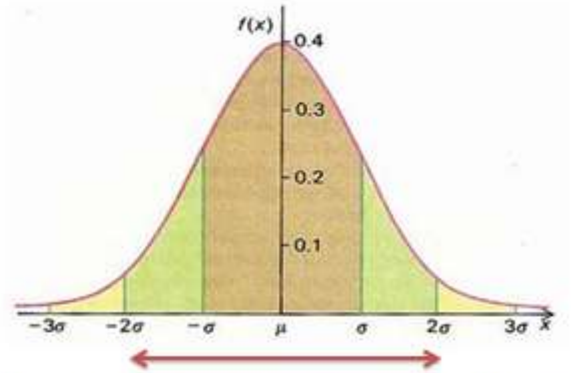
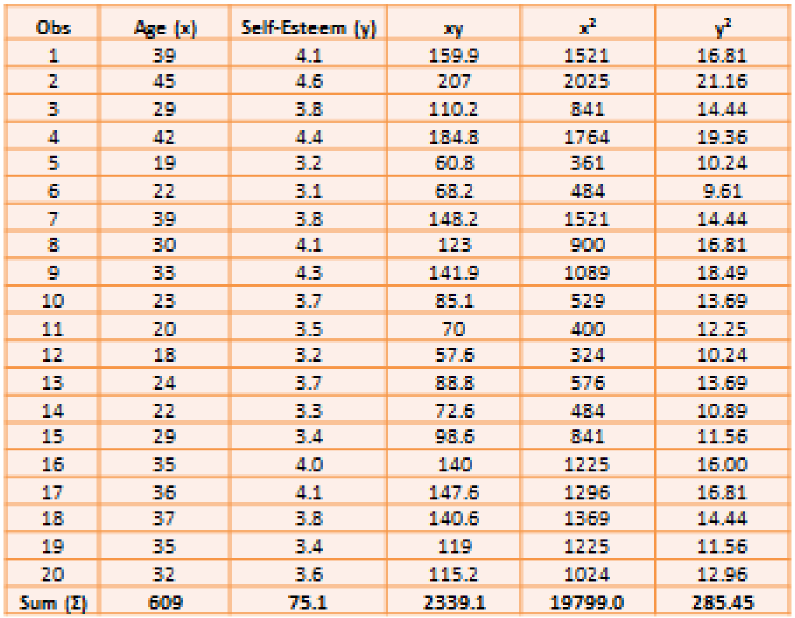


In normal distributions, 95% of the observations fall within two standard deviations of the mean value (μ ± 2σ)

Figure 14.2. Normal distribution



Obs	Age (x)	Self-Esteem (y)	xy	x^2	y^2
1	39	4.1	159.9	1521	16.81
2	45	4.6	207	2025	21.16
3	29	3.8	110.2	841	14.44
4	42	4.4	184.8	1764	19.36
5	19	3.2	60.8	361	10.24
6	22	3.1	68.2	484	9.61
7	39	3.8	148.2	1521	14.44
8	30	4.1	123	900	16.81
9	33	4.3	141.9	1089	18.49
10	23	3.7	85.1	529	13.69
11	20	3.5	70	400	12.25
12	18	3.2	57.6	324	10.24
13	24	3.7	88.8	576	13.69
14	22	3.3	72.6	484	10.89
15	29	3.4	98.6	841	11.56
16	35	4.0	140	1225	16.00
17	36	4.1	147.6	1296	16.81
18	37	3.8	140.6	1369	14.44
19	35	3.4	119	1225	11.56
20	32	3.6	115.2	1024	12.96
Sum (Σ)	609	75.1	2339.1	19799.0	285.45

Table 14.1. Hypothetical data on age and self-esteem

The two variables in this dataset are age (x) and self-esteem (y). Age is a ratio-scale variable, while self-esteem is an average score computed from a multi-item self-esteem scale measured using a 7-point Likert scale, ranging from "strongly disagree" to "strongly agree." The histogram of each variable is shown on the left side of Figure 14.3. The formula for calculating bivariate correlation is:

