1.3 | Exploratory Data Analysis in 2-D

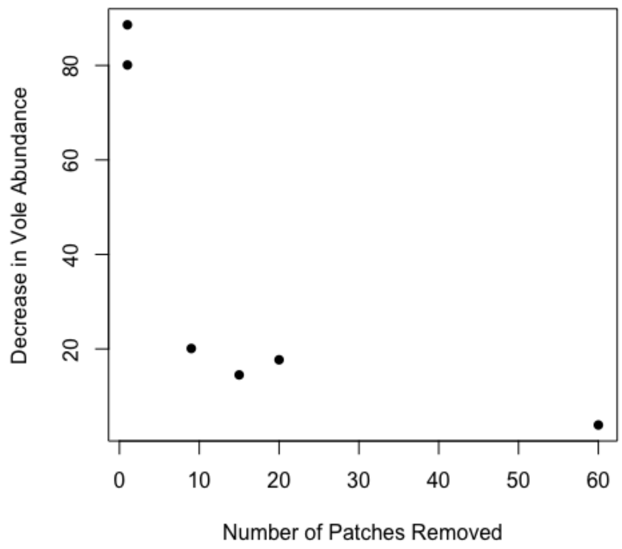
After exploring the central tendency, dispersion, and shape of each dimension separately, the next step is to investigate the relationship between each pair of dimensions or variables. These two-dimensional or *bivariate* analyses typically focus on describing the joint dispersion or covariability of the two variables. If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the lesser values, the covariability is positive. In the opposite case, when the greater values of one variable mainly correspond to the lesser values of the other, the covariability is negative.

1.3.1 | Visualizing Covariability in 2-D

We can visualize the covariability between two quantitative variables using a two-dimensional extension of a dot plot known as a *scatter plot*. We create a scatter plot by displaying the cases as a collection of dots, each having its position on the horizontal axis determined by its value on one dimension and its position on the vertical axis by its value on the other dimension.

In R, we can create a scatter plot by using the plot() command while adjusting a number of arguments to make it look more aesthetically pleasing.

```
plot(MG[,2], MG[,3], pch = 16, xlab = "Number of Patches Removed",
     ylab = "Decrease in Vole Abundance")
```



There are numerous improvements that we can make to this plot, all of which we can explore using the **Help File** that we can access using ?plot.

Even without any additional improvements, however, we can use this scatter plot to describe the *direction*, *shape*, and *strength* of the covariability.

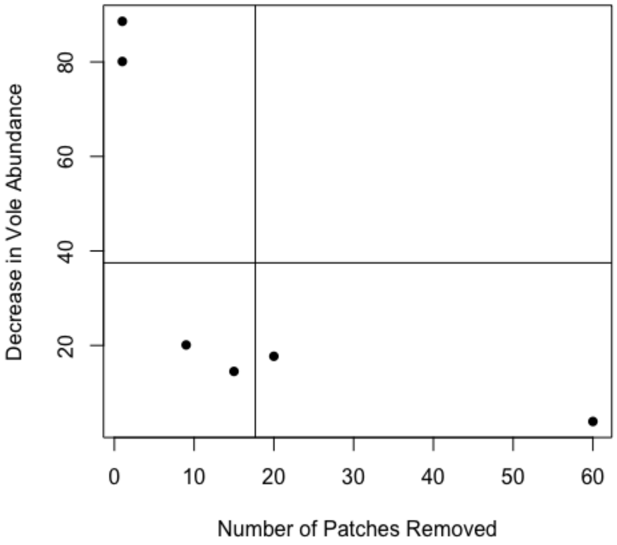
The direction is positive if the cloud of points in the scatter plot goes up from left to right, indicating that higher values on one dimension are accompanied by higher values on the other dimension. If the points instead go down from left to right, as we see in our example data, the direction is negative. If the cloud of points does not seem to follow any clear pattern, or if the pattern goes from left to right without increasing or decreasing, the direction is said to be zero.

The shape of the covariability can take on many forms, including linear, exponential, and quadratic, among others. Given the importance of linear covariability in many statistical models and analytical techniques, we commonly separate the shape into two broad categories – linear if the cloud of points follows a straight line and non-linear if they do not.

The strength of the covariability is an indicator of how closely the cloud of points align with the line or curve that they follow. The closer the data points are to this shape, the stronger the covariability. Although there are some guidelines for what aspects of a scatter plot indicate a strong, moderate, or weak covariability, these distinctions are much easier to identify numerically, as we will do in Section 1.3.2 below.

In addition to exploring the direction, shape, and strength of the covariability, we can use the scatter plot to visualize the two-dimensional *centroid* of the data, which is simply an extension of the one-dimensional mean used earlier to summarize the central tendency of a distribution. Using cartesian coordinates, the centroid can be written as $(\bar{x}_1, \bar{x}_2)$. We can then use the abline() function to divide the scatter plot into four quadrants about this centroid.

```
plot(MG[,2], MG[,3], pch = 16, xlab = "Number of Patches Removed",
     ylab = "Decrease in Vole Abundance")
abline(v = mean(MG[,2]), h = mean(MG[,3]))
```



The `v` and `h` arguments of the `abline()` function are used to denote where to draw the vertical and horizontal lines, respectively.

