alpha$ .

**Definition 156** *Nonrejection of the null hypothesis when it is false is called a **type II error**.*

The probability of committing a type II error, denoted by  $\beta$ , is impossible to compute unless we have a specific alternative hypothesis.

## 5.1 Single Sample: Tests Concerning a Single Mean

In this section, we formally consider tests of hypotheses on a single population mean.

### Tests on a Single Mean (Variance Known)

Let  $X_1, X_2, \dots, X_n$  representing a random sample from a distribution with mean  $\mu$  and variance  $\sigma^2 > 0$ . Consider first the hypothesis

$$H_0 : \mu = \mu_0,$$

$$H_1 : \mu \neq \mu_0.$$

The appropriate test statistic should be based on the random variable  $\bar{X}$ . Recall that the random variable  $\bar{X}$  has approximately a normal distribution with mean  $\mu$  and variance  $\sigma^2/n$  for reasonably large sample sizes. So,  $\mu_{\bar{X}} = \mu$  and  $\sigma_{\bar{X}}^2 = \sigma^2/n$ . We can then determine a critical region based on the computed sample average,  $\bar{x}$ . We know that under  $H_0$ , that is, if  $\mu = \mu_0$ ,  $(\bar{X} - \mu_0)/(\sigma/\sqrt{n})$  follows an  $N(0, 1)$  distribution, and hence the expressions

$$\Pr\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

can be used to write an appropriate nonrejection region.

**Theorem 157 (Test Procedure for a Single Mean (Variance Known))** *If  $-z_\alpha/2 < z < z_\alpha/2$ , do not reject  $H_0$ . Rejection of  $H_0$ , of course, implies acceptance of the alternative hypothesis  $\mu = \mu_0$ . With this definition of the critical region, it should be clear that there will be probability  $\alpha$  of rejecting  $H_0$  (falling into the critical region) when, indeed,  $\mu = \mu_0$ .*

### Tests of one-sided hypotheses on the mean

For example, suppose that we seek to test

$$H_0 : \mu = \mu_0,$$

$$H_1 : \mu > \mu_0.$$

The signal that favors  $H_1$  comes from large values of  $z$ . Thus, rejection of  $H_0$  results when the computed  $z > z_\alpha$ .

**Example 158** *A random sample of 100 recorded deaths in the United States during the past year showed an average life span of 71.8 years. Assuming a population standard deviation of 8.9 years, does this seem to indicate that the mean life span today is greater than 70 years? Use a 0.05 level of significance.*

**Solution 159** 1.  $H_0 : \mu = 70$  years.

2.  $H_1 : \mu > 70$  years.

3.  $\alpha = 0.05$ .

4. Test statistic:  $z = (\bar{x} - \mu_0)/(\sigma/\sqrt{n}) = (71.8 - 70)/(8.9/\sqrt{100}) = 2.02$ .

5. Critical region:  $z > z_\alpha$ , where  $z_\alpha = z_{0.05} = 1.645$

6. Decision: since  $z = 2.02 > 1.645$ , reject  $H_0$  and conclude that the mean life span today is greater than 70 years.

The  $P$ -value corresponding to  $z = 2.02$  is given by the area on the right under the density of standard normal distribution. Using Table A.3, we have  $P\text{-value} = P(Z > 2.02) = 0.0217$ . As a result, the evidence in favor of  $H_1$  is even stronger than that suggested by a 0.05 level of significance.

**Example 160** *A manufacturer of sports equipment has developed a new synthetic fishing line that the company claims has a mean breaking strength of 8 kilograms with a standard deviation*

of 0.5 kilogram. Test the hypothesis that  $\mu = 8$  kilograms against the alternative that  $\mu \neq 8$  kilograms if a random sample of 50 lines is tested and found to have a mean breaking strength of 7.8 kilograms. Use a 0.01 level of significance.

**Solution 161** 1.  $H_0 : \mu = 8$  kilograms.

2.  $H_1 : \mu \neq 8$  kilograms.

3.  $\alpha = 0.01$ .

4. Critical region:  $z > z_{\alpha/2}$  and  $z < -z_{\alpha/2}$ , where  $z = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$  and  $z_{\alpha/2} = 2.575$ .

5. Computations:  $\bar{x} = 7.8$ ,  $\sigma = 0.5$ ,  $n = 50$ , hence  $z = (\bar{x} - \mu_0)/(\sigma/\sqrt{n}) = (7.8 - 8)/(0.5/\sqrt{50}) = -2.83$

6. Decision: since  $z = -2.83 < -2.575$ , hence reject  $H_0$  and conclude that the average breaking strength is not equal to 8 but is, in fact, less than 8 kilograms.

Since the test in this example is two tailed, the desired  $P$ -value is twice the area of the left of  $z = -2.83$ . Therefore, using standard normal table, we have

$$P\text{-value} = P(Z > |2.83|) = 2P(Z > 2.83) = 0.0046 < 0.01$$

which allows us to reject the null hypothesis that  $\mu = 8$  kilograms at a level of significance smaller than 0.01.

### Tests on a Single Sample (Variance Unknown)

**Result 162 (Single Mean (Variance Unknown))** For the two-sided hypothesis

$$H_0 : \mu = \mu_0,$$

$$H_1 : \mu \neq \mu_0$$

we reject  $H_0$  at significance level  $\alpha$  when the computed  $t$ -statistic

$$t = (\bar{x} - \mu_0)/(s/\sqrt{n})$$

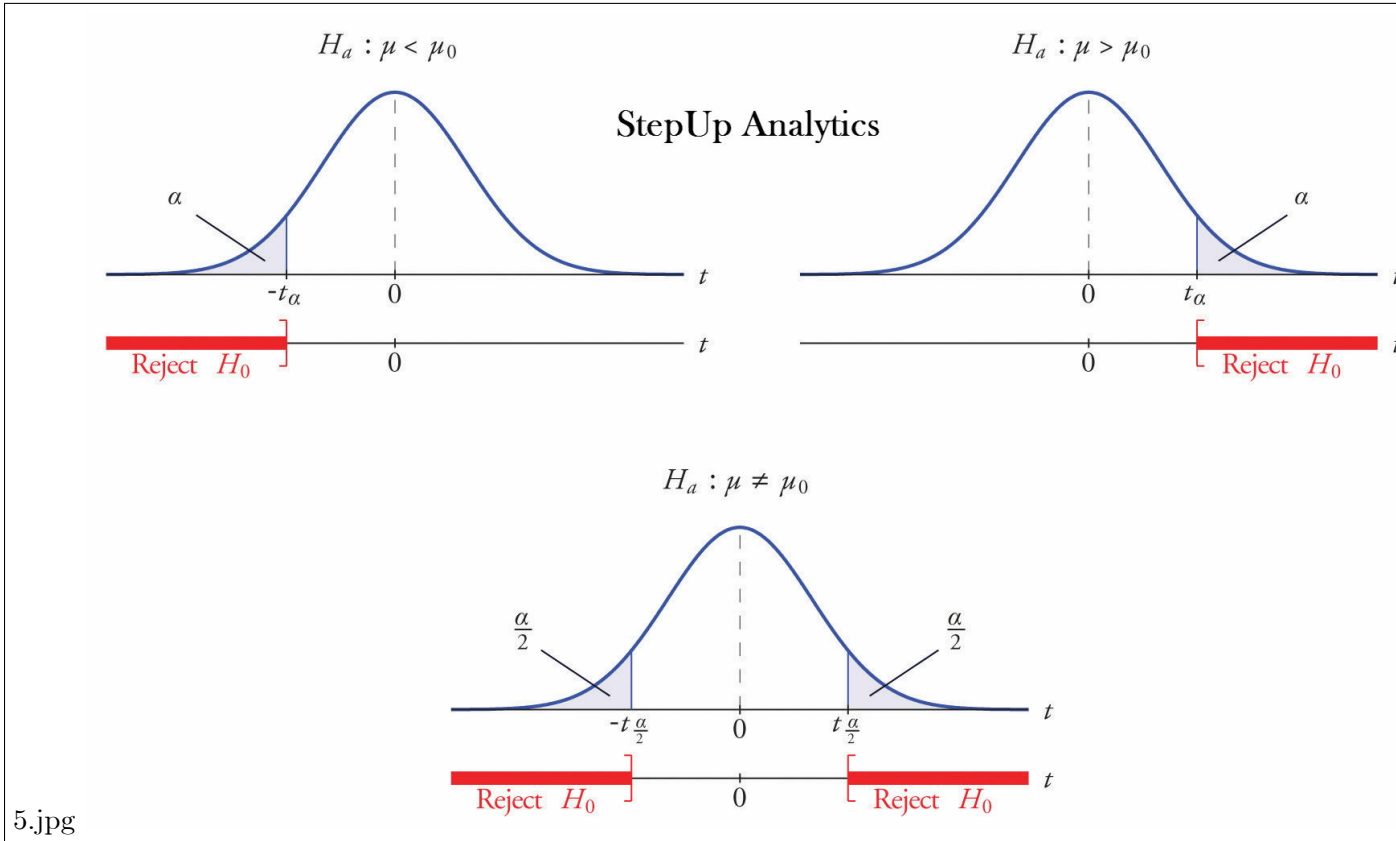


Figure 5-1: Figure 5.1

exceeds  $t_{\alpha/2, n-1}$  or is less than  $-t_{\alpha/2, n-1}$ .

For  $H_1 : \mu > \mu_0$ , rejection results when  $t > t_{\alpha, n-1}$ . For  $H_1 : \mu < \mu_0$ , the critical region is given by  $t < -t_{\alpha, n-1}$ .

**Example 163** *The Edison Electric Institute has published figures on the number of kilowatt hours used annually by various home appliances. It is claimed that a vacuum cleaner uses an average of 46 kilowatt hours per year. If a random sample of 12 homes indicates that vacuum cleaners use an average of 42 kilowatt hours per year with a standard deviation of 11.9 kilowatt hours, does this suggest at the 0.05 level of significance that vacuum cleaners use, on average, less than 46 kilowatt hours annually? Assume the population of kilowatt hours to be normal.*

**Solution 164** 1.  $H_0 : \mu = 46$  kilowatt hours.

2.  $H_1 : \mu < 46$  kilowatt hours.

3.  $\alpha = 0.05$ .

4. Critical region:  $t < -t_{\alpha, n-1}$ , where  $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$  with 11 degrees of freedom and  $t_{\alpha, n-1} = 1.796$ .

5. Computations:  $\bar{x} = 42$  kilowatt hours,  $s = 11.9$  kilowatt hours, and  $n = 12$ .

Hence,  $t = (42 - 46)/(11.9/\sqrt{12}) = -1.16$ .

6. Since  $t > -1.796$ , we do not reject  $H_0$  and conclude that the average number of kilowatt hours used annually by home vacuum cleaners is not significantly less than 46.

Also  $P\text{-value} = \Pr(T < -1.16) = \Pr(T > 1.16) \approx 0.135$ .

## 5.2 Two Samples: Tests on Two Means

The two-sided hypothesis on two means can be written generally as  $H_0 : \mu_1 = \mu_2$ . For  $\sigma_1$  and  $\sigma_2$  known, the test statistic is given by

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

That is, reject  $H_0$  in favor of  $H_1 : \mu_1 \neq \mu_2$  if  $z > z_{\alpha/2}$ , and  $z < -z_{\alpha/2}$ .

One-tailed critical regions are used in the case of the one-sided alternatives. The reader should, as before, study the test statistic and be satisfied that for, say,  $H_1 : \mu_1 > \mu_2$ , the signal favoring  $H_1$  comes from large values of  $z$ . Thus, the upper-tailed critical region applies.

### Unknown But Equal Variances

If we assume that both distributions are normal and that  $\sigma_1 = \sigma_2 = \sigma$ , the two-sample t-test may be used. The test statistic is given by the following test procedure.

**Result 165 (Two-Sample Pooled t-Test)** *For the two-sided hypothesis*

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2,$$

*we reject  $H_0$  at significance level  $\alpha$  when the computed t-statistic*

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{1/n_1 + 1/n_2}}$$

*where*

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

*exceeds  $t_{\alpha/2, n_1+n_2-2}$  or is less than  $-t_{\alpha/2, n_1+n_2-2}$ .*

For  $H_1 : \mu_1 > \mu_2$ , reject  $H_0 : \mu_1 = \mu_2$  when  $t > t_{\alpha, n_1+n_2-2}$ . For  $H_1 : \mu_1 < \mu_2$ , reject  $H_0 : \mu_1 = \mu_2$  when  $t < -t_{\alpha, n_1+n_2-2}$ .

