Chapter 6
Statistical Distributions

Binomial

We have learned that probability and sampling play a role in statistics. In this chapter we show that probability (frequency) distributions exist for different types of statistics; i.e. the **binomial distribution** (frequency distribution of dichotomous data) and **normal distribution** (frequency distribution of continuous data).

Many variables in education, psychology, and business are dichotomous. Examples of dichotomous variables are: boy versus girl; correct versus incorrect answers; delinquent versus non-delinquent; young versus old; part-time versus full-time worker. These variables reflect mutually exclusive and exhaustive categories (i.e., an individual, object, or event can only occur in one or the other category, but not both). Populations that are divided into two exclusive categories are called **dichotomous populations**, which can be represented by the binomial probability distribution. The derivation of the binomial probability is similar to the combination probability derived in Chap. 2.

The **binomial probability distribution** is computed by:

$$P(x\ in\ n) = \left(\frac{n}{x}\right) P^x Q^{n-x}$$

where the following values are used:

n = size of random sample
x = number of events, objects, or individuals in first category
n – x = number of events, objects, or individuals in second category
P = probability of event, object, or individual occurring in the first category
Q = probability of event, object, or individual occurring in the second category, (1 – P).

Since the **binomial distribution** is a theoretical probability distribution based upon objects, events, or individuals belonging in one of only two groups, the values

R. Schumacker and S. Tomek, *Understanding Statistics Using R*,
DOI 10.1007/978-1-4614-6227-9_6, © Springer Science+Business Media New York 2013

for P and Q probabilities associated with group membership must have some basis for selection. An example will illustrate how to use the formula and interpret the resulting binomial distribution.

Students are given five true–false items. The items are scored correct or incorrect with the probability of a correct guess equal to one-half. What is the probability that a student will get four or more true–false items correct? For this example, **n** = 5, P and Q are both .50 (one-half based on guessing the item correct), and **x** ranges from 0 (all wrong) to 5 (all correct) to produce the binomial probability combinations. The calculation of all binomial probability combinations is not necessary to solve the problem, but tabled for illustration and interpretation.

x	$\binom{n}{x}$	P^x	Q^{n-x}	Probability
5	1	$.5^5$	$.5^0$	1/32 = .03
4	5	$.5^4$	$.5^1$	5/32 = .16
3	10	$.5^3$	$.5^2$	10/32 = .31
2	10	$.5^2$	$.5^3$	10/32 = .31
1	5	$.5^1$	$.5^4$	5/32 = .16
0	1	$.5^0$	$.5^5$	1/32 = .03
				32/32 = 1.00

Using the addition rule, the probability of a student getting four or more items correct is: .16 + .03 = .19. The answer is based on the sum of the probabilities for getting four items correct plus five items correct.

The combination formula yields an individual "coefficient" for taking **x** events, objects, or individuals from a group size **n**. Notice that these individual coefficients sum to the total number of possible combinations and are symmetrical across the binomial distribution. The binomial distribution is symmetrical because P = Q = .50. When P does not equal Q, the binomial distribution will not be symmetrical. Determining the number of possible combinations and multiplying it times P and then Q will yield the theoretical probability for a certain outcome. The individual outcome probabilities add to 1.0.

A binomial distribution can be used to compare sample probabilities to theoretical probabilities if:

a. There are only two outcomes, e.g., success or failure.
b. The process is repeated a fixed number of times.
c. The replications are independent of each other.
d. The probability of success in a group is a fixed value, P.
e. The number of successes, x, in group size n, is of interest.

A binomial distribution based on dichotomous data approximates a normal distribution based on continuous data when the sample size is large and P = .50. Consequently, the mean of a binomial distribution is equal to n*P with variance

equal to n*P*Q. A standardized score (**z-score**), which forms the basis for the normal distribution, can be computed from dichotomous data as follows:

$$z = \frac{x - nP}{\sqrt{nPQ}}$$

where:

x = score
nP = mean
nPQ = variance.

A frequency distribution of standard scores (z-scores) has a mean of zero and a standard deviation of one. The z-scores typically range in value from −3.0 to +3.0 in a symmetrical distribution. A graph of the binomial distribution, given P = Q and a large sample size, will be symmetrical and appear normally distributed.

Knowledge of the binomial distribution is helpful in conducting research and useful in practice. Binomial distributions are skewed except for those with a probability of success equal to .50. If P > .50, the binomial distribution is skewed left; if P < .50, the binomial distribution is skewed right. The mean of a binomial distribution is n*P and the variance is n*P*Q. The binomial distribution given by P(x in n) uses the combination formula, multiplication and addition rules of probability. The binomial distribution can be used to compare sample probabilities to expected theoretical probabilities. For P = .50 and large sample sizes, the binomial distribution approximates the normal distribution.

BINOMIAL R Program

The BINOMIAL program simulates binomial probability outcomes. The number of replications, number of trials, and probability value are input to observe various binomial probability outcomes. Trying different values should allow you to observe the properties of the binomial distribution. The program can be replicated any number of times, but extreme values are not necessary to observe the shape of the distribution. The relative frequencies of **x** successes will be used to obtain the approximations of the binomial probabilities. The theoretical probabilities, mean and variance of the relative frequency distribution, and error will be computed and printed.

The program must specify *numReplications* to indicate the number of replications, *numTrials* to indicate the number of respondents (or sampling points) per replication, and *Probability* to indicate the probability of success (or population proportion). The initial values are set at 5 respondents (sample size or number of trials), 500 replications, and a population proportion of .50. The program starts by defining these values and then creates a random sample from the binomial distribution of size *numReplications* with *numTrials* sampling points per replication and a probability of success, *Probability*.

BINOMIAL Program Output

Given the following values:

```
   Number of Trials=5
   Number of Replications=500
   Probability=0.5

Mean number of successes=2.58
Mean expected number of successes=2.5

Sample variance=1.314
Expected variance=1.25

                Rel. Freq.   Probability    Error
   Successes=0   0.024        0.031          -0.007
   Successes=1   0.170        0.156           0.014
   Successes=2   0.264        0.312          -0.048
   Successes=3   0.322        0.312           0.010
   Successes=4   0.184        0.156           0.028
   Successes=5   0.036        0.031           0.005
```

BINOMIAL Exercises

1. Run BINOMIAL for n=5, **P=.50**, and 500 replications. Enter the results below.

NO. SUCCESSES N	REL FREQ.	PROBABILITY P	ERROR
0	_____	_____	_____
1	_____	_____	_____
2	_____	_____	_____
3	_____	_____	_____
4	_____	_____	_____
5	_____	_____	_____

   a. What is the maximum absolute value of the errors?_____
   b. Use the P (x in n) formula for the binomial distribution to prove P(3 in 5) is correct.
   _____
   c. What is the mean number of successes in the simulation?_____
   d. What is the mean expected number of successes?_____
   e. What is the sample variance?_____

f. What is the expected variance?_____

g. Use the probabilities to calculate the expected mean and variance.
Mean = _____ Variance = _____

