

STATISTICAL FUNCTIONS IN EXCEL

In this appendix, we discuss the statistical functions in Excel that are useful for the material in this text. They will be referenced by Chapter within the text.

The statistical functions can be accessed by clicking on f_x at the main menu and then clicking on Statistical Functions in the next submenu. At the next submenu you click on the statistical function you want. Each statistical function has its own submenu where the user supplies the appropriate parameters for the function to work on. These parameters usually include column locations within the spreadsheet where the data lie and in some cases other parameters needed by the function. Before clicking on f_x, the cell where the result from using the function should be placed should be the active cell in the spreadsheet. A nice feature of Excel is that if a cell is made active, then the formula bar displays the function used to compute the result for that cell. Thus, all calculations are self-documented for further reference. There are other statistical commands that are available in the Tools menu, Data Analysis sub-menu. There are beyond the scope of this appendix.

CHAPTER 2 Descriptive Statistics

We have entered the cholesterol levels before adopting a vegetarian diet from Table 2.1 (study guide, Chapter 2) in Table A.1 in cells C5–C28. To compute the arithmetic mean, click on f_x, Statistical Functions, and then Average. Another screen will appear asking for the range of columns defining the dataset where the average is to be computed. Since the 1st value is in C5 and the last value is in C28, we enter C5:C28. The arithmetic mean is displayed to the right of Average(C5:C28). A similar approach is used for the median (median), standard deviation (Stdev), Variance (Var) and Geometric Mean (GeoMean). To compute specific percentiles, we specify the range of data columns (i.e., C5:C28) and then the percentile desired which must be between 0 and 1. The 10th (153.1) and 90th (237.4) percentiles are given in the spreadsheet.

CHAPTER 4 Discrete Probability Distributions

BINOMDIST

BINOMDIST is the statistical function used to compute individual probabilities and cumulative probabilities for the binomial distribution. Suppose we want to compute the probability of 5 successes (k) in 26 trials (n), where the probability of success on one trial is .34 (p). The function BINOMDIST requires

4 parameters to perform this computation. The 1st parameter is k; the 2nd parameter is n; the 3rd parameter is p; the 4th parameter is the word TRUE if a cumulative probability is desired or the word FALSE if an individual probability is desired. Thus, in this case, we specify BINOMDIST(5, 26, .34, FALSE) = .049 (see Table A.2). If instead we want the probability of obtaining ≤ 5 successes, then we specify BINOMDIST(5, 26, .34, TRUE) = .079. To obtain the probability of ≥ 6 successes, we simply subtract from l and obtain .921.

Table A.1 Cholesterol data before adopting a vegetarian diet from Table 2.1

	195
	145
	205
	159
	244
	166
	250
	236
	192
	224
	238
	197
	169
	158
	151
	197
	180
	222
	168
	168
	167
	161
	178
	137
Average (C5:C28)	187.7917
Median (C5:C28)	179
Stdev (C5:C28)	33.15967
Var (C5:C28)	1099.563
GeoMean (C5:C28)	185.0761
Percentile (C5:C28, 0.1)	153.1
Percentile (C5:C28, 0.9)	237.4

Table A.2 Example of BINOMDIST and POISSON

Example of BINOMDIST				
n	k	p		
26	5	0.34	BINOMDIST(5,26,.34, FALSE)	0.048519
26	≤5	0.34	BINOMDIST(5,26,0.34,TRUE)	0.079197
26	≥6	0.34	1-BINOMDIST(5,26,0.34,TRUE)	0.920803
Example of POISSON				
μ	k			
7.4	3		POISSON(3,7.4, FALSE)	0.041282
7.4	≤3		POISSON(3,7.4,TRUE)	0.063153
7.4	≥4		1-POISSON(3,7.4,TRUE)	0.936847

POISSON

This function works similarly to BINOMDIST. There are 3 parameters, μ, k and (TRUE/FALSE). Suppose we want to compute $Pr(X = 3|\mu = 7.4)$. This is given by POISSON(3, 7.4, FALSE) = .041. (See Table A.2). If instead we want $Pr(X \le 3|\mu = 7.4)$, then we use POISSON(3, 7.4, TRUE) = .063. *Note*: If we want $Pr(X \ge 4|\mu = 7.4)$, then we compute 1 – POISSON(3, 7.4, TRUE)= .937. (not 1 – POISSON(4, 7.4, TRUE))

CHAPTER 5 Continuous Probability Distributions

There are 4 commands in Excel that are associated with the normal distribution.