**Example 166** *An experiment was performed to compare the abrasive wear of two materials. Twelve pieces of material 1 were tested and ten pieces of material 2 were similarly tested. The samples of material 1 gave an average (coded) wear of 85 units with a sample standard deviation of 4, while the samples of material 2 gave an average of 83 with a sample standard deviation of 5. Can we conclude at the 0.05 level of significance that the abrasive wear of material 1 exceeds that of material 2? Assume the populations to be approximately normal with equal variances.*

**Solution 167** *Let  $\mu_1$  and  $\mu_2$  represent the population means of the abrasive wear for material 1 and material 2, respectively.*

1.  $H_0 : \mu_1 = \mu_2$ .
2.  $H_1 : \mu_1 > \mu_2$ .
3.  $\alpha = 0.05$ .
4. Critical region:  $t > t_{\alpha, n_1+n_2-2}$ , where  $t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{1/n_1 + 1/n_2}}$  with  $\nu = 20$  degrees of freedom and  $t_{\alpha, n_1+n_2-2} = t_{0.05, 20} = 1.725$ .
5. Computations:

$$\begin{aligned}\bar{x}_1 &= 58 & s_1 &= 4 & n_1 &= 12 \\ \bar{x}_2 &= 81 & s_2 &= 5 & n_2 &= 10\end{aligned}$$

Hence

$$\begin{aligned}s_p &= \sqrt{\frac{(11)(16) + (9)(25)}{12 + 10 - 2}} = 4.478 \\ t &= \frac{(85 - 83)}{4.478 \sqrt{1/12 + 1/10}} = 1.04 < 1.725 \\ P\text{-value} &= \Pr(T > 1.04) \approx 0.16. (\text{See Table A.4.})\end{aligned}$$

6. Decision: Do not reject  $H_0$ . We are unable to conclude that the abrasive wear of material 1 exceeds that of material 2.

### Paired Observations

Testing of two means can be accomplished when data are in the form of paired observations, as discussed in Chapter 4. The statistical test for two means  $\mu_1$  and  $\mu_2$  in the situation with paired observations is based on the random variable

$$T = \frac{\bar{D} - \mu_D}{S_d / \sqrt{n}}$$

where  $\bar{D}$  and  $S_d$  are random variables representing the sample mean and standard deviation of the differences of the observations in the experimental units. As in the case of the pooled t-test, the assumption is that the observations from each population are normal. This two-sample problem is essentially reduced to a one-sample problem by using the computed differences

$d_1, d_2, \dots, d_n$ . Critical regions are constructed using the t-distribution with  $n - 1$  degrees of freedom.

**Example 168 Blood Sample Data:** *A study was conducted to examine the influence of the drug succinylcholine on the circulation levels of androgens in the blood. Blood samples were taken from wild, free-ranging deer immediately after they had received an intramuscular injection of succinylcholine administered using darts and a capture gun. A second blood sample was obtained from each deer 30 minutes after the first sample, after which the deer was released. The levels of androgens at time of capture and 30 minutes later, measured in nanograms per milliliter (ng/mL), for 15 deer are given in Table 6.2. Assuming that the populations of androgen levels at time of injection and 30 minutes later are normally distributed, test at the 0.05 level of significance whether the androgen concentrations are altered after 30 minutes.*

Deer	Androgen (ng/mL)		$d_i$
	At Time of Injection	30 Minutes after Injection	
1	2.76	7.02	4.26
2	5.18	3.10	-2.08
3	2.68	5.44	2.76
4	3.05	3.99	0.94
5	4.10	5.21	1.11
6	7.05	10.26	3.21
7	6.60	13.91	7.31
8	4.79	18.53	13.74
9	7.39	7.91	0.52
10	7.30	4.85	-2.45
11	11.78	11.10	-0.68
12	3.90	3.74	-0.16
13	26.00	94.03	68.03
14	67.48	94.03	26.55
15	17.04	41.70	24.66



**Solution 169** Let  $\mu_1$  and  $\mu_2$  be the average androgen concentration at the time of injection and 30 minutes later, respectively. We proceed as follows:

1.  $H_0 : \mu_1 = \mu_2$  or  $\mu_D = \mu_1 - \mu_2 = 0$ .
2.  $H_1 : \mu_1 \neq \mu_2$  or  $\mu_D = \mu_1 - \mu_2 \neq 0$ .
3.  $\alpha = 0.05$ .
4. Critical region:  $t < -t_{\alpha/2, n-1}$  and  $t > t_{\alpha/2, n-1}$ , where  $t = \frac{\bar{d}}{s_D/\sqrt{n}}$  with  $\nu = 14$  degrees of freedom and  $t_{\alpha/2, n-1} = 2.145$ .
5. Computations: The sample mean and standard deviation for the  $d_i$  are

$$\bar{d} = 9.848 \quad \text{and} \quad s_d = 18.474$$

Therefore

$$t = \frac{9.848}{18.474/\sqrt{15}} = 2.06$$

6. Hence  $-2.145 < t = 2.06 < 2.145$ . Though the  $t$ -statistic is not significant at the 0.05 level, from Table A.4,

$$P = P(|T| > 2.06) \approx 0.06$$

As a result, there is no evidence that there is a difference in mean circulating levels of androgen.

### 5.3 One Sample: Test on a Single Proportion

Tests of hypotheses concerning proportions are required in many areas. We now consider the problem of testing the hypothesis that the proportion of successes in a binomial experiment equals some specified value. That is, we are testing the null hypothesis  $H_0$  that  $p = p_0$ , where  $p$  is the parameter of the binomial distribution. The alternative hypothesis may be one of the usual one-sided or two-sided alternatives:

$$p < p_0 \quad p > p_0 \quad \text{or} \quad p \neq p_0$$

We know that if  $np_0 \geq 5$  and  $n(1 - p_0) \geq 5$ , then the random variable  $\hat{P}$  is approximately a normal distribution with mean  $p_0$  and standard deviation  $\sigma_{\hat{P}} = \sqrt{p_0(1 - p_0)/n}$ . The **z-value**

for testing  $\mathbf{p} = \mathbf{p}_0$  is given by

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Hence, for a two-tailed test at the  $\alpha$ -level of significance, the critical region is  $z < -z_{\alpha/2}$  or  $z > z_{\alpha/2}$ . For the one-sided alternative  $p < p_0$ , the critical region is  $z < -z_{\alpha}$ , and for the alternative  $p > p_0$ , the critical region is  $z > z_{\alpha}$ .

**Example 170** *A commonly prescribed drug for relieving nervous tension is believed to be only 60% effective. Experimental results with a new drug administered to a random sample of 100 adults who were suffering from nervous tension show that 70 received relief. Is this sufficient evidence to conclude that the new drug is superior to the one commonly prescribed? Use a 0.05 level of significance.*

**Solution 171** 1.  $H_0 : p = 0.6$ .

2.  $H_1 : p > 0.6$ .

3.  $\alpha = 0.05$ .

4. Critical region:  $Z > z_{\alpha}$ , where  $z_{\alpha} = 1.645$ . Then, the critical region:  $z > 1.645$ .

5. Computations:  $x = 70, n = 100, \hat{p} = 70/100 = 0.7$ , and

$$\begin{aligned} z &= \frac{0.7 - 0.6}{\sqrt{\frac{(0.6)(0.4)}{100}}} = 2.04 \\ z &= 2.04 > 1.645 \end{aligned}$$

$$P\text{-value} = \Pr(Z > 2.04) < 0.0207.$$

6. Decision: Reject  $H_0$  and conclude that the new drug is superior.

## 5.4 Two Samples: Tests on Two Proportions

Situations often arise where we wish to test the hypothesis that two proportions are equal. That is, we are testing  $p_1 = p_2$  against one of the alternatives  $p_1 < p_2$ ,  $p_1 > p_2$ , or  $p_1 \neq p_2$ . The statistic on which we base our decision is the random variable  $\hat{P}_1 - \hat{P}_2$ . When  $H_0 : p_1 = p_2$  ( $= p$ ) is true, we know that

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{pq(1/n_1 + 1/n_2)}}$$

To compute a value of  $Z$ , however, we must estimate the parameters  $p$  and  $q$  that appear in the radical. Under  $H_0$ , both  $\hat{P}_1$  and  $\hat{P}_2$  are estimators of  $p$ . we use the pooled estimate of the proportion  $p$ , which is

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

where  $x_1$  and  $x_2$  are the numbers of successes in each of the two samples. Substituting  $\hat{p}$  for  $p$  and  $\hat{q} = 1 - \hat{p}$  for  $q$ , the z-value for testing  $p_1 = p_2$  is determined from the formula

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}}$$

The critical regions for the appropriate alternative hypotheses are set up as before, using critical points of the standard normal curve.

**Example 172** *A vote is to be taken among the residents of a town and the surrounding county to determine whether a proposed chemical plant should be constructed. To determine if there is a significant difference in the proportions of town voters and county voters favoring the proposal, a poll is taken. If 120 of 200 town voters favor the proposal and 240 of 500 county residents favor it, would you agree that the proportion of town voters favoring the proposal is higher than the proportion of county voters? Use an  $\alpha = 0.05$  level of significance.*

**Solution 173** *Let  $p_1$  and  $p_2$  be the true proportions of voters in the town and county, respectively, favoring the proposal.  $\hat{p}_1 = x_1/n_1 = 120/200 = 0.6$ ,  $\hat{p}_2 = x_2/n_2 = 240/500 = 0.48$ , and the pooled estimate  $\hat{p} = (x_1 + x_2)/(n_1 + n_2) = (120 + 240)/(200 + 500) = 0.51$ .*

1.  $H_0 : p_1 = p_2$ .
2.  $H_1 : p_1 > p_2$ .
3.  $\alpha = 0.05$ .
4. *The test statistic*

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}} = \frac{0.60 - 0.48}{(0.51)(0.49)(1/200 + 1/500)} = 2.9$$

5. *Critical region:  $z > 1.645$ .  $P$ -value =  $P(Z > 2.9) = 0.0019$ .*
6. *Decision: Reject  $H_0$  and agree that the proportion of town voters favouring the proposal is*

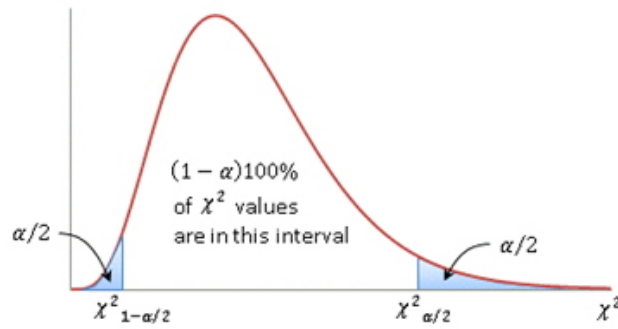


Figure 5-2: Figure 5.2

higher than the proportion of county voters.

## 5.5 One- and Two-Sample Tests Concerning Variances

In this section, we are concerned with testing hypotheses concerning population variances or standard deviations. Let us first consider the problem of testing the null hypothesis  $H_0$  that the population variance  $\sigma^2$  equals a specified value  $\sigma_0^2$  against one of the usual alternatives  $\sigma^2 < \sigma_0^2$ ,  $\sigma^2 > \sigma_0^2$ , or  $\sigma^2 \neq \sigma_0^2$ . If we assume that the distribution of the population being sampled is normal, the chi-squared value for testing  $\sigma^2 = \sigma_0^2$  is given by

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

where  $n$  is the sample size,  $s^2$  is the sample variance, and  $\sigma_0^2$  is the value of  $\sigma^2$  given by the null hypothesis. If  $H_0$  is true,  $\chi^2$  is a value of the chi-squared distribution with  $\nu = n - 1$  degrees of freedom. Hence, for a two-tailed test at the  $\alpha$ -level of significance, the critical region is  $\chi^2 < \chi_{1-\alpha/2}^2$  or  $\chi^2 > \chi_{\alpha/2}^2$  (see figure 5.2). For the one-sided alternative  $\sigma^2 < \sigma_0^2$ , the critical region is  $\chi^2 < \chi_{1-\alpha}^2$ , and for the one-sided alternative  $\sigma^2 > \sigma_0^2$ , the critical region is  $\chi^2 > \chi_{\alpha}^2$ .