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

where r_{xy} is the correlation, x and y are the sample means of x and y, and s_x and s_y are the standard deviations of x and y. The manually computed value of correlation between age and self-esteem, using the above formula as shown in Table 14.1, is 0.79. This figure indicates that age has a strong positive correlation with self-esteem, i.e., self-esteem tends to increase with increasing age, and decrease with decreasing age. Such pattern can also be seen from visually comparing the age and self-esteem histograms shown in Figure 14.3, where it appears that the top of the two histograms generally follow each other. Note here that the vertical axes in Figure 14.3 represent actual observation values, and not the frequency of observations (as was in Figure 14.1), and hence, these are not frequency distributions but rather histograms. The bivariate scatter plot in the right panel of Figure 14.3 is essentially a plot of self-esteem on the vertical axis against age on the horizontal axis. This plot roughly resembles an upward sloping line (i.e., positive slope), which is also indicative of a positive correlation. If the two variables were negatively correlated, the scatter plot would slope down (negative slope), implying that an increase in age would be related to a decrease in self-esteem and vice versa. If the two variables were uncorrelated, the scatter plot would approximate a horizontal line (zero slope), implying than an increase in age would have no systematic bearing on self-esteem.

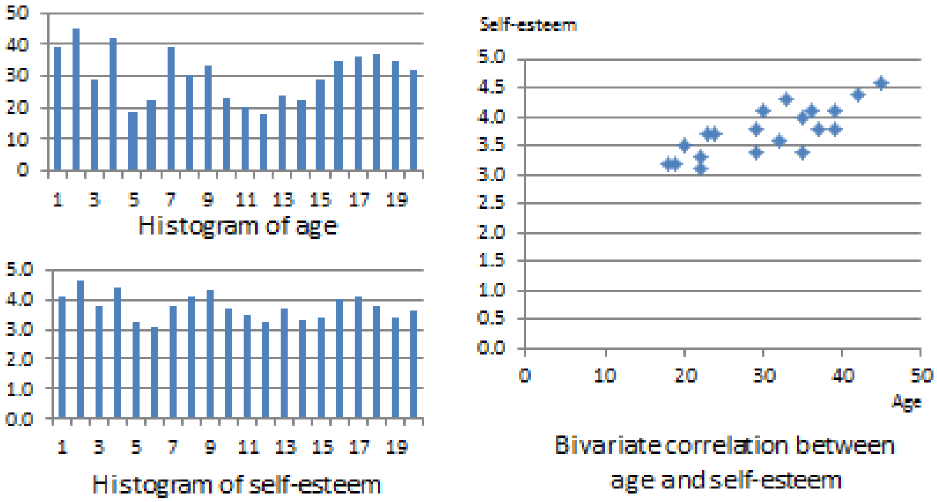


Figure 14.3. Histogram and correlation plot of age and self-esteem

After computing bivariate correlation, researchers are often interested in knowing whether the correlation is significant (i.e., a real one) or caused by mere chance. Answering such a question would require testing the following hypothesis:

$$H_0: r = 0$$
$$H_1: r \neq 0$$

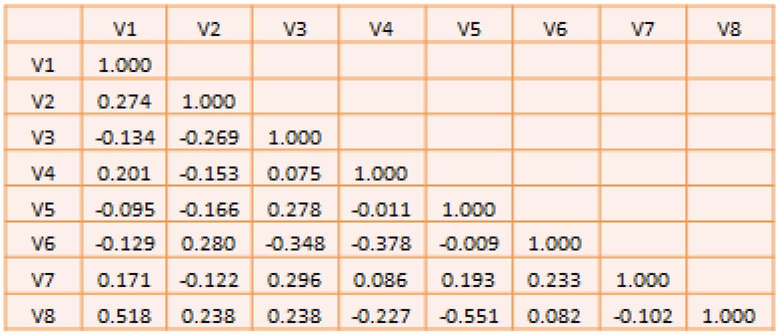
H_0 is called the *null hypotheses*, and H_1 is called the *alternative hypothesis* (sometimes, also represented as H_a). Although they may seem like two hypotheses, H_0 and H_1 actually represent a single hypothesis since they are direct opposites of each other. We are interested in testing H_1 rather than H_0. Also note that H_1 is a non-directional hypotheses since it does not specify whether r is greater than or less than zero. Directional hypotheses will be specified as H_0: $r \leq 0$; H_1: $r > 0$ (if we are testing for a positive correlation). Significance testing of directional hypothesis is done using a one-tailed t-test, while that for non-directional hypothesis is done using a two-tailed t-test.