2. Run BINOMIAL for n = 5, P = **.30**, and 500 replications. Enter the results below.

NO. SUCCESSES N	REL FREQ.	PROBABILITY P	ERROR
0			
1			
2			
3			
4			
5			

a. What is the maximum absolute value of the errors?_____

b. Use the P (x in n) formula for the binomial distribution to prove P(3 in 5) is correct.
_____

c. What is the mean number of successes in the simulation?_____

d. What is the mean expected number of successes?_____

e. What is the sample variance?_____

f. What is the expected variance?_____

NO. SUCCESSES N	REL FREQ.	PROBABILITY P	ERROR
0			
1			
2			
3			
4			
5			

g. Use the probabilities to calculate the expected mean and variance.
Mean = _____ Variance = _____

3. Run BINOMIAL for n = 5, **P = .70**, and 500 replications. Enter the results below.

a. What is the maximum absolute value of the errors? _____

b. Use the P (x in n) formula for the binomial distribution to prove P(3 in 5) is correct.
_____

c. What is the mean number of successes in the simulation? _____

d. What is the mean expected number of successes? _____

e. What is the sample variance? _____

f. What is the expected variance? _____

g. Use the probabilities to calculate the expected mean and variance.
_____

_____

Normal Distribution

In the seventeenth and eighteenth centuries, two mathematicians were asked by gamblers to help them improve their chances of winning at cards and dice. The two mathematicians, who first studied this area of probability, were James Bernoulli and Abraham DeMoivre. James Bernoulli developed the formula for combinations and permutations, and their binomial expansions, which lead to the binomial distribution. Abraham DeMoivre coined the phrase *law of errors* from observing events such as archery matches. The basic idea was that negative errors occurred about as often as positive errors. DeMoivre used this understanding to derive an equation for an *error curve*. DeMoivre in 1733 was credited with developing the mathematical equation for the *normal curve*. In the nineteenth century, Carl Fredrick Gauss (1777–1855), working in the field of astronomy further developed the concept of a mathematical bell-shaped curve and probability. Today, his picture and mathematical equation for the normal curve appear on the deutsche mark currency of Germany.

The normal distribution or normal curve is a mathematical equation for random chance errors. The frequency distribution of many continuous random variables used in research closely approximates the normal distribution. Consequently, the normal distribution is a useful mathematical model in which to study variable relationships in the physical and social sciences.

The mathematical equation for the normal distribution indicates a normal density that is an exponential function of a quadratic form. The normal curve equation defines an infinite number of curves, depending upon the values of the mean and standard deviation. The normal curve equation is defined as:

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

where:

Y = height of the curve at a given score
X = score at a given height
μ = mean of the X variable
σ = standard deviation of X variable
π = constant equal to 3.1416 (pi)
e = constant equal to 2.7183 (base of natural logarithm)

When a set of X scores are transformed to have a mean of zero (0) and a standard deviation of one (1), which are called standard scores or z-scores, the mathematical equation is reduced to:

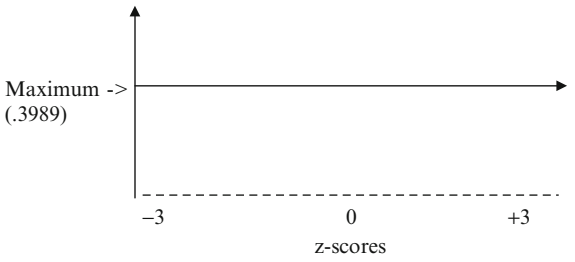
$$Y = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

This equation using standard scores is referred to as the *standard normal curve* with z-score values that range primarily from −3 to +3 and correspond to

the ordinates of Y (density or height of the curve). A z score is calculated as: (X − mean)/standard deviation. The tabled values indicate the z-score values between −3 and +3 corresponding to each y-ordinate value. A graph of these values yields a normal distribution or bell-shaped curve.

Table of z-score and y-ordinate values

z-score	y-ordinates
−3.0	.004
−2.5	.018
−2.0	.054
−1.5	.130
−1.0	.242
−0.5	.352
0.0	.399
+0.5	.352
+1.0	.242
+1.5	.130
+2.0	.054
+2.5	.018
+3.0	.004



The standard normal distribution has a mean of 0, and a standard deviation of 1. The standard normal distribution is bell-shaped and symmetrical. The probability area under the standard normal curve is 1.0 corresponding to 100 % of the area under the curve. The density function is reasonably complex enough that the usual method of finding the probability area between two specified values requires using integral calculus to calculate the probabilities. The standard normal table of z-values in the appendix of textbooks has been derived by this method of integral calculus. Basically, the normal distribution is an exponential function forming a symmetrical curve, with inflection points at −1 and 1. The bell-shaped curve approaches but never touches the X-axis. The normal distribution has an infinite number of different curves based upon the values of the mean and standard deviation. The normal distribution curve using z-scores is called the standard normal curve. The probabilities of the bell-shaped curve are based on the normal distribution.

NORMAL R Program

The NORMAL R program approximates standard normal probabilities. Initial values are set for 1,000 random numbers between the z-score intervals −3 and 3 to correspond to +/−3 standard deviations around the true P value. The *DensityHeight* is fixed at .3989 corresponding to the normal distribution. The normal curve height is calculated as 1/SQRT(2*3.1416) = 0.3989.

The *NumPoints*, *IntervalMin*, and *IntervalMax* variables are user inputted values. *DensityHeight* has a single value since the density function corresponds to a single distribution with z-score values between −3 and +3. These represent the area under the curve for values at the lower end of the interval and the upper end of the interval. These two extra variables are needed because if the lower end of the interval is negative (left of the mean) and the upper end is zero or positive (right of the mean), then the two probabilities are added together to get the total probability. If both are positive, then the lower is subtracted from the upper to leave only the area between the two. If both are negative, then the upper (which would be the smaller absolute difference from the mean) is subtracted from the lower.

The approximation of probabilities in the program are carried out by specifying an interval (A<-->B) and determining the probability, P (A<Z<B). Each approximation is made on the basis of 1,000 random points with an interval A<-->B corresponding to +/−3 and a maximum height of 0.3989. The theoretical probabilities are computed and printed. In order to avoid entering a standard normal probability table into the program, an approximation is used. It was given by Stephen E. Derenzo in "Approximations for Hand Calculators using Small Integer Coefficients," *Mathematics of Computation*, *31*, 1977, pp. 214–225. For A ≥ 0, the approximation is:

$$P(Z < A) = 1 - \frac{1}{2} exp\left[-\frac{((83A + 351)A + 562)A}{703 + 165A} \right]$$

in which 1/2 exp [x] means 1/2 e^x. This approximation has an error of no more than 0.001 for sample sizes of 10,000 or more.

NORMAL Program Output

```
Sample size=1000 Interval width=6

Interval   Sample P   True P   Error
-3<Z<3     0.986      0.997    -0.011
```

NORMAL Distribution Exercises

1. Run the NORMAL program and complete the following table (`NumPoints <- 1000`). Draw small graphs in the last column and shade the area that represents the probability.

Z INTERVAL	APPROXIMATIONS	TRUE PROBABILITY	GRAPH
0<Z<1			
−1<Z<1			
0<Z<2			
−2<Z<2			
0<Z<3			
−3<Z<3			
1.54<Z<2.67			
−0.45<Z<1.15			

2. Run the NORMAL program again with `NumPoints <- 10000` for the Z intervals.

Z INTERVAL	APPROXIMATIONS	TRUE PROBABILITY	GRAPH
0<Z<1			
−1<Z<1			
0<Z<2			
−2<Z<2			
0<Z<3			
−3<Z<3			
1.54<Z<2.67			
−0.45<Z<1.15			

   a. Does `NumPoints <- 1000` or `NumPoints <- 10000` give better approximations?