**Example 174** *A manufacturer of car batteries claims that the life of the company's batteries is approximately normally distributed with a standard deviation equal to 0.9 year. If a random sample of 10 of these batteries has a standard deviation of 1.2 years, do you think that  $\sigma > 0.9$  year? Use a 0.05 level of significance.*

**Solution 175** 1.  $H_0 : \sigma^2 = 0.81$ .

2.  $H_1 : \sigma^2 > 0.81$ .

3.  $\alpha = 0.05$ .

4. *Critical region:* The null hypothesis is rejected when  $\chi^2 > 16.919$ , where  $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$ , with  $\nu = 9$  degrees of freedom.

5. *Computations:*  $s^2 = 1.44$ ,  $n = 10$ , and

$$\chi^2 = \frac{(9)(1.44)}{0.81} = 16.0, P \approx 0.07.$$

6. *Decision:* The  $\chi^2$ -statistic is not significant at the 0.05 level. However, based on the  $P$ -value 0.07, there is evidence that  $\sigma > 0.9$ .

Now let us consider the problem of testing the equality of the variances  $\sigma_1^2$  and  $\sigma_2^2$  of two populations. That is, we shall test the null hypothesis  $H_0$  that  $\sigma_1^2 = \sigma_2^2$  against one of the usual alternatives  $\sigma_1^2 < \sigma_2^2$ ,  $\sigma_1^2 > \sigma_2^2$ , or  $\sigma_1^2 \neq \sigma_2^2$ . For independent random samples of sizes  $n_1$  and  $n_2$ , respectively, from the two populations, the  $f$ -value for testing  $\sigma_1^2 = \sigma_2^2$  is the ratio

$$f = \frac{s_1^2}{s_2^2}$$

where  $s_1^2$  and  $s_2^2$  are the variances computed from the two samples. If the two populations are approximately normally distributed and the null hypothesis is true, then the ratio  $f = s_1^2/s_2^2$  is a value of the  $F$ -distribution with  $\nu_1 = n_1 - 1$  and  $\nu_2 = n_2 - 1$  degrees of freedom. Therefore, the critical regions of size  $\alpha$  corresponding to the one-sided alternatives  $\sigma_1^2 < \sigma_2^2$  and  $\sigma_1^2 > \sigma_2^2$  are, respectively,  $f < f_{1-\alpha}(\nu_1, \nu_2)$  and  $f > f_{\alpha}(\nu_1, \nu_2)$ . For the two-sided alternative  $\sigma_1^2 \neq \sigma_2^2$ , the critical region is  $f < f_{1-\alpha/2}(\nu_1, \nu_2)$  or  $f > f_{\alpha/2}(\nu_1, \nu_2)$ .

**Example 176** *In testing for the difference in the abrasive wear of the two materials in Example 166, we assumed that the two unknown population variances were equal. Were we justified in making this assumption? Use a 0.10 level of significance.*

**Solution 177** *Let  $\sigma_1^2$  and  $\sigma_2^2$  be the population variances for the abrasive wear of material 1 and material 2, respectively.*

1.  $H_0 : \sigma_1^2 = \sigma_2^2$ .

2.  $H_1 : \sigma_1^2 \neq \sigma_2^2$ .

3.  $\alpha = 0.10$ .

4. *Critical region:* We have  $f_{0.05}(11, 9) = 3.11$ , and, by using Theorem 99, we find  $f_{0.95}(11, 9) = \frac{1}{f_{0.05}(9, 11)} = 0.34$ . Therefore, the null hypothesis is rejected when  $f < 0.34$  or  $f > 3.11$ , where  $f = s_1^2/s_2^2$  with  $\nu_1 = 11$  and  $\nu_2 = 9$  degrees of freedom.

5. *Computations:*  $s_1^2 = 16$ ,  $s_2^2 = 25$ , hence  $f = 16/25 = 0.64$ .

6. *Decision:* Do not reject  $H_0$ . Conclude that there is insufficient evidence that the variances differ.

# Chapter 6

## Chi square tests

### 6.1 Goodness-of-Fit Test

we consider a test to determine if a population has a specified theoretical distribution. The test is based on how good a fit we have between the frequency of occurrence of observations in an observed sample and the expected frequencies obtained from the hypothesized distribution. To illustrate, we consider the tossing of a die. We hypothesize that the die is honest, which is equivalent to testing the hypothesis that the distribution of outcomes is the discrete uniform distribution

$$f(x) = \frac{1}{6}, x, 2, 3, 4, 5, 6$$

Suppose that the die is tossed 120 times and each outcome is recorded. Theoretically, if the die is balanced, we would expect each face to occur 20 times. The results are given in the table 6.1.

Face	1	2	3	4	5	6
Observed	20	22	17	18	19	24
Expected	20	20	20	20	20	20

By comparing the observed frequencies with the corresponding expected frequencies, the hypothesis,  $H_0$  : the die is fair, should be rejected or not. A **goodness-of-fit test** between

observed and expected frequencies is based on the quantity

$$\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

where  $\chi^2$  is a value of a random variable whose sampling distribution is approximated very closely by the chi-squared distribution with  $\nu = k - 1$  degrees of freedom. The symbols  $o_i$  and  $e_i$  represent the observed and expected frequencies, respectively, for the  $i$ th cell.

If the observed frequencies are close to the corresponding expected frequencies, the  $\chi^2$ -value will be small, indicating a good fit. If the observed frequencies differ considerably from the expected frequencies, the  $\chi^2$ -value will be large and the fit is poor. A good fit leads to the acceptance of  $H_0$ , whereas a poor fit leads to its rejection. The critical region will, therefore, fall in the right tail of the chi-squared distribution. For a level of significance equal to  $\alpha$ , we find the critical value  $\chi_\alpha^2$  from Table A.5, and then  $\chi^2 > \chi_\alpha^2$  constitutes the critical region. The decision criterion described here should not be used unless each of the expected frequencies is at least equal to 5. This restriction may require the combining of adjacent cells, resulting in a reduction in the number of degrees of freedom.

From Table 6.1, we find the  $\chi^2$ -value to be

$$\begin{aligned} \chi^2 &= \frac{(20-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(18-20)^2}{20} \\ &\quad + \frac{(19-20)^2}{20} + \frac{(24-20)^2}{20} = 1.7 \end{aligned}$$

Using chi-squared table, we find  $\chi_{0.05}^2 = 11.070$  for  $\nu = 5$  degrees of freedom. Since 1.7 is less than the critical value, we fail to reject  $H_0$ . We conclude that there is insufficient evidence that the die is not balanced.

A second example is to test the hypothesis that the frequency distribution of battery lives given in Table 6.2 may be approximated by a normal distribution with mean  $\mu = 3.5$  and standard deviation  $\sigma = 0.7$ . The expected frequencies for the 7 classes (cells), listed in Table 6.2, are obtained by computing the areas under the hypothesized normal curve that fall between the various class boundaries.

Table 6.2: Frequency Distribution of Battery Life.

Class Boundaries	Frequency
------------------	-----------



1.45-1.95	2
1.95-2.45	1
2.45-2.95	4
2.95-3.45	15
3.45-3.95	10
3.95-4.45	5
4.45-4.95	3

The z-values corresponding to the boundaries of the first class are  $z_1 = (1.45 - 3.5)/0.7 = -2.93$  and  $z_2 = (1.95 - 3.5)/0.7 = -2.21$ . From standard normal table, we find the area between  $z_1 = -2.93$  and  $z_2 = -2.21$  to be area equal to  $\Pr(-2.93 < Z < -2.21) = \Pr(Z < -2.21) - \Pr(Z < -2.93) = 0.0136 - 0.0017 = 0.0119$ . Hence, the expected frequency for the first class is  $e_1 = 40(0.0119) = 0.5$ . Similarly, we get

$$e_2 = 40[P((1.95 - 3.5)/0.7 < Z < (2.45 - 3.5)/0.7)] = 40[P(-2.21 < Z < -1.5)] = 2.1$$

$$e_3 = 40[P((2.45 - 3.5)/0.7 < Z < (2.95 - 3.5)/0.7)] = 40[P(-1.5 < Z < -0.79)] = 5.9$$

$$e_4 = 40[P((2.95 - 3.5)/0.7 < Z < (3.45 - 3.5)/0.7)] = 40[P(-0.79 < Z < -0.07)] = 10.3$$

$$e_5 = 40[P((3.45 - 3.5)/0.7 < Z < (3.95 - 3.5)/0.7)] = 40[P(-0.07 < Z < 0.64)] = 10.7$$

$$e_6 = 40[P((3.95 - 3.5)/0.7 < Z < (4.45 - 3.5)/0.7)] = 40[P(0.64 < Z < 1.36)] = 7.0$$

$$e_7 = 40[P((4.45 - 3.5)/0.7 < Z < (4.95 - 3.5)/0.7)] = 40[P(1.36 < Z < 2.07)] = 3.5$$

It is customary to round these frequencies to one decimal. Note that we have to combined adjacent classes in Table 16.2 where the expected frequencies are less than 5 (a rule of thumb in the goodness-of-fit test). Consequently, the total number of intervals is reduced from 7 to 4, resulting in  $\nu = 3$  degrees of freedom. The  $\chi^2$ -value is then given by

$$\chi^2 = \frac{(7-8.5)^2}{8.5} + \frac{(15-10.3)^2}{10.3} + \frac{(10-10.7)^2}{10.7} + \frac{(8-10.5)^2}{10.5} = 3.05$$

Since the computed  $\chi^2$ -value is less than  $\chi_{0.05}^2 = 7.815$  for 3 degrees of freedom, we have no reason to reject the null hypothesis and conclude that the normal distribution with  $\mu = 3.5$  and  $\sigma = 0.7$  provides a good fit for the distribution of battery lives.

## 6.2 Test for Independence (Categorical Data)

The chi-squared test procedure discussed in Section 6.1 can also be used to test the hypothesis  $H_0$  of independence of two variables of classification. A test of independence tests the null hypothesis that there is no association between the two variables in a contingency table where the data is all drawn from one population.

Suppose that we wish to determine whether the opinions of the voting residents of the state of Illinois concerning a new tax reform are independent of their levels of income. Members of a random sample of 1000 registered voters from the state of Illinois are classified as to whether they are in a low, medium, or high income bracket and whether or not they favor the tax reform. The observed frequencies are presented in Table 6.2.

Tax Reform	Low	Medium	High	Total
For	182	213	203	598
Against	154	138	110	402
Total	336	351	313	1000

To find these expected frequencies, let us define the following events:

L: A person selected is in the low-income level.

M: A person selected is in the medium-income level.

H: A person selected is in the high-income level.

F: A person selected is for the tax reform.

A: A person selected is against the tax reform.

We have

$P(L) = 336/1000, P(M) = 351/1000, P(H) = 313/1000, P(F) = 598/1000, P(A) = 402/1000.$

Recall that two events A and B are said independent when  $\Pr(A \cap B) = \Pr(A)\Pr(B)$ . If  $H_0$  is true and the two variables are independent, we should have

$$\Pr(L \cap F) = \Pr(L) \Pr(F) = \left( \frac{336}{1000} \right) \left( \frac{598}{1000} \right)$$

$$\Pr(L \cap A) = \Pr(L) \Pr(A) = \left( \frac{336}{1000} \right) \left( \frac{402}{1000} \right)$$

$$\Pr(M \cap F) = \Pr(M) \Pr(F) = \left( \frac{351}{1000} \right) \left( \frac{598}{1000} \right)$$

$$\Pr(M \cap A) = \Pr(M) \Pr(A) = \left( \frac{351}{1000} \right) \left( \frac{402}{1000} \right)$$

$$\Pr(H \cap F) = \Pr(H) \Pr(F) = \left( \frac{313}{1000} \right) \left( \frac{598}{1000} \right)$$

$$\Pr(H \cap A) = \Pr(H) \Pr(A) = \left( \frac{313}{1000} \right) \left( \frac{402}{1000} \right)$$

The expected frequencies are obtained by multiplying each cell probability by the total number of observations 1000.