NORMDIST

NORMDIST can calculate the pdf and cdf for *any* normal distribution. The parameters of NORMDIST are

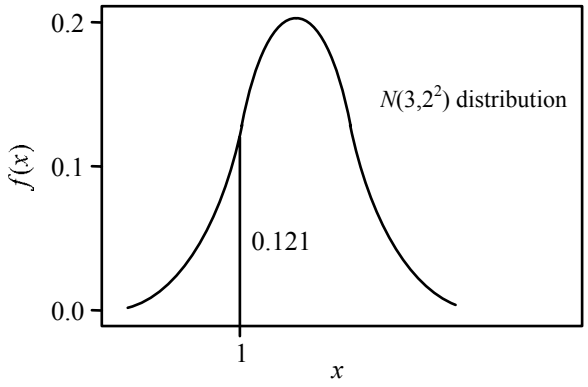
NORMDIST(x, mean, sd, TYPE)

Suppose we want to evaluate the pdf at $x = 1.0$ for a normal distribution with mean = 3 and sd = 2. This is given by

$$\text{NORMDIST(1.0, 3, 2, FALSE)= pdf of } N(3,\ 2^2) \text{ distribution evaluated at } x = 1.0$$
$$= \frac{1}{\sqrt{2\pi}(2)} \exp\left[-\frac{1}{2}\left(\frac{1-3}{2}\right)^2\right] = 0.121$$

(See Figure A.1).

FIGURE A.1

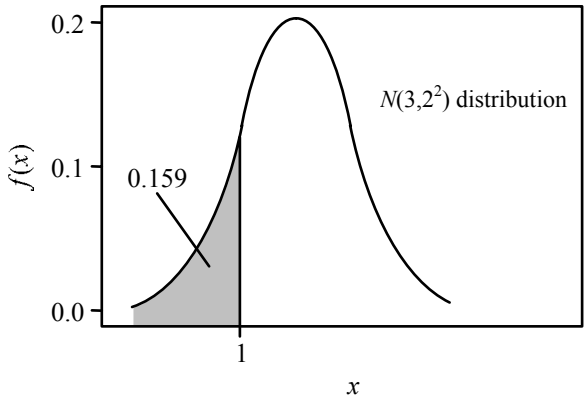


Suppose we want to evaluate the cdf at $x = 1.0$ for a normal distribution with mean = 3 and sd = 2. This is given by:

$$\text{NORMDIST(1.0, 3, 2, TRUE)= cdf of a } N(3, 2) \text{ distribution evaluated at } x = 1.0$$
$$= Pr\left[x \le 1.0 | X \sim N(3,\ 2^2)\right]$$
$$= \Phi\left(\frac{1-3}{2}\right) = \Phi(-1) = .159$$

(See Figure A.2).

FIGURE A.2



NORMINV

NORMINV can calculate percentiles for any normal distribution. The parameters of NORMINV are

$$\text{NORMINV}(p, \text{mean, sd}) = \text{the value } x \text{ such that } Pr[X \le x | X \sim N(\text{mean}, sd^2)] = p$$

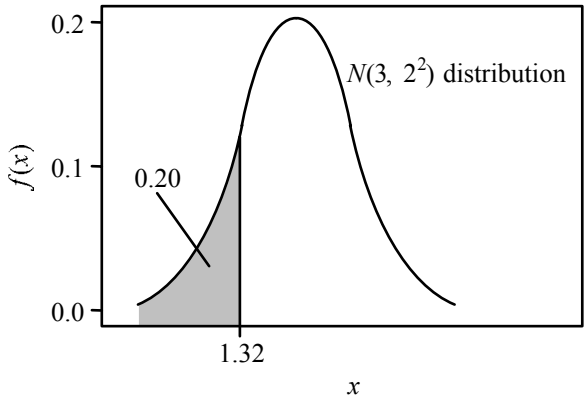
Suppose we want to calculate the 20th percentile of a $N(3,\ 2^2)$ distribution. This is given by

$$\text{NORMINV}(.2, 3, 2) = \text{the value } x \text{ such that } Pr[X \le x | X \sim N(3,\ 2^2)] = 0.2.$$

The corresponding x-value is 1.32.

(See Figure A.3).