   _____

   b. In general will larger sample sizes more closely approximate the normal distribution P value?
   YES _____ NO _____

3. Compare the IntervalWidth (−3<Z<+3) and DensityHeight of .3989 to the IntervalWidth (−4<Z<+4) and DensityHeight of .3989 in the NORMAL program using `NumPoints <- 1000`.

Z INTERVAL	APPROXIMATIONS	TRUE PROBABILITY	DENSITY HEIGHT
−3<Z<+3	_____	_____	.3989
−4<Z<+4	_____	_____	.3989

a. Will the approximations be different?

YES _____ NO _____

b. Will the approximations become more similar as sample size increases?

YES _____ NO _____

Chi-Square Distribution

The chi-square distribution, like the other distributions, is a function of sample size. There are an infinite number of chi-square curves based on sample size. In fact, as sample size increases, the chi-square distribution becomes symmetrical and bell-shaped (normal), but with a mean equal to the degrees of freedom (df) and mode equal to df − 2.

The degrees of freedom are related to sample size, because it takes on a value of N − 1. The degree of freedom concept relates to the number of values or parameters free to vary. If a set of five numbers are given, e.g., 5 4 3 2 1, and the sum of the numbers is known, i.e., ΣX = 15, then knowledge of four numbers implies that the fifth number is not free to vary. For example, four out of five numbers are 10, 15, 25, and 45 with ΣX = 100. Since the four numbers sum to 95, the fifth number must be 5 in order for the sum of the five numbers to equal 100. This same principle applies to a set of numbers and the mean.

Karl Pearson first derived the chi-square distribution as a frequency distribution of squared *z-score* values. The chi-square statistic was computed as:

$$\chi^2 = \sum \left(\frac{X_i - M}{\sigma} \right)^2$$

with df = N − 1. A z-score was calculated as z = [(X − Mean)/Standard Deviation]; where X = a raw score, M = the sample mean, and σ = population standard deviation. The *z-score* transformed a raw score into a standard score based on standard deviation units. The population standard deviation (σ) is indicated in the formula, however, a sample standard deviation estimate was generally used because the population value was not typically known.

Chi-square is related to the variance of a sample. If we squared both the numerator and denominator in the previous formula, we would get:

$$\chi^2 = \frac{\Sigma(X - M)^2 / N - 1}{\sigma^2}$$

The numerator of the formula can be expressed as sample variance because $\Sigma(X-M)^2$ represents the sum of squared deviations, denoted as SS, so the sample variance in the numerator can be written as: $S^2 = SS/N-1$. With a little math, $SS = (N-1)S^2$. Consequently, if samples of size N with variances, S^2, are computed from a normal distribution with variance of σ^2, the χ^2 statistic could be written as:

$$\chi^2 = \frac{(N-1)S^2}{\sigma^2}$$

with $N-1$ degrees of freedom. The chi-square statistic is therefore useful in testing whether a sample variance differs significantly from a population variance because it forms a ratio of sample variance to population variance. Since the chi-square distribution reflects this ratio of variances, all chi-square values are positive and range continuously from zero to positive infinity.

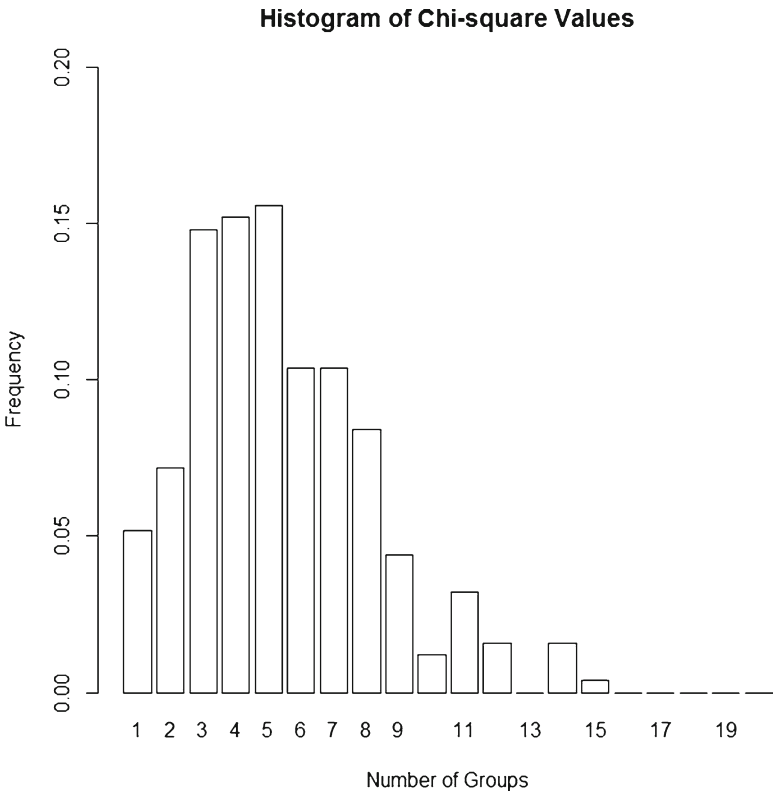
The chi-square statistic, $\chi^2 = (N-1)\ S^2/\sigma^2$, computed by taking random samples from a normal population, produces a different chi-square distribution for each degree of freedom $(N-1)$. The chi-square distribution has a mean equal to the degrees of freedom (df) and a mode equal to $df-2$. The chi-square distribution becomes symmetrical and bell-shaped as sample size increases. The variance of the chi-square distribution is two times the degree of freedom (2*df).

CHISQUARE R Program

The CHISQUARE R program can produce an unlimited number of chi-square distributions, one for each degree of freedom. For small degrees of freedom, the chi-square distribution should be skewed right. As the degrees of freedom increases, that is, sample size increases, the mode of the chi-square distribution moves to the right. The chi-square values should be positive and range continuously from zero to positive infinity. The program permits selection of different sample sizes from a normal distribution. The program will initially select 250 samples of the desired size. Each time, the chi-square statistic will be calculated. The 250 sample chi-square values are graphed to show the chi-square sampling distribution. The chi-square statistics will be recorded in a relative frequency table. A table is printed with the relative frequencies within each interval. The relative frequencies are graphed using the **barplot function** (histogram) to graphically display the chi-square sampling distribution. The group interval and frequency, along with the title and scaling, are set to the default for the Y and X-axis. Finally, the modal chi-square value and the range of the chi-square values are printed.