In statistical testing, the alternative hypothesis cannot be tested directly. Rather, it is tested indirectly by rejecting the null hypotheses with a certain level of probability. Statistical testing is always probabilistic, because we are never sure if our inferences, based on sample data, apply to the population, since our sample never equals the population. The probability that a statistical inference is caused pure chance is called the **p-value**. The p-value is compared with the **significance level** (α), which represents the maximum level of risk that we are willing to take that our inference is incorrect. For most statistical analysis, α is set to 0.05. A p-value less than $\alpha=0.05$ indicates that we have enough statistical evidence to reject the null hypothesis, and thereby, indirectly accept the alternative hypothesis. If $p>0.05$, then we do not have adequate statistical evidence to reject the null hypothesis or accept the alternative hypothesis.

The easiest way to test for the above hypothesis is to look up critical values of r from statistical tables available in any standard text book on statistics or on the Internet (most software programs also perform significance testing). The critical value of r depends on our desired significance level ($\alpha = 0.05$), the degrees of freedom (df), and whether the desired test is a one-tailed or two-tailed test. The **degree of freedom** is the number of values that can vary freely in any calculation of a statistic. In case of correlation, the df simply equals $n - 2$, or for the data in Table 14.1, df is $20 - 2 = 18$. There are two different statistical tables for one-tailed and two-tailed test. In the two-tailed table, the critical value of r for $\alpha = 0.05$ and $df = 18$ is 0.44. For our computed correlation of 0.79 to be significant, it must be larger than the critical value of 0.44 or less than -0.44. Since our computed value of 0.79 is greater than 0.44, we conclude that there is a significant correlation between age and self-esteem in our data set, or in other words, the odds are less than 5% that this correlation is a chance occurrence. Therefore, we can reject the null hypotheses that $r \leq 0$, which is an indirect way of saying that the alternative hypothesis $r > 0$ is probably correct.

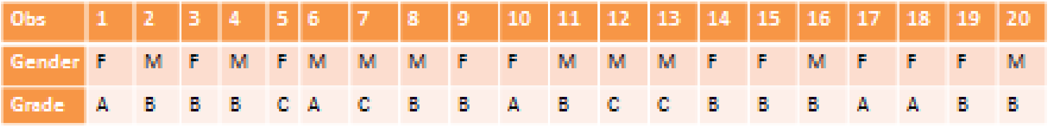
Most research studies involve more than two variables. If there are n variables, then we will have a total of $n*(n-1)/2$ possible correlations between these n variables. Such correlations are easily computed using a software program like SPSS, rather than manually using the formula for correlation (as we did in Table 14.1), and represented using a correlation matrix, as shown in Table 14.2. A correlation matrix is a matrix that lists the variable names along the first row and the first column, and depicts bivariate correlations between pairs of variables in the appropriate cell in the matrix. The values along the principal diagonal (from the top left to the bottom right corner) of this matrix are always 1, because any variable is always perfectly correlated with itself. Further, since correlations are non-directional, the correlation between variables V1 and V2 is the same as that between V2 and V1. Hence, the lower triangular matrix (values below the principal diagonal) is a mirror reflection of the upper triangular matrix (values above the principal diagonal), and therefore, we often list only the lower triangular matrix for simplicity. If the correlations involve variables measured using interval scales, then this specific type of correlations are called **Pearson product moment correlations**.