FIGURE A.3



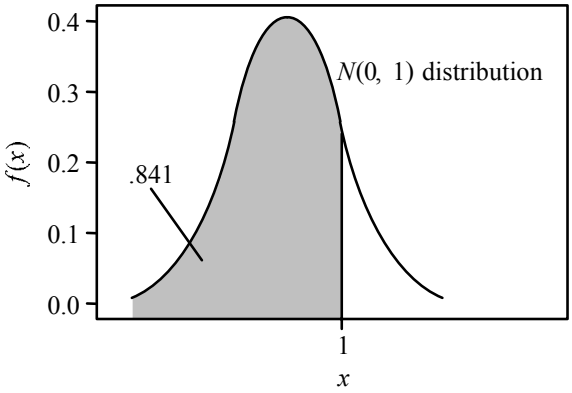
NORMSDIST

NORMSDIST can calculate the cdf for a standard normal distribution. NORMSDIST has a single parameter x, where

$$\text{NORMSDIST}(x) = Pr[X \le x | X \sim N(0,\ 1)]$$

For example, NORMSDIST(1) = $\Phi(1)$ = .841. (See Figure A.4).

FIGURE A.4



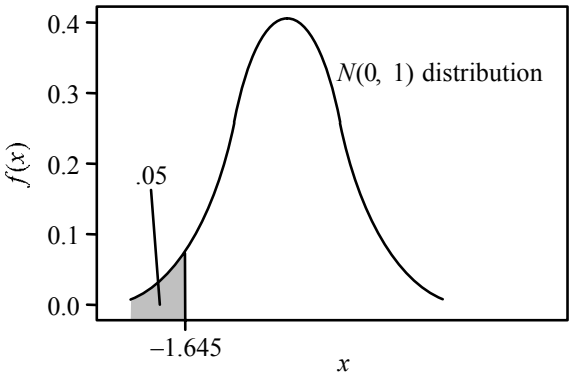
NORMSINV

NORMSINV can calculate percentiles for a standard normal distribution. NORMSINV has a single parameter p, where

$$\text{NORMSINV}(p) = z_p = \text{the value } z \text{ such that } Pr[Z \le z | Z \sim N(0, 1)] = p$$

For example, $\text{NORMSINV}(.05) = z_{.05} = -1.645$. (See Figure A.5).

FIGURE A.5



The spreadsheet below (Table A.3) provides illustrations of all four of the normal distribution functions of Excel.

Table A.3 Illustration of normal distribution commands of Excel

NORMDIST(l.0, 3, 2, FALSE)	0.120985
NORMDIST(l.0, 3, 2, TRUE)	0.158655
NORMINV(.2, 3, 2)	1.316757
NORMSDIST(1)	0.841345
NORMSINV(.05)	−1.64485

CHAPTER 6 Estimation

Excel has two commands for use with the t distribution and two commands for use with the chi-square distribution.

t Distribution Commands

TDIST and TINV are the 2 commands for use with the *t* distribution.

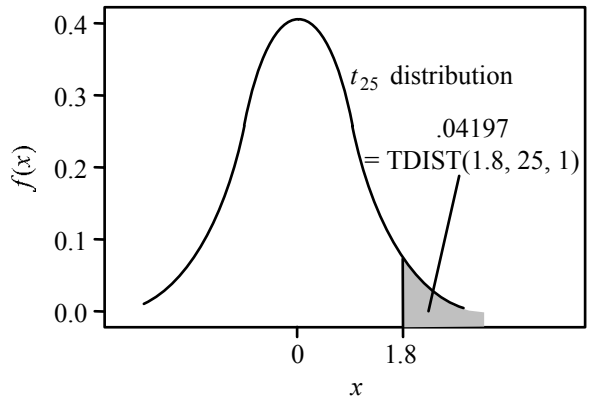
TDIST

TDIST is used to calculate tail areas for a *t* distribution. Suppose we want to calculate $Pr(t \geq 1.8 \mid t \sim t_{25})$. We can compute this by specifying TDIST (1.8, 25, 1) and obtain .04197 (see Table A.4 and Figure A.6). The 1st argument (1.8) is the *t* value, the 2nd argument (25) is the d.f. and the 3rd argument (1) specifies that a one-tailed area is desired.