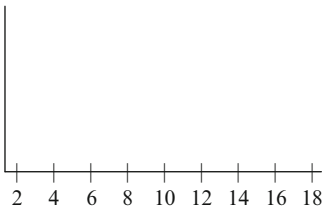
CHISQUARE Program Output

```
Pop. Mean=0
Pop. SD=1
Sample Size=6
N Replications=250

Interval    Rel Freq
( 0.0, 1.0) 0.052
( 1.0, 2.0) 0.072
( 2.0, 3.0) 0.148
( 3.0, 4.0) 0.152
( 4.0, 5.0) 0.156
( 5.0, 6.0) 0.104
( 6.0, 7.0) 0.104
( 7.0, 8.0) 0.084
( 8.0, 9.0) 0.044
( 9.0,10.0) 0.012
(10.0,11.0) 0.032
(11.0,12.0) 0.016
(12.0,13.0) 0.000
(13.0,14.0) 0.016
(14.0,15.0) 0.004
(15.0,16.0) 0.000
(16.0,17.0) 0.000
(17.0,18.0) 0.000
(18.0,19.0) 0.000
(19.0,20.0) 0.000

Modal Group=5 Range of chi-square values=21.1
```

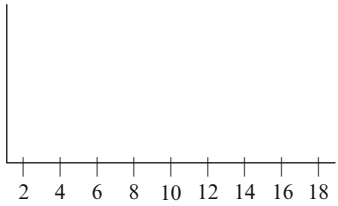


CHISQUARE Exercises

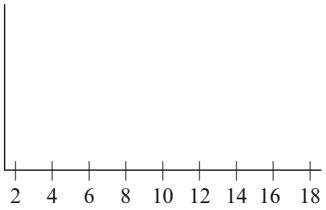
1. Run the CHISQUARE program for each sample size and degrees of freedom listed below; use a population mean of 0, standard deviation of 1, and 250 replications. Graph the shape of the distributions and list the modal group. The modal group is the group with the highest relative frequency.



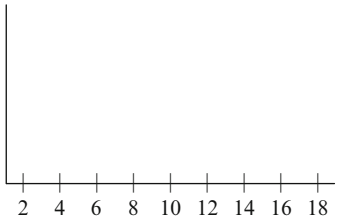
N=2, MODAL GROUP_____



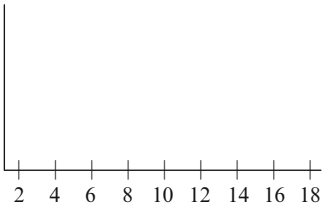
N=3, MODAL GROUP_____



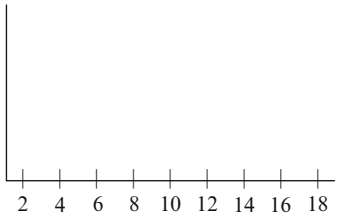
N=4, MODAL GROUP_____



N=5, MODAL GROUP_____



N=6, MODAL GROUP_____



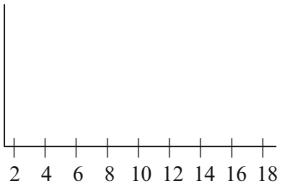
N=7, MODAL GROUP_____

a. Does the shape of the chi-square distribution change as sample size increases? YES _____ NO _____

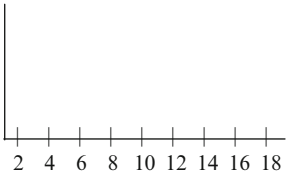
b. List the modal group values from the graphs in the table. The mode of the theoretical chi-square distribution is (DF − 2) when DF > 2. List the theoretical chi-square mode for each sample size. (ERROR = MODAL GROUP − TRUE MODE)

N	DF	MODAL GROUP	TRUE MODE	ERROR
2	1	_____	Not Applicable	_____
3	2	_____	Not Applicable	_____
4	3	_____	_____	_____
5	4	_____	_____	_____
6	5	_____	_____	_____
7	6	_____	_____	_____

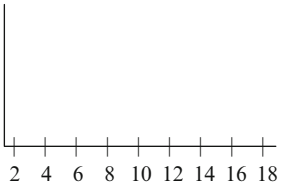
2. Run the CHISQUARE program again for the following sample sizes, but use a population mean of 10, standard deviation of 4, and 250 replications. Graph the shape of the distributions and list the range of the chi-square values. The range is the maximum minus the minimum chi-square value.



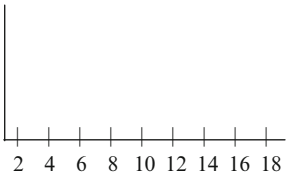
N=5, RANGE_____



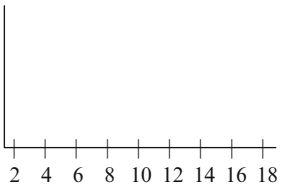
N=10, RANGE_____



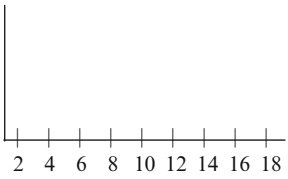
N=15, RANGE_____



N=20, RANGE_____



N=25, RANGE_____



N=30, RANGE_____

a. Compare these graphs with those in Exercise 1. What differences do you see?

_____

_____

b. The variance of the theoretical chi-square distribution is two times the degrees of freedom (2*df). List the theoretical chi-square variance and standard deviation. Divide the range of the chi-square values by four (Range/4) to approximate the standard deviation of the simulated chi-square distribution. (ERROR = APPROXIMATE STANDARD DEVIATION – THEORETICAL STANDARD DEVIATION)

N	σ^2	σ	Range/4	ERROR
5	_____	_____	_____	_____
10	_____	_____	_____	_____
15	_____	_____	_____	_____

N	σ^2	σ	Range/4	ERROR
20	_____	_____	_____	_____
25	_____	_____	_____	_____
30	_____	_____	_____	_____

c. Does the theoretical standard deviation compare to the estimated standard deviation?
YES _____ NO _____

t-Distribution

The early history of statistics involved probability and inference using large samples and the normal distribution. The standard normal curve provided a probability distribution that was bell-shaped for large samples, but was peaked for small samples, which resulted in larger probability areas in the tails of the distribution. At the turn of the century, a chemist named William S. Gossett, who was employed at a brewery in Dublin, Ireland, discovered the inadequacy of the normal curve for small samples. Gossett was concerned with the quality control of the brewing process and took small samples to test the beer, but didn't obtain a normal bell-shaped curve when his results were graphed.

William Gossett empirically established sampling distributions for smaller samples of various sizes using body measurements of 3,000 British criminals. He started with an approximate normal distribution, drew large samples and small samples, to compare the resulting sampling distributions. He quickly discovered that probability distributions for small samples differed markedly from the normal distribution. William Gossett wrote a mathematical expression for these small sample distributions, and in 1908 he published the results under the pen name, "Student." The Student's t-distribution was a major breakthrough in the field of statistics.

The standard normal distribution is bell-shaped, symmetrical, and has a mean of zero and standard deviation of one. The t-distribution is uni-modal, symmetrical, and has a mean of zero, but not a standard deviation of one. The standard deviation of the t-distribution varies, so when small sample sizes are randomly drawn and graphed, the t-distribution is more peaked (leptokurtic). The probability areas in the tails of the t-distribution are consequently higher than those found in the standard normal distribution. For example, the probability area = .046 in the standard normal distribution at two standard deviations from the mean, but the probability area = .140 in the t-distribution at two standard deviations for a sample size of four. This indicates a greater probability of error using smaller samples. As sample sizes become larger, the t-distribution and standard normal distribution take on the same bell-shaped curve. In fact, the t-values and the z-score values become identical around sample sizes of 10,000, which is within .001 error of approximation as indicated in the previous chapter. Researchers today often use the t-distribution for both small

sample and large sample estimation because it becomes identical to the normal distribution as sample size increases.