The general rule for obtaining the expected frequency of any cell is given by the following formula:

$$\text{expected frequency} = \frac{(\text{column total}) \times (\text{row total})}{\text{grand total}}$$

A simple formula providing the correct number of degrees of freedom is

$$\nu = (c - 1) \times (r - 1)$$

Table 6.3: Observed and Expected Frequencies

	Income Level			
Tax Reform	Low	Medium	High	Total
For	182 (200.9)	213 (209.9)	203 (187.2)	598
Against	154 (135.1)	138 (141.1)	110 (125.8)	402
Total	336	351	313	1000

Hence, the expected frequency for the first class is

$$e_{11} = (336 \times 598)/1000 = 200.9.$$

Here  $\nu = (2 - 1)(3 - 1) = 2$  degrees of freedom. If  $\chi^2 > \chi_\alpha^2$ , reject the null hypothesis of independence at the  $\alpha$ -level of significance; otherwise, fail to reject the null hypothesis.

1.  $H_0$ : the two random variables (voter's opinion concerning the tax reform and his or her level of income) are independent.

2.  $H_1$ : voter's opinion concerning the tax reform and his or her level of income are not independent.

3.  $\alpha = 0.05$ .

4. The test statistic:

$$\chi^2 = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \frac{(182-200.9)^2}{200.9} + \frac{(213-209.9)^2}{209.9} + \frac{(203-187.2)^2}{187.2} \\ + \frac{(154-135.1)^2}{135.1} + \frac{(138-141.1)^2}{141.1} + \frac{(110-125.8)^2}{125.8} = 7.85, \text{ P-value} \simeq 0.02.$$

5. From chi-square table we find that  $\chi_{0.05}^2 = 5.991$  for  $\nu = (2 - 1)(3 - 1) = 2$  degrees of freedom. The null hypothesis is rejected and we conclude that a voter's opinion concerning the tax reform and his or her level of income are not independent.

### 6.3 Test for Homogeneity

Now, rather than test for independence, this test determines if two or more populations have the same distribution of a single categorical variable. The key difference from the test of independence is that there are multiple populations that the data is drawn from. Suppose, for example, that we decide in advance to select 200 Democrats, 150 Republicans, and 150 Independents from the voters of the state of North Carolina and record whether they are for a proposed abortion law, against it, or undecided. We assume that the row totals in table 6.4 were random, while the column totals were presumably fixed in advance, since they represented numbers of voters sampled from different political affiliations.

Table 6.4: Observed Frequencies

Abortion Law	Political Affiliation			Total
	Democrat	Republican	Independent	
For	82	70	62	214
Against	93	62	67	222
Undecided	25	18	21	64
Total	200	150	150	500

we want to test the hypothesis that the proportions of Democrats, Republicans, and Independents favoring the abortion law are the same; the proportions of each political affiliation against the law are the same; and the proportions of each political affiliation that are undecided are

the same. Then the null hypothesis is

$$H_0 : \text{For each row } i, p_{i1} = \dots = p_{ic}$$

The alternative hypothesis  $H_1$  is that at least one of the null hypothesis statements is false. Such a test is called a test for homogeneity.

If homogeneity true, we should have

$$\Pr(\text{For} | \text{Democrat}) = \Pr(\text{For} | \text{Republican}) = \Pr(\text{For} | \text{Independent})$$

$$\begin{aligned} \text{But } \Pr(\text{For}) &= \Pr(\text{For} | \text{Democrat}) \Pr(\text{Democrat}) + \Pr(\text{For} | \text{Republican}) \Pr(\text{Republican}) \\ &+ \Pr(\text{For} | \text{Independent}) \Pr(\text{Independent}) = \Pr(\text{For} | \text{Democrat}) \end{aligned}$$

Consequently

$$\Pr(\text{For} | \text{Democrat}) = \Pr(\text{For} | \text{Republican}) = \Pr(\text{For} | \text{Independent}) = \Pr(\text{For}) \text{ estimated by } \frac{214}{500}$$

But  $\Pr(\text{For} \cap \text{Democrat}) = \Pr(\text{For} | \text{Democrat}) \Pr(\text{Democrat}) \simeq \frac{214}{500} \frac{200}{500}$ . Hence  $e_{11} = \left(\frac{214}{500}\right) \left(\frac{200}{500}\right) (500) = \frac{214 \times 200}{500} = 85.6$ . The expected cell frequencies is then obtained by multiplying the corresponding row and column totals and then dividing by the grand total. The analysis then proceeds using the same chi-squared statistic as before. Use of the chi-square distribution is appropriate whenever the expected values are all greater than or equal to 5.

1.  $H_0$ : For each opinion, the proportions of Democrats, Republicans, and Independents are the same.

2.  $H_1$ : For at least one opinion, the proportions of Democrats, Republicans, and Independents are not the same.

3.  $\alpha = 0.05$ .

4. Critical region:  $\chi^2 > \chi_\alpha^2$  with  $\nu = (3-1)(3-1) = 4$  degrees of freedom. Then the critical region:  $\chi^2 > 9.488$ .

5. Computations: The observed and expected cell frequencies are displayed in Table 6.5.

Table 6.5: Observed and Expected Frequencies

Political Affiliation

Abortion Law	Democrat	Republican	Independent	Total
For	82 (85.6)	70 (64.2)	62 (64.2)	214
Against	93 (88.8)	62 (66.6)	67 (66.6)	222
Undecided	25 (25.6)	18 (19.2)	21 (19.2)	64
Total	200	150	150	500

We obtain that  $\chi^2 = 1.53$ .

6. Decision: Do not reject  $H_0$ . There is insufficient evidence to conclude that the proportions of Democrats, Republicans, and Independents differ for each stated opinion.

## Chapter 7

# Simple linear regression and correlation

### 7.1 Joint Probability Distributions

If  $X$  and  $Y$  are two discrete random variables, the probability distribution for their simultaneous occurrence can be represented by a function with values  $f(x, y)$  for any pair of values  $(x, y)$  within the range of the random variables  $X$  and  $Y$ . It is customary to refer to this function as the joint probability distribution of  $X$  and  $Y$ .

Hence, in the discrete case,

$$f(x, y) = \Pr(X = x, Y = y)$$

**Definition 178** *The function  $f(x, y)$  is a joint probability distribution or probability mass function of the discrete random variables  $X$  and  $Y$  if*

1.  $f(x, y) \geq 0$  for all  $(x, y)$ ,
2.  $\sum_x \sum_y f(x, y) = 1$ ,
3.  $\Pr(X = x, Y = y) = f(x, y)$

When  $X$  and  $Y$  are continuous random variables, we have

**Definition 179** *The function  $f(x, y)$  is a joint density function of the continuous random*

variables  $X$  and  $Y$  if

1.  $f(x, y) \geq 0$  for all  $(x, y)$ ,
2.  $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$ ,
3.  $\Pr[(X, Y) \in A] = \int \int_A f(x, y) dx dy$ , for any region  $A$  in the  $xy$  plane.

**Example 180** Let a joint density function of  $X$  and  $Y$  given by

$$f(x, y) = \begin{cases} \frac{2}{5}(2x + 3y), & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0, & \text{elsewhere.} \end{cases}$$

- 1) Verify condition 2 of Definition 179.
- 2) Find  $\Pr[(X, Y) \in A]$ , where  $A = \{(x, y) \mid 0 < x < \frac{1}{2}, \frac{1}{4} < y < \frac{1}{2}\}$

**Definition 181** The marginal distributions of  $X$  alone and of  $Y$  alone are

$$g(x) = \sum_y f(x, y) \text{ and } h(y) = \sum_x f(x, y)$$

for the discrete case, and

$$g(x) = \int_{-\infty}^{\infty} f(x, y) dy \text{ and } h(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

for the continuous case.

**Example 182** Find  $g(x)$  and  $h(y)$  for the joint density function of Example 180.

**Definition 183** Let  $X$  and  $Y$  be two random variables, discrete or continuous, with joint probability distribution  $f(x, y)$  and marginal distributions  $g(x)$  and  $h(y)$ , respectively. The random variables  $X$  and  $Y$  are said to be statistically independent if and only if

$$f(x, y) = g(x)h(y)$$

for all  $(x, y)$ .



**Definition 184** Let  $X$  be a random variable with probability distribution  $f(x)$  and mean  $\mu$ . The variance of  $X$  is

$$\begin{aligned}\sigma^2 &= E[(X - \mu)^2] = \sum_x (x - \mu)^2 f(x), \text{ if } X \text{ is discrete, and} \\ \sigma^2 &= E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx, \text{ if } X \text{ is continuous.}\end{aligned}$$

The positive square root of the variance,  $\sigma$ , is called the standard deviation of  $X$ .

**Example 185** The weekly demand for a drinking-water product, in thousands of liters, from a local chain of efficiency stores is a continuous random variable  $X$  having the probability density

$$f(x) = \begin{cases} 2(x - 1), & 1 < x < 2, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the mean and variance of  $X$ .

**Definition 186** Let  $X$  be a random variable with probability distribution  $f(x)$ . The variance of the random variable  $g(X)$  is

$$\sigma_{g(X)}^2 = E[(g(X) - \mu_{g(X)})^2] = \sum_x (g(x) - \mu)^2 f(x)$$

if  $X$  is discrete, and

$$\sigma_{g(X)}^2 = E[(g(X) - \mu_{g(X)})^2] = \int_{-\infty}^{\infty} (g(x) - \mu)^2 f(x) dx$$

if  $X$  is continuous.

**Definition 187** Let  $X$  and  $Y$  be random variables with joint probability distribution  $f(x, y)$ . The mean, or expected value, of the random variable  $g(X, Y)$  is

$$\mu_{g(X,Y)} = E[g(X, Y)] = \sum_x \sum_y g(x, y) f(x, y)$$

if  $X$  and  $Y$  are discrete, and

$$\mu_{g(X,Y)} = E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)f(x,y)dx dy$$

if  $X$  and  $Y$  are continuous.

**Exercise 188** Let  $X$  and  $Y$  be the random variables with joint probability distribution indicated in the following table. Find the expected value of  $g(X,Y) = XY$  .

$y \backslash x$	0	1	2	
0	$\frac{3}{28}$	$\frac{9}{29}$	$\frac{3}{28}$	$\frac{15}{28}$
1	$\frac{3}{14}$	$\frac{3}{14}$	0	$\frac{3}{7}$
2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
	$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

**Solution 189** By definition,

$$\begin{aligned} E(XY) &= \sum_{x=0}^2 \sum_{y=0}^2 xyf(x,y) \\ &= (0)(0)f(0,0) + (0)(1)f(0,1) \\ &\quad + (1)(0)f(1,0) + (1)(1)f(1,1) + (2)(0)f(2,0) \\ &= f(1,1) = \frac{3}{14}. \end{aligned}$$

**Definition 190** Let  $X$  and  $Y$  be random variables with joint probability distribution  $f(x,y)$ . The covariance of  $X$  and  $Y$  is

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = \sum_x \sum_y (x - \mu_X)(y - \mu_Y)f(x,y)$$

if  $X$  and  $Y$  are discrete, and

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y)f(x,y)dx dy$$

if  $X$  and  $Y$  are continuous.

**Theorem 191** The covariance of two random variables  $X$  and  $Y$  with means  $\mu_X$  and  $\mu_Y$  ,

respectively, is given by

$$\sigma_{XY} = E(XY) - \mu_X\mu_Y$$

**Example 192** Let the joint density function given by

$$f(x, y) = \begin{cases} 8xy, & 0 \leq y \leq x \leq 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the covariance of  $X$  and  $Y$ .

Although the covariance between two random variables does provide information regarding the nature of the relationship, the magnitude of  $\sigma_{XY}$  does not indicate anything regarding the strength of the relationship, since  $\sigma_{XY}$  is not scale-free. Its magnitude will depend on the units used to measure both  $X$  and  $Y$ . There is a scale-free version of the covariance called the correlation coefficient that is used widely in statistics.

**Definition 193** Let  $X$  and  $Y$  be random variables with covariance  $\sigma_{XY}$  and standard deviations  $\sigma_X$  and  $\sigma_Y$ , respectively. The correlation coefficient of  $X$  and  $Y$  is

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

It should be clear to the reader that  $\rho_{XY}$  is free of the units of  $X$  and  $Y$ . The correlation coefficient satisfies the inequality  $-1 \leq \rho_{XY} \leq 1$ . It assumes a value of zero when  $\sigma_{XY} = 0$ . Where there is an exact linear dependency, say  $Y \equiv a + bX$ ,  $\rho_{XY} = 1$  if  $b > 0$  and  $\rho_{XY} = -1$  if  $b < 0$ .