Table A.4 Use of *t* distribution commands in Excel

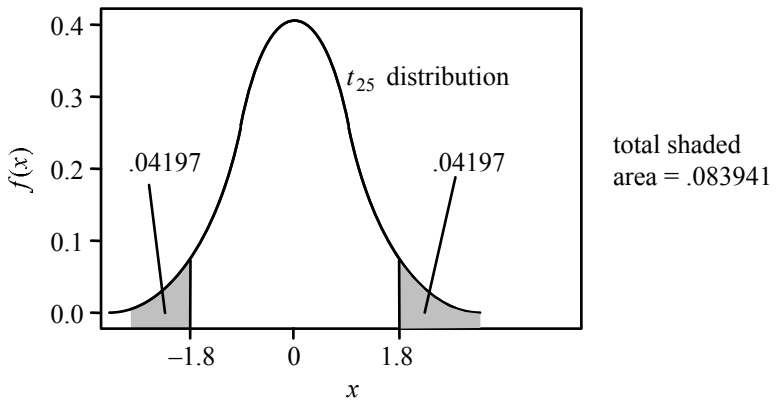
TDIST	TDIST(1.8, 25, 1)	0.04197
	TDIST(1.8, 25, 2)	0.083941
TINV	TINV(.2, 93)	1.290721

FIGURE A.6



In some instances, particularly in hypothesis testing, we will want 2-tailed areas such as $Pr(|t| \geq 1.8 | t \sim t_{25})$. This is obtained by specifying TDIST(1.8, 25, 2) = .083941 (see Table A.4 and Figure A.7). Please note that the 1st argument of TDIST can only be a positive number.

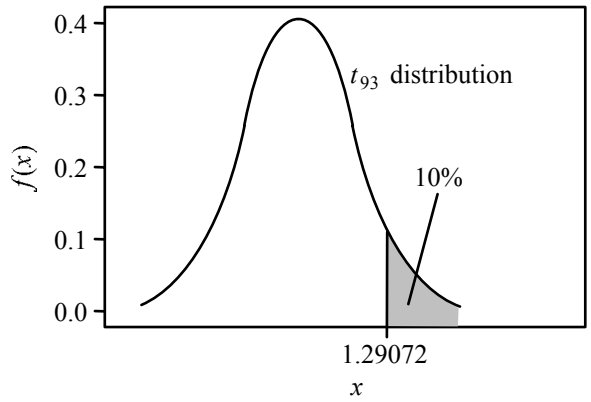
FIGURE A.7



TINV

TINV is used to calculate percentiles of a t distribution. Suppose we want to calculate the 90th percentile of a t distribution with 93 df. This is obtained by specifying TINV(0.2, 93) = 1.291 (see Table A.4 and Figure A.8). In general, for the 100% × (1-α)th percentile of a t distribution with d degrees of freedom, we specify TINV $(2\alpha, d)$. Note that the TINV function can only be used to calculate upper percentiles (i.e., percentiles greater than 50%).

FIGURE A.8



Chi-square distribution commands

CHIDIST and CHIINV are the 2 commands for use with the chi-square distribution.

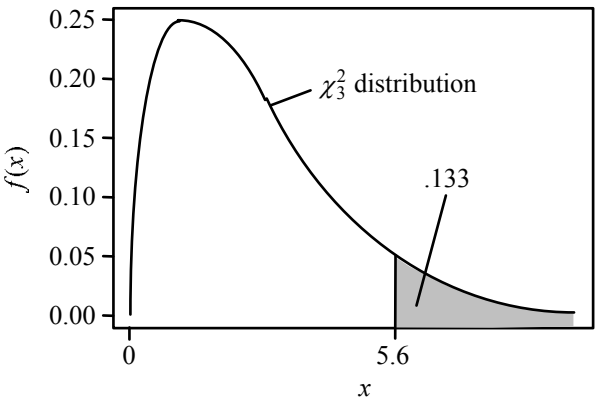
CHIDIST

CHIDIST is used to calculate tail areas for a chi-square distribution. Suppose we want to calculate the probability that a chi-square distribution with 3 d.f. exceeds 5.6. We can obtain this probability by computing CHIDIST(5.6, 3) = .133. In general, $Pr(\chi^2_d > X^2) = \text{CHIDIST}(X^2, d)$. (See Table A.5 and Figure A.9).