In many disciplines, such as education, psychology, and business, variable values are normally distributed. Achievement tests, psychological tests, the height or weight of individuals, and the time to complete a task are examples of variables commonly used in these disciplines. In many instances, the population mean and standard deviation for these variables are not known, but rather estimated from sample data. This forms the basis for making an inference about the population parameters (e.g., mean and standard deviation) from the sample statistics (sample mean and standard deviation). Given small random samples, the t-distribution would better estimate the probability under the frequency distribution curve. Given large random samples, both the standard normal distribution and t-distribution would both yield similar probabilities under the frequency distribution curve.

If the population standard deviation is known, the z-score can be computed as:

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

Otherwise, the sample standard deviation is used to compute a t-value:

$$t = \frac{\overline{X} - \mu}{S / \sqrt{n}}$$

The sample standard deviation, *S*, as an estimate of the population standard deviation, σ, is typically in error, thus the sample means are not distributed as a standard normal distribution, but rather a t-distribution. When sample sizes are larger, the sample standard deviation estimate becomes similar to the population standard deviation. Consequently, the shape of the t-distribution is similar to the standard normal distribution. This points out why the estimate of the population standard deviation is critical in the field of statistics. Many researchers attempt to estimate better the unknown population standard deviation by one of the following methods:

1. Use test publisher norms when available (μ, σ)
2. Take an average value from several research studies using the same variable
3. Take large samples of data for better representation
4. Divide the range of sample data by six (see Chap. 5)

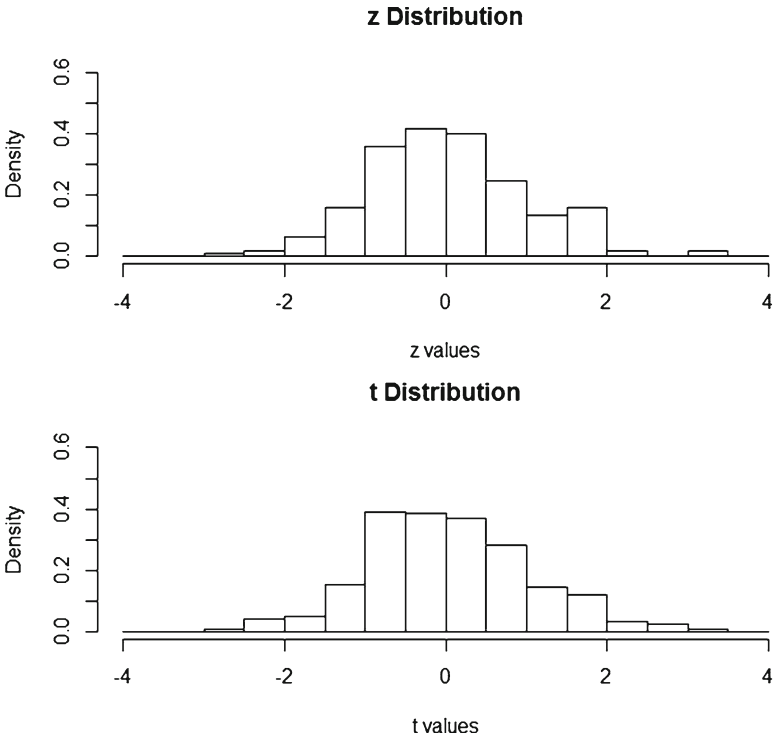
The t-distribution is symmetrical, unimodal, and has a mean of zero. The t-distribution has a greater probability area in its tails than the standard normal distribution due to sample estimation of the population standard deviation. The shape of the t-distribution is not affected by the mean and variance of the population from which random sampling occurs. As the sample size increases, the t-distribution becomes similar to the standard normal distribution.

t-DISTRIBUTION R Program

The t-DISTRIBUTION program creates a z distribution of z values and a t distribution of t-values. The program specifies a population mean and standard deviation, sample size, and the number of replications (samples to be taken), which are initially set but can be changed. The program then selects a random sample of that size, computes the sample mean and sample standard deviation, and then the z- and t-statistics. This process will be repeated 250 times. The 250 z- and t-statistics, which arise from these simulations, will be tabulated and printed in a frequency table. By comparing the frequency tables for t and z, you will be able to observe the higher probability in the heavier tails of the t-distribution. By varying the sample size, you will be able to observe how the shape of the t-distribution changes and becomes more normally distributed as the sample size increases.

t-DISTRIBUTION Program Output

```
Pop. Mean=50 Pop. SD=15
Sample Size=30
N Replications=250

Interval        Freq t    Freq z
(-4.0,-3.5)     0.000     0.000
(-3.5,-3.0)     0.000     0.000
(-3.0,-2.5)     0.004     0.004
(-2.5,-2.0)     0.020     0.008
(-2.0,-1.5)     0.024     0.032
(-1.5,-1.0)     0.076     0.080
(-1.0,-0.5)     0.196     0.180
(-0.5, 0.0)     0.192     0.208
( 0.0, 0.5)     0.184     0.200
( 0.5, 1.0)     0.140     0.124
( 1.0, 1.5)     0.072     0.068
( 1.5, 2.0)     0.060     0.080
( 2.0, 2.5)     0.016     0.008
( 2.5, 3.0)     0.012     0.000
( 3.0, 3.5)     0.004     0.008
( 3.5, 4.0)     0.000     0.000
```



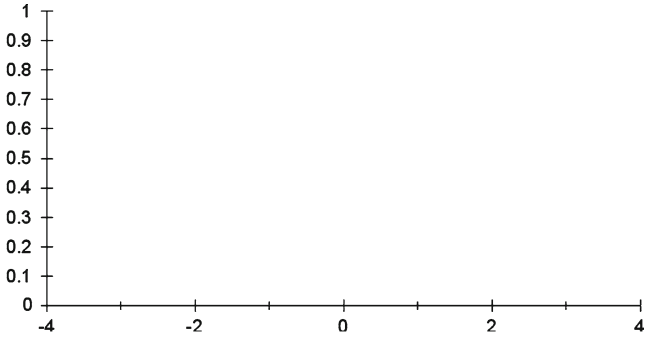
t-DISTRIBUTION Exercises

1. Run t-DISTRIBUTION for population mean of 0 and a standard deviation of 1. Use a sample size of 5 and perform 1,000 replications (`popMean <- 0, popStdDev <- 1, sampleSize <- 5, replicationSize <- 1000`). Record the results below.