## 7.2 Means and Variances of Linear Combinations of Random Variables

**Theorem 194** If  $a$  and  $b$  are constants, then

$$E(aX + b) = aE(X) + b.$$

**Proof.** As exercise. ■

**Corollary 195** Setting  $a = 0$ , we see that  $E(b) = b$ .

**Corollary 196** Setting  $b = 0$ , we see that  $E(aX) = aE(X)$ .

**Theorem 197** The expected value of the sum or difference of two or more functions of a random variable  $X$  is the sum or difference of the expected values of the functions. That is,

$$E[g(X) \pm h(X)] = E[g(X)] \pm E[h(X)].$$

**Example 198** Let  $X$  be a random variable with probability distribution as follows:

$x$	0	1	2	3
$f(x)$	$\frac{1}{3}$	$\frac{1}{2}$	0	$\frac{1}{6}$

Find the expected value of  $Y = (X - 1)^2$ .

**Example 199** The weekly demand for a certain drink, in thousands of liters, at a chain of convenience stores is a continuous random variable  $g(X) = X^2 + X - 2$ , where  $X$  has the density function

$$f(x) = \begin{cases} 2(x - 1), & 1 < x < 2, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the expected value of the weekly demand for the drink.

**Theorem 200** The expected value of the sum or difference of two or more functions of the random variables  $X$  and  $Y$  is the sum or difference of the expected values of the functions. That is,

$$E[g(X, Y) \pm h(X, Y)] = E[g(X, Y)] \pm E[h(X, Y)].$$

**Corollary 201** Setting  $g(X, Y) = g(X)$  and  $h(X, Y) = h(Y)$ , we see that

$$E[g(X) \pm h(Y)] = E[g(X)] \pm E[h(Y)].$$

**Corollary 202** Setting  $g(X, Y) = X$  and  $h(X, Y) = Y$ , we see that

$$E[X \pm Y] = E[X] \pm E[Y].$$

**Theorem 203** *Let  $X$  and  $Y$  be two independent random variables. Then*

$$E(XY) = E(X)E(Y).$$

**Proof.** By Definition

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y)dx dy$$

Since  $X$  and  $Y$  are independent, we may write

$$f(x, y) = g(x)h(y),$$

where  $g(x)$  and  $h(y)$  are the marginal distributions of  $X$  and  $Y$ , respectively. Hence,

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyg(x)h(y)dx dy = \int_{-\infty}^{\infty} g(x)dx \int_{-\infty}^{\infty} h(y)dy = E(X)E(Y)$$

■

**Corollary 204** *Let  $X$  and  $Y$  be two independent random variables. Then  $cov(X, Y) = 0$ .*

**Theorem 205** *If  $X$  and  $Y$  are random variables with joint probability distribution  $f(x, y)$  and  $a$ ,  $b$ , and  $c$  are constants, then  $var(aX + bY + c) = a^2var(X) + b^2var(Y) + 2abcov(X, Y)$ .*

**Corollary 206** *Setting  $b = 0$ , we see that*

$$var(aX + c) = a^2var(X).$$

**Corollary 207** *Setting  $a = 1$  and  $b = 0$ , we see that*

$$var(X + c) = var(X).$$

**Corollary 208** *Setting  $b = 0$  and  $c = 0$ , we see that*

$$var(aX) = var(X)$$

**Corollary 209** *If  $X$  and  $Y$  are independent random variables, then*

$$\text{var}(aX + bY) = a^2\text{var}(X) + b^2\text{var}(Y).$$

**Corollary 210** *If  $X_1, X_2, \dots, X_n$  are independent random variables, then*

$$\text{var}(a_1X_1 + a_2X_2 + \dots + a_nX_n) = a_1^2\text{var}(X_1) + a_2^2\text{var}(X_2) + \dots + a_n^2\text{var}(X_n).$$

**Example 211** *If  $X$  and  $Y$  are random variables with variances  $\text{var}(X) = 2$  and  $\text{var}(Y) = 4$  and covariance  $\text{cov}(X, Y) = -2$ , find the variance of the random variable  $Z = 3X - 4Y + 8$ .*

### 7.3 Correlation

We consider the problem of measuring the relationship between the two variables  $X$  and  $Y$ . We want to determine whether large values of  $X$  are associated with large values of  $Y$ , and vice versa. Correlation analysis attempts to measure the strength of such relationships between two variables  $X$  and  $Y$  by means of a single number called a correlation coefficient.

**Definition 212 (Coefficient)** *The measure of linear association  $\rho$  between two variables  $X$  and  $Y$  is estimated by the sample correlation coefficient  $r$ , where*

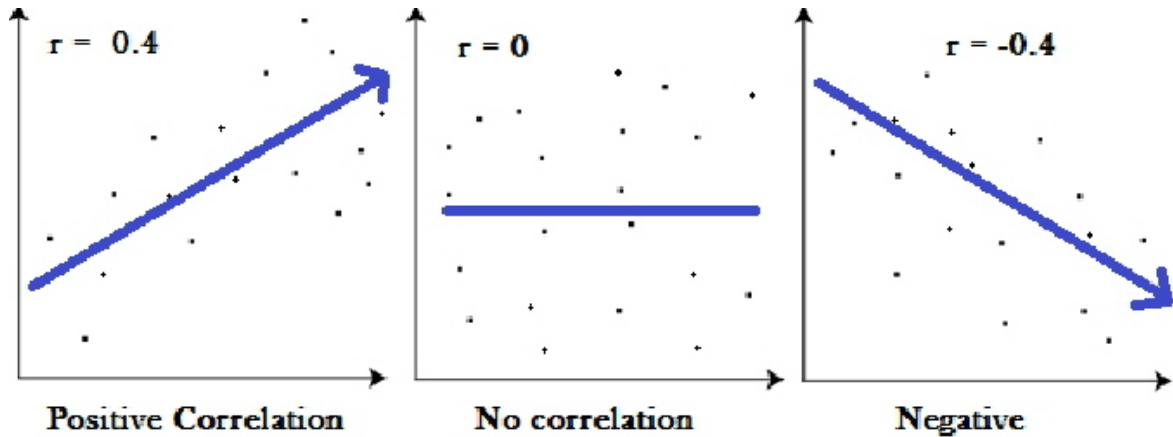
$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

with  $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ ,  $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$  and  $S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$ .

**Example 213** *Let consider the following grades of 6 students selected at random*

<i>Mathematics grade</i>	70	92	80	74	65	83
<i>English grade</i>	74	84	63	87	78	90

We have  $n = 6$ ,  $S_{xy} = 115.33$ ,  $S_{xx} = 471.33$ ,  $S_{yy} = 491.33$ . Hence  $r = \frac{115.33}{\sqrt{(471.33)(491.33)}} = 0.24$ .



### 7.3.1 Properties of $r$

$r = 1$  iff all  $(x_i, y_i)$  pairs lie on straight line with positive slope, and  $r = -1$  iff all  $(x_i, y_i)$  pairs lie on a straight line with negative slope.

## 7.4 Simple linear regression

The form of a relationship between the response  $Y$  (the dependent or the response variable) and the regressor  $X$  (the independent variable) is in mathematical the linear relationship

$$Y = \beta_0 + \beta_1 X + \varepsilon_i$$

where,  $\beta_0$  is the intercept,  $\beta_1$  the slope and  $\varepsilon_i$ , the error term in the model, is a random variable with mean 0 and constant variance. The relationship is illustrated in Figure 7.1.

An important aspect of regression analysis is to estimate the parameters  $\beta_0$  and  $\beta_1$  (i.e., estimate the so-called regression coefficients). The method of estimation will be discussed in the next section. Suppose we denote the estimates  $b_0$  for  $\beta_0$  and  $b_1$  for  $\beta_1$ . Then the estimated or fitted regression line is given by

$$\hat{y} = b_0 + b_1 x$$

where  $\hat{y}$  is the predicted or fitted value.

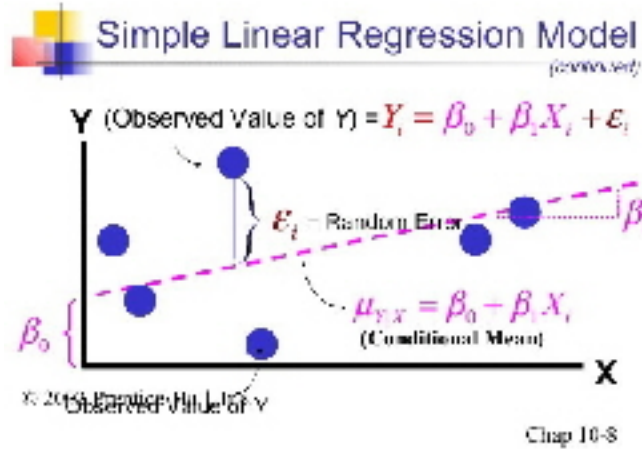


Figure 7-1: Figure 7.1

### 7.4.1 Least Squares and the Fitted Model

**Definition 214** Given a set of regression data  $\{(x_i, y_i); i = 1, 2, \dots, n\}$  and a fitted model,  $\hat{y}_i = b_0 + b_1 x_i$ , the  $i$ th residual  $e_i$  is given by

$$e_i = y_i - \hat{y}_i, i = 1, 2, \dots, n.$$

We shall find  $b_0$  and  $b_1$ , the estimates of  $\beta_0$  and  $\beta_1$ , so that the sum of the squares of the residuals is a minimum. This minimization procedure for estimating the parameters is called the method of least squares. Hence, we shall find  $b_0$  and  $b_1$  so as to minimize

$$\text{the error sum of squares} = SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y - \hat{y}_i)^2 = \sum_{i=1}^n (y - b_0 - b_1 x_i)^2$$

Differentiating SSE with respect to  $b_0$  and  $b_1$ , we get:

**Theorem 215** Given the sample  $\{(x_i, y_i); i = 1, 2, \dots, n\}$ , the least squares estimates  $b_0$  and  $b_1$



of the regression coefficients  $\beta_0$  and  $\beta_1$  are computed from the formulas

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

**Example 216** Consider the experimental data in Table 7.1, which were obtained from 33 samples of chemically treated waste in a study conducted at Virginia Tech. Readings on  $x$ , the percent reduction in total solids, and  $y$ , the percent reduction in chemical oxygen demand, were recorded.

<i>Solids Reduction,</i>	<i>Oxygen Demand</i>	<i>Solids Reduction,</i>	<i>Oxygen Demand</i>
<i>x (%)</i>	<i>Reduction, y (%)</i>	<i>x (%)</i>	<i>Reduction, y (%)</i>
3	5	36	34
7	11	37	36
11	21	38	38
15	16	39	37
18	16	39	36
27	28	39	45
29	27	40	39
30	25	41	41
30	35	42	40
31	30	42	44
31	40	43	37
32	32	44	44
33	34	45	46
33	32	46	46
34	34	47	49
36	37	50	51
36	38		

The estimated regression line is given by

$$\hat{y} = 3.8296 + 0.9036x.$$

Using the regression line, we would predict a 31% reduction in the chemical oxygen demand when the reduction in the total solids is 30%. The 31% reduction in the chemical oxygen demand may be interpreted as an estimate of the population mean  $\mu_Y|_{30}$  or as an estimate of a new observation when the reduction in total solids is 30%.

### 7.4.2 Properties of the Least Squares Estimators

We are interested in the expectation and variance the estimator  $B_1$  of  $\beta_1$  and the expectation of  $B_0$  the estimator of  $\beta_0$ .

**Theorem 217** We have:  $E(B_0) = \beta_0$ ,  $E(B_1) = \beta_1$ ,  $\text{var}(B_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}}$

**Theorem 218** An unbiased estimate of  $\sigma^2$ , named the mean squared error, is

$$\hat{\sigma}^2 = \frac{SSE}{n - 2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}$$

### 7.4.3 Inferences Concerning the Regression Coefficients

We assume now that the errors  $\varepsilon_i$  are normally distributed. ++++++

**Theorem 219** A  $100(1 - \alpha)\%$  confidence interval for the parameter  $\beta_1$  in the regression line

$$b_1 - t_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}} < \beta_1 < b_1 + t_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

where  $t_{\alpha/2}$  is a value of the  $t$ -distribution with  $n - 2$  degrees of freedom.

**Example 220** Find a 95% confidence interval for  $\beta_1$  in the regression line, based on the pollution data of Table 7.1.