Table A.5 Use of chi-square distribution commands in Excel

CHIDIST	CHIDIST(5.6, 3)	0.132778
CHIINV	CHIINV(.1, 31)	41.42175

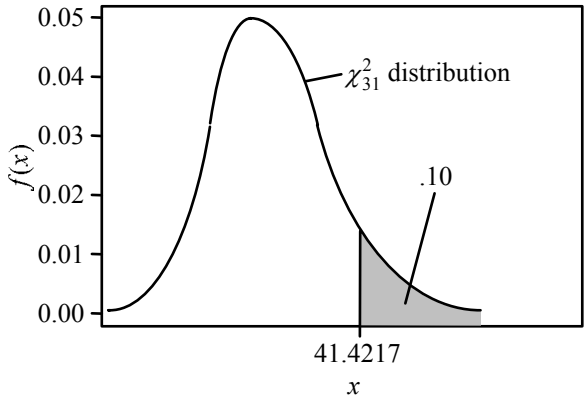
FIGURE A.9



CHIINV

CHIINV is used to calculate percentile for a chi-square distribution. Suppose we want to calculate the upper 10% point of a chi-square distribution with 31 d.f. We can obtain this percentile by computing CHIINV (.1, 31) = 41.4217. In general $\chi^2_{d,\,p} = \text{CHIINV}(p,\, d)$. (See Table A.5 and Figure A.10). To obtain the lower 10th percentile of a chi-square distribution with 31 d.f. we specify CHIINV (.9, 31) = 21.4336.

FIGURE A.10



CHAPTER 8 Hypothesis Testing-2 Sample Inference

In this chapter, we will discuss Excel's commands to perform *t* tests as well as Excel commands to work with the *F* distribution.

TTEST

With the TTEST command; we can obtain *p*-values from both the paired *t* test as well as the 2-sample *t* test with either equal or unequal variances.

Paired *t* test

We refer to data illustrating the effect of using a treadmill on heart rate. Suppose a sedentary individual begins an exercise program and uses a treadmill for 10 minutes (with a 1 minute warm-up and 9 minutes at 2.5 miles per hour). Heart rate is taken before starting the treadmill as well as after 5 minutes while using the treadmill. Measurements are made on 10 days. The baseline heart rate is stored in cells C5:C14 (see Table A.6). The 5 minute heart rate is stored in cells D5:D14. The mean and sd for baseline and 5 min heart rate were also computed using Excel's AVERAGE and STDEV commands and are given in the spreadsheet. We want to test the hypothesis that there has been a significant change in heart rate after using the treadmill for 5 minutes. The appropriate test is the paired *t* test. To implement this test, we use TTEST(C5:C14, D5:D14, 2, 1), where 2 indicates a 2-tailed test and 1 indicates that the paired *t* test is used. The *p*-value from using this test is stored in cell C20. The *p*-value is 4.76×10^{-6} indicating that there has been a significant increase in heart rate after using the treadmill.

Two-sample *t* test

After using the treadmill, the subject either uses the reclining bike for 18 minutes or the Stairmaster for 10 minutes, approximately on alternate days. The heart rate immediately after these activities are given in columns D10:D13 for the bike ($n = 4$) and E10:E15 for the Stairmaster ($n = 6$). (See Table A.7). We wish to compare the mean final heart rate after these 2 activities. Since these are independent samples we will use a 2 sample *t* test. We should perform the *F* test to decide which *t* test to use. However, to illustrate the TTEST command we will perform both tests. For the equal variance *t* test we use TTEST(D10:D13, E10:E15, 2, 2) with a 2-tailed *p*-value indicated by the 3rd argument (2) and the equal variance *t*-test by the 4th argument (2) (see Table A.7). For the unequal variance *t*-test we use TTEST(D10:D13, E10:E15, 2, 3) with a 2-tailed *p*-value indicated by the 3rd argument (2) and the unequal variance *t*-test by the 4th argument (3) (see Table A.7). The results indicate a significant difference in heart rate for both the equal variance *t*-test ($p = .016$) and the unequal variance *t*-test ($p = .013$). Please note that the unequal variance *t* test implementation by Excel is different from the approach used in this text. I haven't been able to find documentation of the formula used.