INTERVAL	FREQ t	FREQ z
(−4.0, −3.5)	_____	_____
(−3.5, −3.0)	_____	_____
(−3.0, −2.5)	_____	_____
(−2.5, −2.0)	_____	_____
(−2.0, −1.5)	_____	_____
(−1.5, −1.0)	_____	_____
(−1.0, −0.5)	_____	_____
(−0.5, 0.0)	_____	_____
(0.0, 0.5)	_____	_____
(0.5, 1.0)	_____	_____

INTERVAL	FREQ t	FREQ z
(1.0, 1.5)	_____	_____
(1.5, 2.0)	_____	_____
(2.0, 2.5)	_____	_____
(2.5, 3.0)	_____	_____
(3.0, 3.5)	_____	_____
(3.5, 4.0)	_____	_____

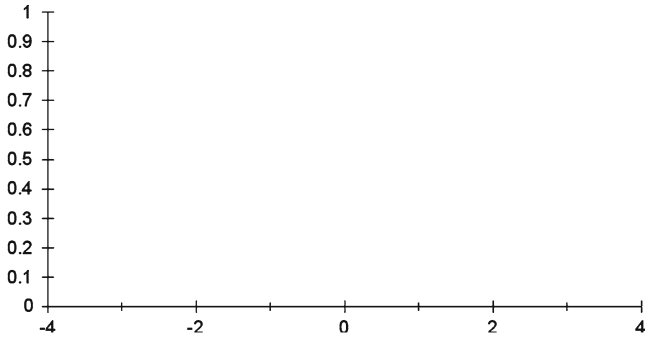
a. Does the t-statistic distribution have higher frequencies in the tails of the distribution than the z-statistic distribution? YES _____ NO _____
b. Graph the z-statistic distribution with a *solid line* and the t-statistic distribution with a *dashed line*. Are the two distributions the same? YES _____ NO _____



2. Run t-DISTRIBUTION again for population mean of 0 and a standard deviation of 1. Use a sample size of 100 and perform 1,000 replications (`popMean <- 0, popStdDev <- 1, sampleSize <- 100, replicationSize <- 1000`). Record the results below.

INTERVAL	FREQ t	FREQ z
(–4.0, –3.5)	_____	_____
(–3.5, –3.0)	_____	_____
(–3.0, –2.5)	_____	_____
(–2.5, –2.0)	_____	_____
(–2.0, –1.5)	_____	_____
(–1.5, –1.0)	_____	_____
(–1.0, –0.5)	_____	_____
(–0.5, 0.0)	_____	_____
(0.0, 0.5)	_____	_____
(0.5, 1.0)	_____	_____
(1.0, 1.5)	_____	_____
(1.5, 2.0)	_____	_____
(2.0, 2.5)	_____	_____
(2.5, 3.0)	_____	_____
(3.0, 3.5)	_____	_____
(3.5, 4.0)	_____	_____

a. Does the t-statistic distribution have higher frequencies in the tails of the distribution than the z-statistic distribution? YES _____ NO _____
b. Graph the z-statistic distribution with a *solid line* and the t-statistic distribution with a *dashed line*. Are the two distributions the same? YES _____ NO _____
c. As sample size increased from n = 5 to n = 100, did the t-statistic distribution more closely approximate a normal distribution? YES _____ NO _____

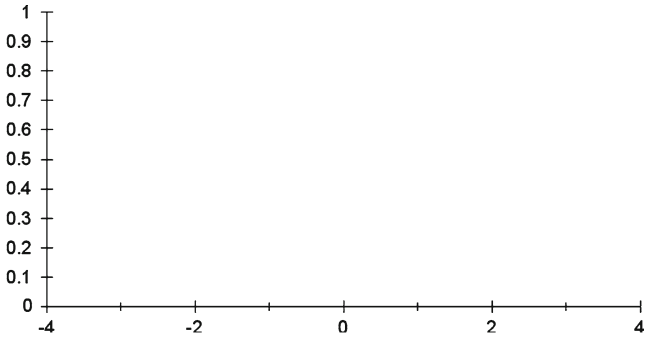


3. Run t DISTRIBUTION again for population mean of 0 and a standard deviation of 15. Use a sample size of 5 and perform 1,000 replications (`popMean <- 0, popStdDev <- 15, sampleSize <- 5, replicationSize <- 1000`). Record the results below.

INTERVAL	FREQ t	FREQ z
(–4.0, –3.5)	_____	_____
(–3.5, –3.0)	_____	_____
(–3.0, –2.5)	_____	_____
(–2.5, –2.0)	_____	_____
(–2.0, –1.5)	_____	_____
(–1.5, –1.0)	_____	_____
(–1.0, –0.5)	_____	_____
(–0.5, 0.0)	_____	_____
(0.0, 0.5)	_____	_____
(0.5, 1.0)	_____	_____
(1.0, 1.5)	_____	_____
(1.5, 2.0)	_____	_____
(2.0, 2.5)	_____	_____
(2.5, 3.0)	_____	_____
(3.0, 3.5)	_____	_____
(3.5, 4.0)	_____	_____

a. Does the t-statistic distribution have higher frequencies in the tails of the distribution than the z-statistic distribution? YES _____ NO _____

b. Graph the z-statistic distribution with a *solid line* and the t-statistic distribution with a *dashed line*. Are the two distributions the same? YES _____ NO _____
c. Is the t-statistic distribution affected by the population standard deviation value? YES _____ NO _____
d. Is the t-statistic distribution affected by the population mean value? YES _____ NO _____

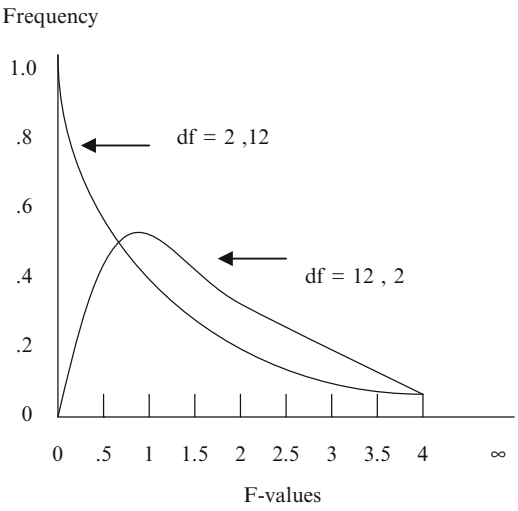


F-Distribution

Sir Ronald Fisher was interested in extending our knowledge of testing mean differences to an analysis of the variability of scores, i.e., variance. He was specifically interested in comparing the variance of two random samples of data. For example, if a random sample of data was drawn from one population and a second random sample of data was drawn from a second population, the two sample variances could be compared as an **F-ratio**: $F = S^2_1/S^2_2$. The F-ratio is equal to one if the variances of the two random samples are the same. The F-distribution in the appendix reveals this for F-values with df = ∞, ∞ in the numerator and denominator. The F-ratio could be less than one, depending upon which sample variances were in the numerator and denominator, but F-values less than one are not considered, so we always place the larger sample variance in the numerator.

If several random samples of data were drawn from each population and the F-ratio computed on the variances for each pair of samples, a sampling distribution of the F's would create the F-distribution. Sir Ronald Fisher determined that like the t-distribution and chi-square distribution, the F-distribution was a function of sample size; specifically the sizes of the two random samples. Consequently, a family of F-curves can be formed based on the degrees of freedom in the numerator and denominator. An **F-curve** is positively skewed with F-ratio values ranging from zero to infinity (∞). If the degrees of freedom for both samples are large, then the F-distribution approaches symmetry (bell-shaped).

Example F-curves for certain degree of freedom pairs can be illustrated as:



Because there are two degrees of freedom associated with an F-ratio, F-tables were constructed to list the F-values expected by chance with the degrees of freedom for the numerator across the top (column values) and the degrees of freedom for the denominator along the side (row values). The corresponding intersection of a column and row degrees of freedom would indicate the tabled F-value. If the computed F-value is greater than the tabled F-value, we conclude that the two sample variances are statistically different at a specified level of probability, e.g., .05 level of significance.