**Solution 221** We show that

$$\hat{\sigma}^2 = \frac{SSE}{n - 2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2} = 0.4299.$$

Therefore, taking the square root, we obtain  $\hat{\sigma} = 3.2295$ . Also,  $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = 4152.18$ . Using Table A.4, we find that  $t_{0.025} \approx 2.045$  for 31 degrees of freedom. Therefore, a 95% confidence interval for  $\beta_1$  is

$$0.903643 - (2.045) \frac{3.2295}{\sqrt{4152.18}} < \beta_1 < 0.903643 + (2.045) \frac{3.2295}{\sqrt{4152.18}}$$

which simplifies to

$$0.8012 < \beta_1 < 1.0061.$$

## Hypothesis Testing on the Slope

To test the null hypothesis  $H_0$  that  $\beta_1 = \beta_{10}$ , we again use the t-distribution with  $n - 2$  degrees of freedom to establish a critical region and then base our decision on the value of

$$t = \frac{b_1 - \beta_{10}}{\hat{\sigma} / \sqrt{S_{xx}}}$$

**Example 222** Using the estimated value  $b_1 = 0.903643$  of Example 7.1, test the hypothesis that  $\beta_1 = 1.0$  against the alternative that  $\beta_1 < 1.0$ .

**Solution 223** The hypotheses are  $H_0 : \beta_1 = 1.0$  and  $H_1 : \beta_1 < 1.0$ . So

$$t = \frac{0.903643 - 1.0}{3.2295 / \sqrt{4152.18}} = -1.92,$$

with  $n - 2 = 31$  degrees of freedom ( $P \approx 0.03$ ).

*Decision:*  $P$ -value  $< 0.05$ , suggesting strong evidence that  $\beta_1 < 1.0$

One important t-test on the slope is the test of the hypothesis  $H_0 : \beta_1 = 0$  versus  $H_1 : \beta_1 \neq 0$ . When the null hypothesis is not rejected, the conclusion is that there is no significant linear relationship between  $E(y)$  and the independent variable  $x$ . Rejection of  $H_0$  above implies that a significant linear regression exists.

## Measuring Goodness-of-Fit: the Coefficient of Determination

A goodness-of-fit statistic is a quantity that measures how well a model explains a given set of data. A linear model fits well if there is a strong linear relationship between  $x$  and  $y$ .

**Definition 224** *The coefficient of determination,  $R^2$ , is given by*

$$R^2 = 1 - \frac{SSE}{SST}$$

where  $SSE = \sum_{i=1}^n (y - \hat{y}_i)^2$  and  $SST = \sum_{i=1}^n (y - \bar{y})^2$ .

Note that if the fit is perfect, all residuals  $y - \hat{y}_i$  are zero, and thus  $R^2 = 1.0$ . But if  $SSE$  is only slightly smaller than  $SST$ ,  $R^2 \approx 0$ . In the example of table 7.1, the coefficient of determination  $R^2 = 0.913$ , suggests that the model fit to the data explains 91.3% of the variability observed in the response, the reduction in chemical oxygen demand.

**Theorem 225** *The square  $r^2$  of the sample correlation coefficient gives the value of the coefficient of determination  $R^2$  that would result from fitting the simple linear regression model.*

# Chapter 8

## 8.1 One Way ANOVA

Suppose in an industrial experiment that an engineer is interested in how the mean absorption of moisture in concrete varies among 5 different concrete aggregates. The samples are exposed to moisture for 48 hours. It is decided that 6 samples are to be tested for each aggregate, requiring a total of 30 samples to be tested. The data are recorded in Table 13.1.

The model for this situation may be set up as follows. There are 6 observations taken from each of 5 populations with means  $\mu_1, \mu_2, \dots, \mu_5$ , respectively. We may wish to test

$H_0: \mu_1 = \mu_2 = \dots = \mu_5$ ,

$H_1$ : At least two of the means are not equal.

Table 8.1: Absorption of Moisture in Concrete Aggregates

Aggregate	1	2	3	4	5	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
<b>Total</b>	3320	3416	3663	2791	3664	16854
<b>Mean</b>	553.33	569.33	610.5	465.17	610.67	561.8

### Two Sources of Variability in the Data

In the analysis-of-variance procedure, it is assumed that whatever variation exists among the aggregate averages is attributed to (1) variation in absorption among observations *within* aggregate types and (2) variation *among* aggregate types, that is, due to differences in the chemical composition of the aggregates. The **within aggregate variation** is, of course, brought about by various causes. Perhaps humidity and temperature conditions were not kept entirely constant throughout the experiment. It is possible that there was a certain amount of heterogeneity in the batches of raw materials that were used. At any rate, we shall consider the within-sample variation to be **chance or random variation**. Part of the goal of the analysis of variance is to determine if the differences among the 5 sample means are what we would expect due to random variation alone or, rather, due to variation beyond merely random effects, i.e., differences in the chemical composition of the aggregates.

### One-Way Analysis of Variance: Completely Randomized Design (One-Way ANOVA)

Random samples of size  $n$  are selected from each of  $k$  populations. The  $k$  different populations are classified on the basis of a single criterion such as different treatments or groups. Today the term **treatment** is used generally to refer to the various classifications,

whether they be different aggregates, different analysts, different fertilizers, or different regions of the country.

### Assumptions and Hypotheses in One-Way ANOVA

It is assumed that the  $k$  populations are independent and normally distributed with means  $\mu_1, \mu_2, \dots, \mu_k$  and common variance  $\sigma^2$ .

We wish to derive appropriate methods for testing the hypothesis

$H_0: \mu_1 = \mu_2 = \dots = \mu_k$ ,

$H_1$ : At least two of the means are not equal.

Let  $y_{ij}$  denote the  $j$ th observation from the  $i$ th treatment and arrange the data as in Table 13.2. Here,  $Y_i$  is the total of all observations in the sample from the  $i$ th treatment,  $\bar{y}_i$  is the mean of all observations in the sample from the  $i$ th treatment,  $Y_{..}$  is the total of all  $nk$  observations, and  $\bar{y}_{..}$  is the mean of all  $nk$  observations.

Treatment:	1	2	...	$i$	...	$k$	
	$y_{11}$	$y_{21}$	...	$y_{i1}$	...	$y_{k1}$	
	$y_{12}$	$y_{22}$	...	$y_{i2}$	...	$y_{k2}$	
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
	$y_{1n}$	$y_{2n}$	...	$y_{in}$	...	$y_{kn}$	
Total	$Y_1$	$Y_2$	...	$Y_i$	...	$Y_k$	$Y_{..}$
Mean	$\bar{y}_1$	$\bar{y}_2$	...	$\bar{y}_i$	...	$\bar{y}_k$	$\bar{y}_{..}$

### Model for One-Way ANOVA

Each observation may be written in the form

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$

where  $\varepsilon_{ij}$  measures the deviation of the  $j$ th observation of the  $i$ th sample from the corresponding treatment mean. The  $\varepsilon_{ij}$ -term represents random error and plays the same role as the error terms in the regression models.

### Theorem: Sum-of-Squares Identity

**Sum-of-Squares Identity**

$$\sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 = n \sum_{i=1}^k (\bar{y}_i - \bar{y}_{..})^2 + \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2$$

It will be convenient in what follows to identify the terms of the sum-of-squares identity by the following notation:

Three Important Measures of Variability

$$SST = \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 = \text{total sum of squares,}$$

$$SSA = n \sum_{i=1}^k (\bar{y}_i - \bar{y}_{..})^2 = \text{treatment sum of squares,}$$

$$SSE = \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2 = \text{error sum of squares.}$$

The sum-of-squares identity can then be represented symbolically by the equation  $SST = SSA + SSE$ .

The identity above expresses how between-treatment and within-treatment variation add to the total sum of squares.

### Analysis of Variance for the One-Way ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Computed $f$
Treatments	$SSA$	$k - 1$	$s_1^2 = \frac{SSA}{k - 1}$	$\frac{s_1^2}{s^2}$
Error	$SSE$	$k(n - 1)$	$s^2 = \frac{SSE}{k(n - 1)}$	
Total	$SST$	$kn - 1$		

### F-Ratio for Testing Equality of Means

When  $H_0$  is true, the ratio  $f = s_1^2/s^2$  is a value of the random variable  $F$  having the  $F$ -distribution with  $k-1$  and  $k(n-1)$  degrees of freedom.

The null hypothesis  $H_0$  is rejected at the  $\alpha$ -level of significance when  $f > f_{\alpha}[k - 1, k(n - 1)]$ .

Example:

Test the hypothesis  $\mu_1 = \mu_2 = \dots = \mu_5$  at the 0.05 level of significance for the data of Table 13.1 on absorption of moisture by various types of cement aggregates.

**Solution :** The hypotheses are

$H_0: \mu_1 = \mu_2 = \dots = \mu_5$ ,

$H_1$ : At least two of the means are not equal.

$\alpha = 0.05$ .

Critical region:  $f > 2.76$  with  $v_1 = 4$  and  $v_2 = 25$  degrees of freedom. The sum-of-squares computations give

$SST = 209,377$ ,  $SSA = 85,356$ ,

$SSE = 209,377 - 85,356 = 124,021$ . the ratio  $f = s_1^2/s^2 = 4.30$  These results and the remaining computations are exhibited in the next figure in the SAS ANOVA procedure.

The GLM Procedure					
Dependent Variable: moisture					
Source	DF	Squares	Sum of Mean Square	F Value	Pr > F
Model	4	85356.4667	21339.1167	4.30	0.0088
Error	25	124020.3333	4960.8133		
Corrected Total	29	209376.8000			
R-Square	Coeff Var	Root MSE	moisture Mean		
0.407669	12.53703	70.43304	561.8000		
Source	DF	Type I SS	Mean Square	F Value	Pr > F
aggregate	4	85356.46667	21339.11667	4.30	0.0088

Decision: Reject  $H_0$  and conclude that the aggregates do not have the same mean absorption. The  $P$ -value for  $f = 4.30$  is 0.0088, which is smaller than 0.05.

## Comparing a Set of Treatments in Blocks

where  $y_{11}$  represents the response obtained by using treatment 1 in block 1,  $y_{12}$  represents the response obtained by using treatment 1 in block 2, ..., and  $y_{34}$  represents the response obtained by using treatment 3 in block 4.

Let us now generalize and consider the case of  $k$  treatments assigned to  $b$  blocks.

Treatment	Block							
	1	2	j				b	
1	$y_{11}$	$y_{12}$	...	...	$y_{1j}$	...	$y_{1b}$	$\bar{y}_{1.}$
2	$y_{21}$	$y_{22}$	...	...	$y_{2j}$	...	$y_{2b}$	$\bar{y}_{2.}$
...								
...								
...								
...								
...								
...								
...								
...								
i	$y_{i1}$	$y_{i2}$	...		$y_{ij}$	...	$y_{ib}$	$\bar{y}_{i.}$
...								
...								
...								
...								
...								
...								
...								
k	$y_{k1}$	$y_{k2}$	...		$y_{kj}$	...	$y_{kb}$	$\bar{y}_{k.}$
Mean	$\bar{y}_{.1}$	$\bar{y}_{.2}$	...		$\bar{y}_{.j}$	...	$\bar{y}_{.b}$	$\bar{y}_{..}$

$y_{ij}$  = the observations in the ( $ij$ )th cell,  
 $\bar{y}_{i.}$  = mean of the observations for the  $i$ th level,  
 $\bar{y}_{.j}$  = mean of the observations for the  $j$ th block,  
 $\bar{y}_{..}$  = mean of all  $kn$  observations.

### Model and Hypotheses

The analysis of variance model is based on the following

$$y_{ij} = \mu_{ij} + \varepsilon_{ij} \quad i = 1, \dots, k; \quad j = 1, \dots, b$$

where  $y_{ij}$  represents the observation of the  $i$ th treatment in the  $j$ th block,  $\mu_{ij}$  is the mean response and  $\varepsilon_{ij}$  are errors which are independent  $N(0, \sigma^2)$ . The response mean for the level  $i$  of the factor is  $\mu_{i.} = E(\bar{y}_{i.})$  and the response mean for the block  $j$  is  $\mu_{.j} = E(\bar{y}_{.j})$  and the overall mean is  $\mu_{..} = E(\bar{y}_{..})$ .