Table A.6 Example ol Excel, Paired *t* test function

Row		Column C Baseline Heart Rate	Column D 5 min Heart Rate
5		84	87
6		87	92
7		90	93
8		94	98
9		98	100
10		86	92
11		88	93
12		84	90
13		86	92
14		98	104
15			
16	mean	89.5	94.1
17	sd	5.36	5.07
18	*n*	10	10
19			
20	paired *t*, *p*-value = TTEST(C5:C14, D5:D14, 2, 1)	4.76E-06	
21			

Table A.7 Example of Excel two-sample *t* test and *f* test commands

Row			Column D Heart Rate, 18 Minutes of Reclining Bike	Column E Heart Rate, 10 min of Stairmaster
10			93	106
11			102	111
12			98	109
13			97	102
14				99
15				110
16				
17				
18		mean	97.50	106.17
19		sd	3.70	4.79
20		*n*	4	6
21				
22				
23	2 sample *t* test	TTEST(D10:D13, E10:E15, 2, 2)	0.016	
24	equal variance, *p*-value			
25	2 sample *t* test	TTEST(D10:D13, E10:E15, 2, 3)	0.013	
26	unequal variance, *p*-value			
27		*F* statistic	1.680	
28		FDIST(D27,5,3)	0.355	
29	*F* test for the equality	2 × FDIST(D27,5,3)	0.710	
30	of 2 variances, *p*-value			
31		FTEST(E10:E15, D10:D13)	0.710	

F distribution commands

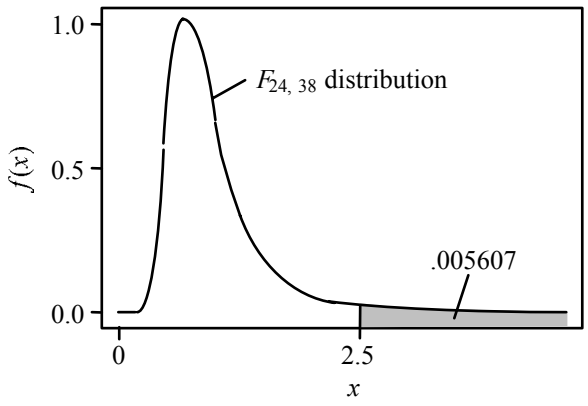
FDIST

The FDIST command is used to calculate tail areas for an F distribution. Specifically, FDIST(x, a, b) calculates $Pr(F_{a,\,b} > x)$. For example, to compute $Pr(F_{24,\,38} > 2.5)$ we specify FDIST(2.5, 24, 38) = .005607 (see Table A.8 and Figure A.11).

Table A.8 Illustration of Excel *F* distributon commands

FDIST(2.5, 24, 38)	0.005607
FINV(.025, 24, 38)	2.026965

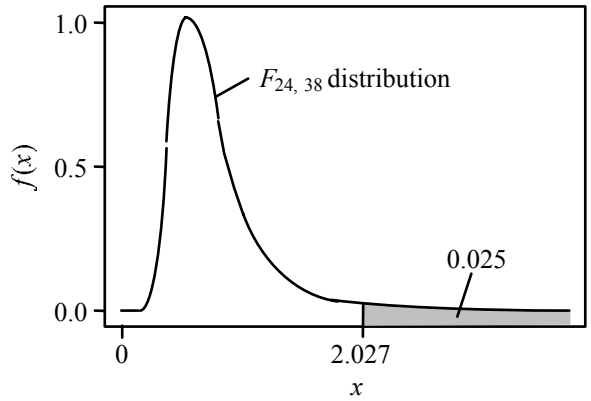
FIGURE A.11



FINV

The FINV command is used to calculate percentiles for an F distribution. Specifically, FINV(p, a, b) calculates the upper pth percentile of an F distribution with a and b df, i.e., $F_{a,\,b,\,1-p}$. For example, to compute the upper 2.5th percentile of an F distribution with 24 and 38 df, we specify FINV(.025, 24, 38) = 2.026965 (see Table A.8 and Figure A.12).

FIGURE A.12



FTEST

This function performs the F test for the equality of 2 variances. For example, referring to Table A.7, to test the hypothesis that the variance of heart rate is the same after using the reclining bike and the stairmaster, we obtain the p-value by specifying FTEST(E10:E15, D10:D13) or alternatively, FTEST(D10:D13, E10:E15) = .710. If we want to see the F statistic, then we alternatively can compute it using ordinary spreadsheet commands (where the larger variance has been placed in the numerator) and obtain 1.680 which is located in cell D27. We then can obtain the p-value for the F test by specifying 2 × FDIST(D27, 5, 3) = .710. To obtain the p-value using this method, the larger variance must be placed in the numerator.

CHAPTER 10 Hypothesis Testing: Categorical Data

HYPGEOMDIST

The HYPGEOMDIST function in Excel is used to compute probabilities under a hypergeometric distribution. This is especially useful in performing Fisher's exact test.