Relationship of F-Distribution to Chi-Square Distribution and t-Distribution

In previous chapters, distributions were graphed based on various sample sizes. We learned that with large sample sizes, sample estimates were closer to population values (z-values) and the sampling distributions (frequency distributions) were more normally distributed. Similarly, the chi-square and t-distributions are also a function of sample size. In fact, as sample size increases, the t, z, and chi-square sampling distributions became symmetrical and bell-shaped (normal). The sampling distributions of the F-ratio operate similar to the t, z, and chi-square family of curves based on sample size.

The t-distribution with degrees of freedom equal to infinity is the normal distribution. Consequently, t-values become equal to z-values when sample sizes are large ($n > 10{,}000$ to infinity). Check this by referring to the last row of the tabled t-values in the Appendix where you will find that the t-values are the same as the z-values in the normal distribution table. For example, $t = 1.96$ is equal to $z = 1.96$ at the .05 level of significance. The normal distribution can be considered a special case of the t-distribution, because as sample size increases, the t-distribution becomes the normal distribution, i.e., t-values = z-values.

The F-distribution, with *one* degree of freedom in the numerator and the same degree of freedom in the denominator as the t-test, is equal to the square of the t-distribution value. To check this, refer to the first column of the tabled F-values ($df_1 = 1$) in the Appendix where you will find that the F-values are the square of the t-values in the t-test table (df_2 = degrees of freedom for t-test). For example, if $F = 3.84$, then $t = 1.96$ for $df_1 = 1$ and $df_2 = \infty$. In fact, since $t^2 = z^2 = F$ for one degree of freedom given large samples, the t-distribution and normal distribution are special cases of the F-distribution.

The F-distribution values, with degrees of freedom in the denominator equal to infinity, can be multiplied by the F-value numerator degrees of freedom to compute a chi-square value. To check this, refer to the last row of the tabled F-values in the Appendix where you will find that the F-values multiplied by the corresponding numerator degrees of freedom (df_1) equals the chi-square value in the chi-square distribution table. For example, $F = 2.21$ for $df_1 = 5$ and $df_2 = \infty$, therefore, chi-square = 11.05 with 5 degrees of freedom, i.e., $5*2.21 = 11.05$! Consequently, the chi-square distribution is also a special case of the F-distribution.

Test of Difference Between Two Independent Variances

The sampling distribution of the F-ratio of two variances cannot be approximated by the normal, t, or chi-square sampling distributions because sample sizes seldom approach infinity, and unlike the normal and t-distributions, F sampling distributions range from zero to infinity rather than from negative infinity to positive infinity. Consequently, the F-distribution, named after Sir Ronald A. Fisher, is used to test whether two independent sample variances are the same or different.

If the variances from two randomly drawn samples are equal, then the F-ratio equals one (largest sample variance over the smallest sample variance), otherwise it increases positively to infinity. The ratio of the two sample variances is expressed as $F = S^2_1/S^2_2$, with a numerator and denominator degrees of freedom. For example, the distance *twenty suburban* housewives traveled to the grocery store varied by 2 miles and the distance *ten rural* housewives traveled to the grocery store varied by 10 miles. We want to test if the suburban and rural mileage variance is equal: $F = 10/2 = 5.0$ with $df_1 = 9$ and $df_2 = 19$ (Note: degrees of freedom are one less than the respective sample sizes). We compare this computed F-ratio to the tabled F-value in the Appendix for a given level of significance, e.g., .01 level of significance. We find the 9 degrees of freedom (df_1) across the top of the F-table and the 19 degrees of freedom (df_2) along the side of the table. The intersection of the column and row indicates an F-value equal to 3.52. Since $F = 5.0$ is greater than tabled $F = 3.52$, we conclude that the *rural* housewives vary more in their mileage to the grocery store than *suburban* housewives. Another way of saying this is that the sample variances are not homogeneous (equal) across the two groups.

Since we conducted this test for only the larger variance in the numerator of the F-ratio, we must make a correction to the level of significance. This is accomplished for any tabled F-value by simply doubling the level of significance, e.g., .01 to .02 level of significance. Therefore, $F = 5.0$ is statistically different from the tabled

F = 3.52 at the .02 level of significance (even though we looked up the F value in the .01 level of significance table).

Test of Difference Between Several Independent Variances

H.O. Hartley extended the F-ratio test to the situation in which three or more sample variances were present, which was aptly named the *Hartley F-max test*. A separate F-max distribution table was therefore created (see Appendix). The Hartley F-max test is limited to using equal sample sizes and sample data randomly drawn from a normal population. However, Henry Winkler in 1967 at Ohio University compared the Hartley F-max, Bartlett, Cochran, and Levene's tests for equal variances in his master's thesis and concluded that the Hartley F-max test was the most robust (best choice) when sample sizes were equal.

Extending our previous example, the distance *twenty-one suburban* housewives traveled to the grocery store varied by 2 miles, the distance *twenty-one rural* housewives traveled to the grocery store varied by 10 miles, and the distance *twenty-one urban* housewives traveled to the grocery store varied by 5 miles. The Hartley F-max test is computed for the following example as follows:

Step 1: Calculate the sample variances.

Urban	$S^2 = 5$
Suburban	$S^2 = 2$
Rural	$S^2 = 10$

Step 2: Calculate F-max test by placing largest variance over smallest variance.

F-max = 10/2 = 5.0

Step 3: Determine the two separate degrees of freedom for the F-max Table.

k = number of sample variances (column values in table)
k = 3
df = sample size − 1 (row values in the table)
df = 21 − 1 = 20

Step 4: Compare computed F-max to tabled F-max values.

F-max = 5.0 is greater than tabled F-max = 2.95 for 3 and 20 degrees of freedom at the .05 level of significance. We conclude that the sample variances are *not* homogeneous (not the same) across the three groups.

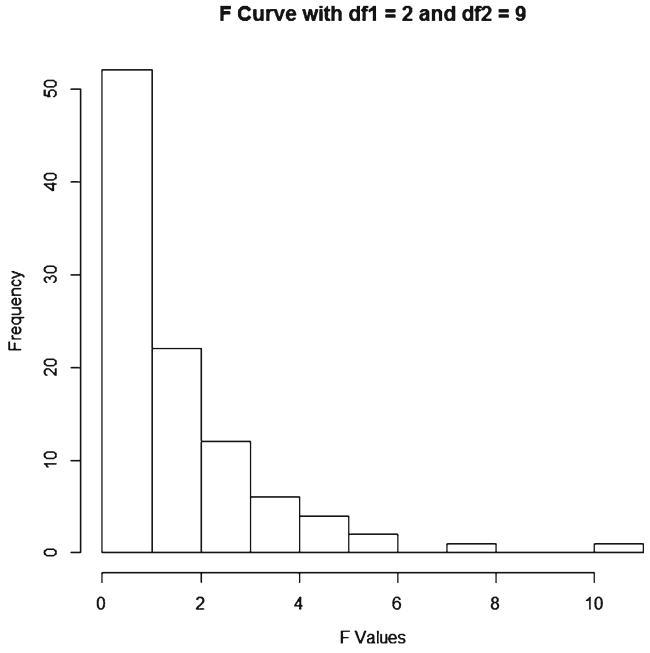
In summary, we find that the F-distribution is a family of frequency curves based upon sample size. The normal (z), t, and chi-square distributions are special cases of the F-distribution. The F-ratio can test whether two independent sample variances are homogeneous. The F-max can test whether three or more independent sample variances are homogeneous. The F-distribution is positively skewed for different numerator and denominator degrees of freedom. As sample sizes increase, the shape of the F-distribution becomes symmetrical.