Another useful way of presenting bivariate data is cross-tabulation (often abbreviated to cross-tab, and sometimes called more formally as a contingency table). A **cross-tab** is a table that describes the frequency (or percentage) of all combinations of two or more nominal or categorical variables. As an example, let us assume that we have the following observations of gender and grade for a sample of 20 students, as shown in Figure 14.3. Gender is a nominal variable (male/female or M/F), and grade is a categorical variable with three levels (A, B, and C). A simple cross-tabulation of the data may display the joint distribution of gender and grades (i.e., how many students of each gender are in each grade category, as a raw frequency count or as a percentage) in a 2 x 3 matrix. This matrix will help us see if A, B, and C grades are equally distributed across male and female students. The cross-tab data in Table 14.3 shows that the distribution of A grades is biased heavily toward female students: in a sample of 10 male and 10 female students, five female students received the A grade compared to only one male students. In contrast, the distribution of C grades is biased toward male students: three male students received a C grade, compared to only one female student. However, the distribution of B grades was somewhat uniform, with six male students and five female students. The last row and the last column of this table are called marginal totals because they indicate the totals across each category and displayed along the margins of the table.



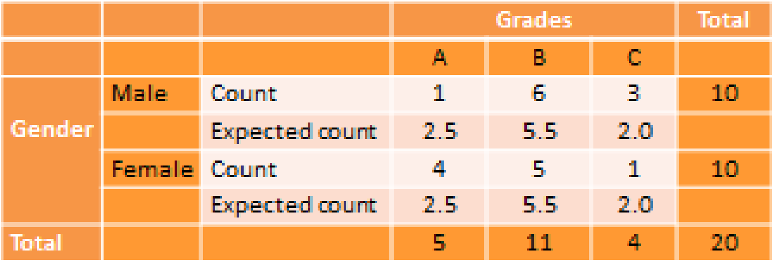
	V1	V2	V3	V4	V5	V6	V7	V8
V1	1.000							
V2	0.274	1.000						
V3	-0.134	-0.269	1.000					
V4	0.201	-0.153	0.075	1.000				
V5	-0.095	-0.166	0.278	-0.011	1.000			
V6	-0.129	0.280	-0.348	-0.378	-0.009	1.000		
V7	0.171	-0.122	0.296	0.086	0.193	0.233	1.000	
V8	0.518	0.238	0.238	-0.227	-0.551	0.082	-0.102	1.000

Table 14.2. A hypothetical correlation matrix for eight variables



Obs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Gender	F	M	F	M	F	M	M	M	F	F	M	M	M	F	F	M	F	F	F	M
Grade	A	B	B	B	C	A	C	B	B	A	B	C	C	B	B	B	A	A	B	B

Hypothetical data set



			Grades			Total
			A	B	C	
Gender	Male	Count	1	6	3	10
		Expected count	2.5	5.5	2.0	
	Female	Count	4	5	1	10
		Expected count	2.5	5.5	2.0	
Total			5	11	4	20

Cross-tabulation of gender versus age

Table 14.3. Example of cross-tab analysis

Although we can see a distinct pattern of grade distribution between male and female students in Table 14.3, is this pattern real or "statistically significant"? In other words, do the above frequency counts differ from that that may be expected from pure chance? To answer this question, we should compute the expected count of observation in each cell of the 2 x 3 cross-tab matrix. This is done by multiplying the marginal column total and the marginal row total for each cell and dividing it by the total number of observations. For example, for the male/A grade cell, expected count = 5 * 10 / 20 = 2.5. In other words, we were expecting 2.5 male students to receive an A grade, but in reality, only one student received the A grade. Whether this difference between expected and actual count is significant can be tested using a *chi-square test*. The chi-square statistic can be computed as the average difference between observed and expected counts across all cells. We can then compare this number to the critical value associated with a desired probability level ($p < 0.05$) and the degrees of freedom, which is simply (m-1)*(n-1), where m and n are the number of rows and columns respectively. In this example, df = (2 – 1) * (3 – 1) = 2. From standard chi-square tables in any statistics book, the critical chi-square value for p=0.05 and df=2 is 5.99. The computed chi-square value, based on our observed data, is 1.00, which is less than the critical value. Hence, we must conclude that the observed grade pattern is not statistically different from the pattern that can be expected by pure chance.