Suppose we have a 2×2 table with cell counts a, b, c and d as shown below

a	b	$a + b$
c	d	$c + d$
$a + c$	$b + d$	n

To compute the exact probability of this table (see text, Equation 10.7, Chapter 10), we specify HYPGEOMDIST(x_1, n_1, x, n), where x_1 = the (1, 1) cell count in the table $= a$, n_1 = 1st row total $= a + b$, x = the 1st column total in the table $= a + c$ and $n =$ the grand total $= a + b + c + d$. The result is the probability of obtaining a table with cells a, b, c and d given the fixed margins. It can also be interpreted as the probability of obtaining x_1 successes in a sample of n_1 trials where the total population is finite and consists of n total sampling units, of which x are successes.

Suppose we consider Problem 10.5 (Study Guide, Chapter 10). Find the probability of obtaining the observed table, viz.

2	13	15
4	15	19
6	28	34

This is obtained from HYPGEOMDIST(2, 15, 6, 34) = .303 (see Table A.9).

Table A.9 Example of HYPGEOMDIST function of Excel

HYPGEOMDIST(2, 15, 6, 34)	0.302609

By repeated use of the HYPGEOMDIST function we can calculate *p*-values for Fisher's exact test. Referring to Problem 10.6, we first enumerate each possible table with the fixed margins (see solution to Problem 10.6). We then use the HYPGEOMDIST function to evaluate the probability of each table (see Table A.10). The 2-tailed *p*-value

$$\begin{aligned} &2 \times \min[Pr(0) + Pr(1) + Pr(2),\ Pr(2) + Pr(3) + \ldots + Pr(6),\ .5] \\ &= 2 \times \min(.020 + .130 + .303,\ .303 + .328 + \ldots + .004,\ .5) \\ &= 2 \times .452 = .905 \end{aligned}$$

CHITEST

The CHITEST function is used to compute the chi-square goodness-of-fit test statistic given in Equation 10.26 (text, Chapter 10). For example, suppose we have observed and expected cell counts stored in B8:B12 and C8:C12, respectively (see Table A.11). If we specify CHITEST(B8:B12, C8:C12), then we obtain the *p*-value = $Pr(\chi_4^2 > X^2)$). In this case the *p*-value = .798 (see Table A.11).

Table A.10 Example of use of HYPGEOMDIST function of Excel to perform Fisher's exact test

HYPGEOMDIST(0, 15, 6, 34)	0.020174
HYPGEOMDIST(l, 15, 6, 34)	0.12969
HYPGEOMDIST(2, 15, 6, 34)	0.302609
HYPGEOMDIST(3, 15, 6, 34)	0.327826
HYPGEOMDIST(4, 15, 6, 34)	0.173555
HYPGEOMDIST(5, 15, 6, 34)	0.042425
HYPGEOMDIST(6, 15, 6, 34)	0.003721
p-value(2-tail)	0.904945

Table A.11 Example of CHITEST function of Excel

Observed Counts (B8:B12)	Expected Counts (C8:C12)
10	8
15	17
23	21
12	11
11	14
CHITEST(B8:B12,C8:C12)	0.798054061

Note that the degrees of freedom used by CHITEST $= g - 1$ if there are g groups. This degrees of freedom is only valid if there are no parameters estimated from the data (i.e., from an externally specified model). Otherwise, the degrees of freedom has to be reduced to $g - p - 1$, where $g =$ the number of groups and $p =$ the number of parameters estimated from the data. In the latter case, the chi-square statistic (X^2) can be computed from the observed and expected counts using ordinary spreadsheet commands and we can

then compute the *p*-value using the CHIDIST function given by CHIDIST$(X^2,\ g-p-1)$. A similar approach can be used to compute the *p*-value for the chi-square test for 2×2 or $r\times c$ contingency tables. In this case, the *p*-value is given by CHIDIST$(X^2,\ 1)$ for a 2×2 table or CHIDIST$\left[X^2,\ (r-1)(c-1)\right]$ for an $r\times c$ table.

CHAPTER 11 Regression and Correlation Methods

There are a number of Excel functions which are useful for calculating correlation and regression statistics. As an example, we have entered the pulse rate data in Table 11.5 (Study Guide, Chapter 11) in Excel (see Table A.12). The dependent variable (pulse rate) is stored in cells D7:D28. The independent variable (age) is stored in cells C7:C28. The following functions are available:

CORR

This function calculates the correlation coefficient between 2 variables. We have

$$\text{CORR(C7:C28, D7:D28)} = -.682,$$

where the *x*-variable (age) is in C7:C28 and the *y*-variable(pulse rate) is in D7:D28 (or vice-versa).