F-DISTRIBUTION R Programs

The **F-Curve** program simulates F-distributions for given degrees of freedom. It begins by defining the degrees of freedom for the numerator (df1) and denominator (df2) of an F-ratio. Next, the number of replications for the simulation is defined and the random F-values for that number of replications is taken from an F-distribution with the given degrees of freedom. These F-values are plotted in a histogram to show a representation of the F-curve. The F-curve will vary depending upon the degrees of freedom entered in the program.

The **F-Ratio** program inputs the group sizes and variances of two groups, as well as, the alpha level for the significance test. Next the F-ratio is calculated and placed into a display string along with a representation of the actual ratio as a fraction. Then the critical F is determined using the *qf* function, based on the alpha level and degrees of freedom. If the F-ratio is greater than the critical F then the decision is set to "reject," otherwise it stays at "accept." Finally, the display string for the F-ratio, the critical F-value, and the decision are placed into a matrix, labels are applied, and the matrix is displayed.

F-Curve Program Output



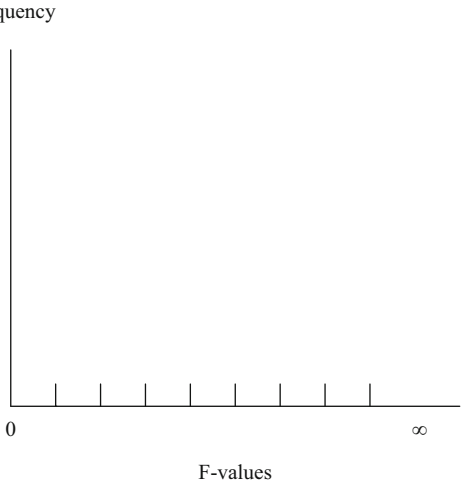
F-Ratio Program Output

```
Sample 1: Size=20 Variance=10
Sample 2: Size=20 Variance=10
alpha=0.01

F ratio Tabled F Decision
10/10=1 3.03 accept
```

F-DISTRIBUTION Exercises

1. Run the **F-Curve** program for each pair of degrees of freedom listed below based on 100 replications. Graph the F-curves on the chart. Note: The two samples are randomly drawn from a normal population, sample variances calculated, and the F-ratio computed.

	Sample 1	Sample 2
	df_1	df_2
Run 1:	5	15
Run 2:	15	15
Run 3:	15	5



   a. Does the shape of the F-curve change based on the numerator and denominator degrees of freedom? YES _____ NO _____
   b. Are the F-values always positive? YES _____ NO _____

c. Why are some of the F-values below 1.0?

_____

_____

2. Run the **F-Ratio** program for the following pairs of sample variances listed below. List the F-ratio, Tabled F-value, and Decision at the .01 level of significance. The *Decision* indicates whether the sample variances are homogeneous.

	Sample 1	Sample 2	F-ratio	Tabled F	Decision
Run 1	n = 20 $S^2 = 10$	n = 20 $S^2 = 10$	_____	_____	_____
Run 2	n = 20 $S^2 = 10$	n = 20 $S^2 = 100$	_____	_____	_____
Run 3	n = 40 $S^2 = 100$	n = 20 $S^2 = 10$	_____	_____	_____
Run 4	n = 20 $S^2 = 10$	n = 40 $S^2 = 100$	_____	_____	_____
Run 5	n = 20 $S^2 = 100$	n = 40 $S^2 = 10$	_____	_____	_____

a. Does the sample size affect the accept or reject decision? YES _____ NO _____

b. Can you estimate the sample variance ratio that would yield an F-value, which would lead to a reject decision? YES _____ NO _____

c. Explain how you determined that the ratio of sample variances led to a reject decision.

_____

_____

_____

3. For the following list of sample variances, compute the F-max test. Find the Tabled F-max value at the .01 level of significance. Decide whether sample variances are homogeneous. Sample size is n = 31 fr each sample.

	Sample 1	Sample 2	Sample 3	F-max	Tabled F	Decision
a.	$S^2 = 10$	$S^2 = 10$	$S^2 = 10$	_____	_____	_____
b.	$S^2 = 10$	$S^2 = 100$	$S^2 = 100$	_____	_____	_____
c.	$S^2 = 40$	$S^2 = 10$	$S^2 = 20$	_____	_____	_____

True or False Questions

Binomial Distribution

T F a. All binomial distributions are symmetrical.
T F b. If n = 10 and P = .50, then 50 % of the time x = 5.
T F c. The binomial distribution approximates the normal distribution when sample size is large and P = .50.
T F d. If a binomial process consists of n trials, then the number of successes, x, will range from 0 to n.
T F e. A binomial distribution is created based on dichotomous variables.
T F f. As sample size increases for P = .50, the mean of the binomial distribution (nP) more closely approximates the population mean.
T F g. As the number of replications increase the absolute value of the error decreases.
T F h. If P < .50, the binomial distribution is skewed right.
T F i. If P > .50, the binomial distribution is skewed left.

Normal Distribution

T F a. The standard normal distribution is a skewed distribution.
T F b. The value of the standard normal density Y at point Z is the probability that the random variable has a value equal to Z.
T F c. The standard normal distribution has a mean of 0, and standard deviation of 1.
T F d. The probability area under the standard normal curve can be approximated in the interval −4 to +4.
T F e. The Monte Carlo approximations of the standard normal probabilities are close to the integral calculus exact theoretical probabilities.
T F f. As sample size increases, the probabilities more closely approximate a standard normal distribution.

Chi-Square Distribution

T F a. The chi-square distribution is independent of the mean and variance of the normal distribution.
T F b. As the degrees of freedom increases, the variance of the chi-square distribution decreases.
T F c. The location of the mode in a chi-square distribution moves to the left as sample size increases.
T F d. For small degrees of freedom, the chi-square distribution is skewed right.
T F e. Some chi-square values are negative.

t-Distribution

T F a. The shape of the t-distribution depends on the standard deviation of the population distribution.
T F b. The smaller the sample size, the larger the probability area in the tails of the t-distribution.
T F c. The population mean value has *no* effect on the shape of the t-distribution.
T F d. The t-distribution is symmetrical, unimodal, and mean of zero.
T F e. For large sample sizes, the t-distribution is the same as the standard normal distribution.
T F f. The z-statistic distribution will always be normally distributed.

F-Distribution

T F a. The normal(z), t, and chi-square distributions are special cases of the F-distribution.
T F b. As sample size increases for both samples, the F-curve becomes symmetrical.
T F c. The F-ratio tests whether two sample variances are homogeneous.
T F d. The word “homogeneous” implies that sample variances are different.
T F e. The F-distribution ranges from zero to infinity.
T F f. The Hartley F-max test requires equal sample sizes in groups.
T F g. The F Ratio program could be used to compute a Hartley F-max test.
T F h. Sample size affects the F-ratio test of equal variances.