The hypothesis to be tested is as follows:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

$$H_1 : \text{at least two treatments have different means}$$

### Partitioning of Variability in the Two-Factor Case



Theorem: Sum of squares Identity

$$\sum_{i=1}^k \sum_{j=1}^b (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^k b(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{j=1}^b k(\bar{y}_{.j} - \bar{y}_{..})^2 + \sum_{i=1}^k \sum_{j=1}^b (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$$

The sum-of-squares identity may be presented symbolically by the equation

$$SST = SSA + SSB + SSE,$$

where

$$SST = \sum_{i=1}^k \sum_{j=1}^b (y_{ij} - \bar{y}_{..})^2$$

$$SSA = \sum_{i=1}^k b(\bar{y}_{i.} - \bar{y}_{..})^2$$

$$SSB = \sum_{j=1}^b k(\bar{y}_{.j} - \bar{y}_{..})^2$$

$$SSE = \sum_{i=1}^k \sum_{j=1}^b (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$$

The ANOVA table for the randomized block design appears in the following table.

Source	Degrees of freedom	Sum of Squares
Treatment	k-1	$\sum_{i=1}^k b(\bar{y}_{i.} - \bar{y}_{..})^2$
Block	b-1	$\sum_{j=1}^b k(\bar{y}_{.j} - \bar{y}_{..})^2$
Residual	(k-1)(b-1)	$\sum_{i=1}^k \sum_{j=1}^b (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$
Total	bk-1	$\sum_{i=1}^k \sum_{j=1}^b (y_{ij} - \bar{y}_{..})^2$

The null hypothesis of no treatment effect difference  $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ , can be tested by using the  $F$  statistic

$$F = \frac{SSTr / (k-1)}{SSE / (k-1)(b-1)}$$

where  $SSTr$  and  $SSE$  are the treatment and error sums of squares.

The  $F$  test rejects  $H_0$  at level  $\alpha$  if the  $F$  value exceeds  $F_{\alpha}[k-1, (k-1)(b-1)]$ .

**Example:**

**Table 8.2**

Machine	Operator(block)						Total
	1	2	3	4	5	6	
1	42.5	39.3	39.6	39.9	42.9	43.6	247.8
2	39.8	40.1	40.5	42.3	42.5	43.1	248.3
3	40.2	40.5	41.3	43.4	44.9	45.1	255.4
4	41.3	42.2	43.5	44.2	45.9	42.3	259.4
Total	163.8	162.1	164.9	169.8	176.2	174.1	1010.9

Analysis of Variance for the Data of Table 8.2

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Computed <i>f</i>
Machines	15.93	3	5.31	3.34
Operators	42.09	5	8.42	
Error	23.84	15	1.59	
Total	81.86	23		

The null hypothesis is rejected at the 5%-level of significance since  $f=3.34 > f_{\alpha}[k - 1, (k - 1)(b - 1)] = f_{0.05}[3, 15]=3.29$ .

Similarly for the null hypothesis  $H_0 : \mu_1 = \mu_j = \dots = \mu_b$ , we can compute the value of the *F* statistic  $F_{b-1,(k-1)(b-1)}$  for testing the difference between blocks:

$$\frac{8.42}{1.59} = 5.30$$

Since  $5.30 > 2.90 = f_{\alpha}[b - 1, (k - 1)(b - 1)]$ , suggesting that there is a significant difference between the six blocks.

## Chapter 9

# Nonparametric Statistics

Most of the hypothesis-testing procedures discussed in previous chapters are based on the assumption that the random samples are selected from normal populations. In this chapter, we consider a number of alternative test procedures, called nonparametric or distribution-free methods, that often assume no knowledge whatsoever about the distributions of the underlying populations.

### 9.1 Wilcoxon signed rank test

We assume that the random variable  $X$  is symmetric around its median  $\theta$ . We consider testing the null hypothesis that the median of  $X$  is zero; that is  $H_0 : \theta = 0$ . Let consider the following data:

$$X_1 = -0.5 \quad X_2 = 2 \quad X_3 = 1$$

When  $H_0$  is true, we expect the magnitudes of the negative observations in the sample to be comparable to the magnitudes of the positive observations. let compute disregard the signs of the observations and rank the absolute magnitudes from 1 to 3, with the smallest getting rank 1, the second smallest rank 2, and so on. Then apply the sign of each observation to the

corresponding rank . The test statistic will be  $S^+$  the sum of the positively signed ranks.

$X$	-0.5	1	2
Rank	1	2	3
Signed rank	-1	2	3

$s^+ = 2 + 3 = 5$ . Let consider  $H_1 : \theta > 0$ . When the median of the distribution is much greater than 0, most of the observations with large absolute magnitudes should be positive, resulting in positively signed ranks and a large value of  $S^+$ . Let compute all the  $S^+$ .

$i$	1	2	3	$S^+$
1	1	3	2	6
2	1	3	-2	4
3	1	-3	2	3
4	-1	3	2	5
5	1	-3	-2	1
6	-1	-3	2	2
7	-1	3	-2	3
8	-1	-3	-2	0

We deduce the null distribution of  $S^+$  when  $n = 3$ .

$s^+$	0	1	2	3	4	5	6
distribution of $S^+$	1/8	1/8	1/8	2/8	1/8	1/8	1/8

P-value =  $\Pr(S^+ \geq s_{\text{observed}}^+) = \Pr(S^+ \geq 5) = 2/8 > 0.05$ . We then not reject  $H_0$ . One cal also use the table A.13. We have  $\Pr(S^+ \geq 6) = 0.125$ . Then  $\Pr(S^+ \geq 5) > \Pr(S^+ \geq 6) > 0.05$ . The critical values  $c_1$  obtained from Table A.13 satisfy  $\Pr(S^+ \geq c_1) \approx \alpha (= 0.05)$ .

## 9.2 Rank correlation coefficient

In Chapter 7, we used the sample correlation coefficient  $r$  to measure the population correlation coefficient  $\rho$ , the linear relationship between two continuous variables  $X$  and  $Y$ . If ranks  $1, 2, \dots, n$  are assigned to the  $x$  observations in order of magnitude and similarly to the  $y$  observations, and if these ranks are then substituted for the actual numerical values in the formula for the correlation coefficient in Chapter 7, we obtain the nonparametric counterpart of the conventional correlation coefficient. A correlation coefficient calculated in this manner is known as the Spearman rank correlation coefficient and is denoted by  $r_s$ .

**Definition 1 (Rank Correlation Coefficient)** *A nonparametric measure of association between two variables  $X$  and  $Y$  is given by the rank correlation coefficient by the rank correlation coefficient*

$$r_s = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^n d_i^2$$

where  $d_i$  is the difference between the ranks assigned to  $x_i$  and  $y_i$  and  $n$  is the number of pairs of data.

In practice, the preceding formula is also used when there are ties among either the  $x$  or  $y$  observations. The ranks for tied observations are assigned as in the signed-rank test by averaging the ranks that would have been assigned if the observations were distinguishable.

The value of  $r_s$  will usually be close to the value obtained by finding  $r$  based on numerical measurements and is interpreted in much the same way. As before, the value of  $r_s$  will range from  $-1$  to  $+1$ . A value of  $+1$  or  $-1$  indicates perfect association between  $X$  and  $Y$ , the plus sign occurring for identical rankings and the minus sign occurring for reverse rankings. When  $r_s$  is close to zero, we conclude that the variables are uncorrelated.

**Example 2** *Table 9.1 shows the milligrams of tar (██████) and nicotine found in 10 brands of cigarettes. Calculate the rank correlation coefficient to measure the degree of relationship*

between tar and nicotine content in cigarettes.

<b><i>Cigarette Brand</i></b>	<b><i>Tar Content</i></b>	<b><i>Nicotine Content</i></b>
<i>Viceroy</i>	14	0.9
<i>Marlboro</i>	17	1.1
<i>Chesterfield</i>	28	1.6
<i>Kool</i>	17	1.3
<i>Kent</i>	16	1.0
<i>Raleigh</i>	13	0.8
<i>Old Gold</i>	24	1.5
<i>Philip Morris</i>	25	1.4
<i>Oasis</i>	18	1.2
<i>Players</i>	31	2.0

**Solution 3** Here the ranking of Tar and Nicotine Content.

<b><i>Cigarette Brand</i></b>	$x_i$	$y_i$	$d_i$
<i>Viceroy</i>	2	2	0
<i>Marlboro</i>	4.5	4	0.5
<i>Chesterfield</i>	9	9	0
<i>Kool</i>	4.5	6	-1.5
<i>Kent</i>	3	3	0
<i>Raleigh</i>	1	1	0
<i>Old Gold</i>	7	8	-1
<i>Philip Morris</i>	8	7	1
<i>Oasis</i>	6	5	1
<i>Players</i>	10	10	0

Substituting into the formula for  $r_s$ , we find that

$$r_s = 1 - \frac{(6)(5.50)}{(10)(100 - 1)} = 0.967,$$

*indicating a high positive correlation between the amounts of tar and nicotine found in cigarettes.*

Some advantages to using  $r_s$  rather than  $r$  do exist. For instance, when the data possess a distinct curvilinear relationship, the rank correlation coefficient will likely be more reliable than the conventional measure. Perhaps the greatest advantage occurs when we are unable to make meaningful numerical measurements but nevertheless can establish rankings. Such is the case, for example, when different

judges rank a group of individuals according to some attribute. The rank correlation coefficient can be used in this situation as a measure of the consistency of the two judges.

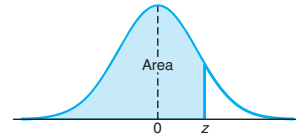
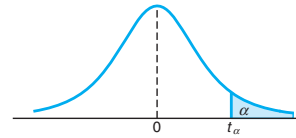


Table A.3 Areas under the Normal Curve

<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641







**Table A.4** Critical Values of the *t*-Distribution

<i>v</i>	$\alpha$						
	<b>0.40</b>	<b>0.30</b>	<b>0.20</b>	<b>0.15</b>	<b>0.10</b>	<b>0.05</b>	<b>0.025</b>
<b>1</b>	0.325	0.727	1.376	1.963	3.078	6.314	12.706
<b>2</b>	0.289	0.617	1.061	1.386	1.886	2.920	4.303
<b>3</b>	0.277	0.584	0.978	1.250	1.638	2.353	3.182
<b>4</b>	0.271	0.569	0.941	1.190	1.533	2.132	2.776
<b>5</b>	0.267	0.559	0.920	1.156	1.476	2.015	2.571
<b>6</b>	0.265	0.553	0.906	1.134	1.440	1.943	2.447
<b>7</b>	0.263	0.549	0.896	1.119	1.415	1.895	2.365
<b>8</b>	0.262	0.546	0.889	1.108	1.397	1.860	2.306
<b>9</b>	0.261	0.543	0.883	1.100	1.383	1.833	2.262
<b>10</b>	0.260	0.542	0.879	1.093	1.372	1.812	2.228
<b>11</b>	0.260	0.540	0.876	1.088	1.363	1.796	2.201
<b>12</b>	0.259	0.539	0.873	1.083	1.356	1.782	2.179
<b>13</b>	0.259	0.538	0.870	1.079	1.350	1.771	2.160
<b>14</b>	0.258	0.537	0.868	1.076	1.345	1.761	2.145
<b>15</b>	0.258	0.536	0.866	1.074	1.341	1.753	2.131
<b>16</b>	0.258	0.535	0.865	1.071	1.337	1.746	2.120
<b>17</b>	0.257	0.534	0.863	1.069	1.333	1.740	2.110
<b>18</b>	0.257	0.534	0.862	1.067	1.330	1.734	2.101
<b>19</b>	0.257	0.533	0.861	1.066	1.328	1.729	2.093
<b>20</b>	0.257	0.533	0.860	1.064	1.325	1.725	2.086
<b>21</b>	0.257	0.532	0.859	1.063	1.323	1.721	2.080
<b>22</b>	0.256	0.532	0.858	1.061	1.321	1.717	2.074
<b>23</b>	0.256	0.532	0.858	1.060	1.319	1.714	2.069
<b>24</b>	0.256	0.531	0.857	1.059	1.318	1.711	2.064
<b>25</b>	0.256	0.531	0.856	1.058	1.316	1.708	2.060
<b>26</b>	0.256	0.531	0.856	1.058	1.315	1.706	2.056
<b>27</b>	0.256	0.531	0.855	1.057	1.314	1.703	2.052
<b>28</b>	0.256	0.530	0.855	1.056	1.313	1.701	2.048
<b>29</b>	0.256	0.530	0.854	1.055	1.311	1.699	2.045
<b>30</b>	0.256	0.530	0.854	1.055	1.310	1.697	2.042
<b>40</b>	0.255	0.529	0.851	1.050	1.303	1.684	2.021
<b>60</b>	0.254	0.527	0.848	1.045	1.296	1.671	2.000
<b>120</b>	0.254	0.526	0.845	1.041	1.289	1.658	1.980
$\infty$	0.253	0.524	0.842	1.036	1.282	1.645	1.960