PEARSON

This is the same as the CORR function.

Table A.12 Example of Excel Regression and Correlation Commands

Row	Age Column C	Pulse rate Column D	Correlation and Regression Statistics	Column G
7	1	103	Correlation	–0.681663 CORR(C7:C28, D7:D28) or PEARSON(C7:C28, D7:D28)
8	0	125	Covariance	–69.15083 COVAR(C7:C28, D7:D28)
9	3	102		
10	3	86	Intercept	96.77431 INTERCEPT(D7:D28, C7:C28)
11	5	88	Slope	–1.734055 SLOPE(D7:D28, C7:C28)
12	5	78		
13	6	77	*R*-square	0.464665 RSQ(D7:D28, C7:C28)
14	6	68		
15	9	90	*t*-value	–4.166505 *r**sqrt(n–2)/sqrt(l –*r*^2)
16	8	75	*p*-value	0.000477 TDIST(abs(t-value),20,2)
17	11	78		
18	11	66	Fisher's *z*	–0.832214 FISHER(G7)
19	12	76	transform	
20	13	82		
21	14	58	Inverse	–0.681663 FISHERINV(G18)
22	14	56	Fisher's *z*	
23	16	72	transform	
24	17	70		
25	19	56	*sd*(*y*.*x*)	12.32734 STEYX(D7:D28, C7:C28)
26	18	64		
27	21	81		
28	21	74		

COVAR

This function calculates the covariance between 2 variables. We have COVAR(C7:C28, D7:D28) = −69.15, where C7:C28 is the *x*-variable and D7:D28 is the *y*-variable (or vice-versa).

INTERCEPT

This function calculates the least squares intercept. For example, INTERCEPT(D7:D28, C7:C28) = 96.8, where D7:D28 is the *y*-variable and C7:C28 is the *x*-variable.

SLOPE

This function calculates the least squares slope. For example, SLOPE(D7:D28, C7:C28) $= -1.73$, where D7:D28 is the *y*-variable and C7:C28 is the *x*-variable.

RSQ

This function calculates the R^2 from a univariate regression of *y* on *x*. For example,

$$\text{RSQ(D7:D28, C7:C28)} = .46,$$

where D7:D28 is the *y*-variable and C7:C28 is the *x*-variable.

p-value for the 1-sample *t* test for correlation

There is not an Excel function to obtain the *p*-value for the 1-sample *t* test for correlation. However, this can be done by using standard spreadsheet commands to calculate the *t* statistic (see *t*-value in Table A.12) and then the TDIST function to calculate the *p*-value (see *p*-value in Table A.12). It is possible to obtain an analysis of variance table for both univariate and multiple regression using Excel, but only by using the Tools menu, Data Analysis sub-menu, which is beyond the scope of this Appendix.

FISHER

This function computes Fisher's *z* statistic $= 0.5\ln[(1+r)/(1-r)]$. For example, FISHER(G7) $= -0.832$, where G7 = the sample correlation coefficient (–.682) computed above.

FISHERINV

This function computes the inverse Fisher's *z* transformation $= (e^{2z}-1)/(e^{2z}+1)$. For example, FISHERINV(G18) = –.682, where G18 = the sample *z* statistic (–0.832). This function can be used to compute confidence limits for ρ (see Equation 11.23, Chapter 11, text).

STEYX

This function calculates the standard deviation of *y* given *x*, i.e. $s_{y.x} = \sqrt{\text{Res MS}}$ from a univariate regression of *y* on *x*. For example, STEYX(D7:D28, C7:C28) = 12.33, where the *y*-variable (pulse) is in D7:D28 and the *x*-variable (age) is in C7:C28 and a univariate regression is run of *y* on *x*. This quantity is useful for obtaining standard errors and confidence limits for regression parameters (see Equation 11.9, Chapter 11, text) as well as for obtaining confidence limits for predictions made from regression lines (Equations 11.11 and 11.12, Chapter 11, text).

One-way ANOVA and two-way ANOVA

Excel can also perform multiple regression, one-way ANOVA and two-way ANOVA analyses, using the Tools menu and Data Analysis sub-menu. This is beyond the scope of this Appendix.