Table A.4 (continued) Critical Values of the  $t$ -Distribution

$v$	$\alpha$						
	0.02	0.015	0.01	0.0075	0.005	0.0025	0.0005
1	15.894	21.205	31.821	42.433	63.656	127.321	636.578
2	4.849	5.643	6.965	8.073	9.925	14.089	31.600
3	3.482	3.896	4.541	5.047	5.841	7.453	12.924
4	2.999	3.298	3.747	4.088	4.604	5.598	8.610
5	2.757	3.003	3.365	3.634	4.032	4.773	6.869
6	2.612	2.829	3.143	3.372	3.707	4.317	5.959
7	2.517	2.715	2.998	3.203	3.499	4.029	5.408
8	2.449	2.634	2.896	3.085	3.355	3.833	5.041
9	2.398	2.574	2.821	2.998	3.250	3.690	4.781
10	2.359	2.527	2.764	2.932	3.169	3.581	4.587
11	2.328	2.491	2.718	2.879	3.106	3.497	4.437
12	2.303	2.461	2.681	2.836	3.055	3.428	4.318
13	2.282	2.436	2.650	2.801	3.012	3.372	4.221
14	2.264	2.415	2.624	2.771	2.977	3.326	4.140
15	2.249	2.397	2.602	2.746	2.947	3.286	4.073
16	2.235	2.382	2.583	2.724	2.921	3.252	4.015
17	2.224	2.368	2.567	2.706	2.898	3.222	3.965
18	2.214	2.356	2.552	2.689	2.878	3.197	3.922
19	2.205	2.346	2.539	2.674	2.861	3.174	3.883
20	2.197	2.336	2.528	2.661	2.845	3.153	3.850
21	2.189	2.328	2.518	2.649	2.831	3.135	3.819
22	2.183	2.320	2.508	2.639	2.819	3.119	3.792
23	2.177	2.313	2.500	2.629	2.807	3.104	3.768
24	2.172	2.307	2.492	2.620	2.797	3.091	3.745
25	2.167	2.301	2.485	2.612	2.787	3.078	3.725
26	2.162	2.296	2.479	2.605	2.779	3.067	3.707
27	2.158	2.291	2.473	2.598	2.771	3.057	3.689
28	2.154	2.286	2.467	2.592	2.763	3.047	3.674
29	2.150	2.282	2.462	2.586	2.756	3.038	3.660
30	2.147	2.278	2.457	2.581	2.750	3.030	3.646
40	2.123	2.250	2.423	2.542	2.704	2.971	3.551
60	2.099	2.223	2.390	2.504	2.660	2.915	3.460
120	2.076	2.196	2.358	2.468	2.617	2.860	3.373
$\infty$	2.054	2.170	2.326	2.432	2.576	2.807	3.290

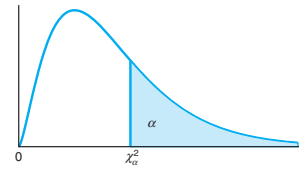


Table A.5 Critical Values of the Chi-Squared Distribution

<i>v</i>	$\alpha$									
	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.75	0.70	0.50
1	0.0 <sup>4</sup> 393	0.0 <sup>3</sup> 157	0.0 <sup>3</sup> 628	0.0 <sup>3</sup> 982	0.00393	0.0158	0.0642	0.102	0.148	0.455
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	0.575	0.713	1.386
3	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	1.213	1.424	2.366
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	1.923	2.195	3.357
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	2.675	3.000	4.351
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	3.455	3.828	5.348
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	4.255	4.671	6.346
8	1.344	1.647	2.032	2.180	2.733	3.490	4.594	5.071	5.527	7.344
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	5.899	6.393	8.343
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	6.737	7.267	9.342
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	7.584	8.148	10.341
12	3.074	3.571	4.178	4.404	5.226	6.304	7.807	8.438	9.034	11.340
13	3.565	4.107	4.765	5.009	5.892	7.041	8.634	9.299	9.926	12.340
14	4.075	4.660	5.368	5.629	6.571	7.790	9.467	10.165	10.821	13.339
15	4.601	5.229	5.985	6.262	7.261	8.547	10.307	11.037	11.721	14.339
16	5.142	5.812	6.614	6.908	7.962	9.312	11.152	11.912	12.624	15.338
17	5.697	6.408	7.255	7.564	8.672	10.085	12.002	12.792	13.531	16.338
18	6.265	7.015	7.906	8.231	9.390	10.865	12.857	13.675	14.440	17.338
19	6.844	7.633	8.567	8.907	10.117	11.651	13.716	14.562	15.352	18.338
20	7.434	8.260	9.237	9.591	10.851	12.443	14.578	15.452	16.266	19.337
21	8.034	8.897	9.915	10.283	11.591	13.240	15.445	16.344	17.182	20.337
22	8.643	9.542	10.600	10.982	12.338	14.041	16.314	17.240	18.101	21.337
23	9.260	10.196	11.293	11.689	13.091	14.848	17.187	18.137	19.021	22.337
24	9.886	10.856	11.992	12.401	13.848	15.659	18.062	19.037	19.943	23.337
25	10.520	11.524	12.697	13.120	14.611	16.473	18.940	19.939	20.867	24.337
26	11.160	12.198	13.409	13.844	15.379	17.292	19.820	20.843	21.792	25.336
27	11.808	12.878	14.125	14.573	16.151	18.114	20.703	21.749	22.719	26.336
28	12.461	13.565	14.847	15.308	16.928	18.939	21.588	22.657	23.647	27.336
29	13.121	14.256	15.574	16.047	17.708	19.768	22.475	23.567	24.577	28.336
30	13.787	14.953	16.306	16.791	18.493	20.599	23.364	24.478	25.508	29.336
40	20.707	22.164	23.838	24.433	26.509	29.051	32.345	33.66	34.872	39.335
50	27.991	29.707	31.664	32.357	34.764	37.689	41.449	42.942	44.313	49.335
60	35.534	37.485	39.699	40.482	43.188	46.459	50.641	52.294	53.809	59.335

Table A.5 (continued) Critical Values of the Chi-Squared Distribution

<i>v</i>	$\alpha$									
	0.30	0.25	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	1.074	1.323	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	2.408	2.773	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	3.665	4.108	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.266
4	4.878	5.385	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.466
5	6.064	6.626	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.515
6	7.231	7.841	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	8.383	9.037	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.321
8	9.524	10.219	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.124
9	10.656	11.389	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	11.781	12.549	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	12.899	13.701	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	14.011	14.845	15.812	18.549	21.026	23.337	24.054	26.217	28.300	32.909
13	15.119	15.984	16.985	19.812	22.362	24.736	25.471	27.688	29.819	34.527
14	16.222	17.117	18.151	21.064	23.685	26.119	26.873	29.141	31.319	36.124
15	17.322	18.245	19.311	22.307	24.996	27.488	28.259	30.578	32.801	37.698
16	18.418	19.369	20.465	23.542	26.296	28.845	29.633	32.000	34.267	39.252
17	19.511	20.489	21.615	24.769	27.587	30.191	30.995	33.409	35.718	40.791
18	20.601	21.605	22.760	25.989	28.869	31.526	32.346	34.805	37.156	42.312
19	21.689	22.718	23.900	27.204	30.144	32.852	33.687	36.191	38.582	43.819
20	22.775	23.828	25.038	28.412	31.410	34.170	35.020	37.566	39.997	45.314
21	23.858	24.935	26.171	29.615	32.671	35.479	36.343	38.932	41.401	46.796
22	24.939	26.039	27.301	30.813	33.924	36.781	37.659	40.289	42.796	48.268
23	26.018	27.141	28.429	32.007	35.172	38.076	38.968	41.638	44.181	49.728
24	27.096	28.241	29.553	33.196	36.415	39.364	40.270	42.980	45.558	51.179
25	28.172	29.339	30.675	34.382	37.652	40.646	41.566	44.314	46.928	52.619
26	29.246	30.435	31.795	35.563	38.885	41.923	42.856	45.642	48.290	54.051
27	30.319	31.528	32.912	36.741	40.113	43.195	44.140	46.963	49.645	55.475
28	31.391	32.620	34.027	37.916	41.337	44.461	45.419	48.278	50.994	56.892
29	32.461	33.711	35.139	39.087	42.557	45.722	46.693	49.588	52.335	58.301
30	33.530	34.800	36.250	40.256	43.773	46.979	47.962	50.892	53.672	59.702
40	44.165	45.616	47.269	51.805	55.758	59.342	60.436	63.691	66.766	73.403
50	54.723	56.334	58.164	63.167	67.505	71.420	72.613	76.154	79.490	86.660
60	65.226	66.981	68.972	74.397	79.082	83.298	84.58	88.379	91.952	99.608

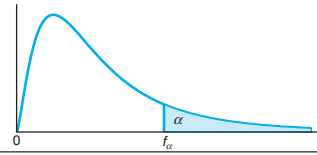


Table A.6 Critical Values of the F-Distribution

		$f_{0.05}(v_1, v_2)$								
		$v_1$								
$v_2$	1	2	3	4	5	6	7	8	9	
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	

Reproduced from Table 18 of *Biometrika Tables for Statisticians*, Vol. I, by permission of E.S. Pearson and the Biometrika Trustees.

Table A.6 (continued) Critical Values of the  $F$ -Distribution

		$f_{0.05}(v_1, v_2)$								
		$v_1$								
$v_2$	10	12	15	20	24	30	40	60	120	$\infty$
1	241.88	243.91	245.95	248.01	249.05	250.10	251.14	252.20	253.25	254.31
2	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
$\infty$	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

**Table A.6** (continued) Critical Values of the F-Distribution

		$f_{0.01}(v_1, v_2)$								
		$v_1$								
$v_2$	1	2	3	4	5	6	7	8	9	
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	
$\infty$	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	



Table A.6 (continued) Critical Values of the  $F$ -Distribution

		$f_{0.01}(v_1, v_2)$								
		$v_1$								
$v_2$	10	12	15	20	24	30	40	60	120	$\infty$
1	6055.85	6106.32	6157.28	6208.73	6234.63	6260.65	6286.78	6313.03	6339.39	6365.86
2	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
$\infty$	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

**Table A.13 Critical Values for the Wilcoxon Signed-Rank Test**

$$P_0(S_+ \geq c_1) = P(S_+ \geq c_1 \text{ when } H_0 \text{ is true})$$

$n$	$c_1$	$P_0(S_+ \geq c_1)$	$n$	$c_1$	$P_0(S_+ \geq c_1)$
3	6	.125		78	.011
4	9	.125		79	.009
	10	.062		81	.005
5	13	.094	14	73	.108
	14	.062		74	.097
	15	.031		79	.052
6	17	.109		84	.025
	19	.047		89	.010
	20	.031		92	.005
	21	.016	15	83	.104
7	22	.109		84	.094
	24	.055		89	.053
	26	.023		90	.047
	28	.008		95	.024
8	28	.098		100	.011
	30	.055		101	.009
	32	.027		104	.005
	34	.012	16	93	.106
	35	.008		94	.096
	36	.004		100	.052
9	34	.102		106	.025
	37	.049		112	.011
	39	.027		113	.009
	42	.010		116	.005
	44	.004	17	104	.103
10	41	.097		105	.095
	44	.053		112	.049
	47	.024		118	.025
	50	.010		125	.010
	52	.005		129	.005
11	48	.103	18	116	.098
	52	.051		124	.049
	55	.027		131	.024
	59	.009		138	.010
	61	.005		143	.005
12	56	.102	19	128	.098
	60	.055		136	.052
	61	.046		137	.048
	64	.026		144	.025
	68	.010		152	.010
	71	.005		157	.005
13	64	.108	20	140	.101
	65	.095		150	.049
	69	.055		158	.024
	70	.047		167	.010
	74	.024		172	.005