Now that we know which data points fall into which quadrants, we can use an analytical technique known as *scatter plot quadrant analysis* to quickly determine the direction of the covariability. Data points in the bottom left and top right quadrant, where being above average on one dimension is accompanied by being above average on the other dimension (or being below average on one is accompanied by below-average values on the other), contribute to a positive covariability. Data points in the top left and bottom right quadrants, on the other hand, where being above average on one dimension is accompanied by being below average on the other dimension, contribute to a negative covariability. Data points that fall exactly on either the vertical or horizontal line contribute no information to the direction of the covariability.

Although it is not always so simple, a quick check to determine direction is to determine if there are more points indicative of a positive covariability or more points indicative of a negative covariability. In our example, data there are four data points in the negative quadrants (two sets of points have identical coordinates and as such are plotted one on top of the other) and only two in the negative quadrants, confirming our initial conclusion that the covariability is indeed negative.

1.3.2 | Quantifying Covariability in 2-D

In addition to visually inspecting and qualitatively discussing the covariability between two dimensions, we can quantify this covariability by simply expanding our earlier calculations of squared deviations and variances into a two-dimensional space.

First, we consider the two-dimensional deviation of each value, which we compute as the product of the deviation of one variable around its means times the deviation of the other variable around its mean:

$$(x_{i,1} - \bar{x}_1)(x_{i,2} - \bar{x}_2) \tag{1.9}$$

The sum of these values across all cases is known as the *cross-product* or *sum of products*:

$$SP = \sum_{i=1}^{n} (x_{i,1} - \bar{x}_1)(x_{i,2} - \bar{x}_2) \tag{1.10}$$

We can use R to quickly calculate this value:

```
SP = sum((MG[,2] - mean(MG[,2])) * (MG[,3] - mean(MG[,3])))
SP
[1] -2818.133
```

Note that if we compute the sum of products of a variable with itself (i.e., $x_1 = x_2$), the calculation simplifies to the sum of squares defined in equation (1.4) earlier:

$$SP = \sum_{i=1}^{n} \left(x_{i,1} - \bar{x}_1\right)\left(x_{i,1} - \bar{x}_1\right) = \sum_{i=1}^{n} (x_i - \bar{x})^2 = SS \qquad (1.11)$$

Note also that these calculations follow the same pattern as the scatter plot quadrant analysis above. Data points that are above average or below average on both dimensions yield positive contributions to the sum of products and would appear in either the bottom left or top right quadrants. Points that are above average on one dimension but below average on the other dimension yield negative contributions and would appear in either the top left or bottom right quadrants. The distance that these points are from each of the two averages then determines the magnitude of the covariability. Points that are at the exact average on at least one of the dimensions yield a value of zero, and as such do not contribute to the calculated covariability.

By dividing the sum of products by $n - 1$, which in effect averages the two-dimensional deviations, we arrive at a common measure of covariability known as the *covariance*:

$$Cov(x_1, x_2) = \frac{SP}{n-1} = \frac{\sum_{i=1}^{n}\left(x_{i,1} - \bar{x}_1\right)\left(x_{i,2} - \bar{x}_2\right)}{n-1} \qquad (1.12)$$

In R, we can use the `cov()` function or calculate this value manually:

```
COV <- SP / (nrow(MG) - 1)
COV
[1] -563.6267

cov(MG[,2],MG[,3])
[1] -563.6267
```

Note that the covariance of a variable with itself simplifies to the variance in equation (1.5).

The units of measurement for the covariance $Cov(x_1, x_2)$ are those of x_1 times those of x_2, which makes comparing the covariability between two different pairs of variables challenging. How do you compare a covariance of 1.3 feet-dollars to 3.7 degree-pounds?

The *correlation* between two variables, on the other hand, provides a *standardized* and *unitless* measure of covariability that can easily be used to make such comparisons. Although there are several correlation coefficients that are used to measure the degree of covariability, the most common by far is the *Pearson's correlation coefficient*, which is often denoted as r:

$$Cor(x_1, x_2) = \frac{Cov(x_1, x_2)}{S_1 S_2} \qquad (1.13)$$

We can again use R to calculate this value, either manually or by using the `cor()` function.

```
COR <- COV / (sd(MG[,2]) * sd(MG[,3]))
COR
[1] -0.6935986

cor(MG[,2],MG[,3])
[1] -0.6935986
```

It is a corollary of the *Cauchy-Schwarz inequality* that the absolute value of the correlation coefficient cannot exceed 1; therefore, the correlation always ranges between -1 and +1. The sign of the coefficient again indicates the direction of the covariability, while the magnitude indicates its strength. The closer the coefficient is to either -1 or +1, the stronger the relationship, whereas values approaching zero suggest a lack of covariability.

One important thing to note is that this correlation coefficient is only appropriate for data whose covariability exhibits a linear shape. If the shape is substantially non-linear, the Pearson's correlation coefficient will often produce values that do not accurately capture the covariability (e.g., a correlation of zero when there is a clear but not non-linear covariability). In such instances, we can either use a different correlation coefficient or simply transform our data until the covariability is more linear.

1.3.3 | Using Matrices to Summarize Covariability

Given the similarity between the sum of squares and the sum of products, we often combine the two measures into the *sum of squares and products matrix*, $\boldsymbol{S}$ (note the bold), which is also known as the *sum of squares and cross product matrix*:

$$\boldsymbol{S} = \begin{array}{cc} & \begin{array}{cc} x_1 & x_2 \end{array} \\ \begin{array}{c} x_1 \\ x_2 \end{array} & \begin{bmatrix} SS(x_1) & SP(x_1,x_2) \\ SP(x_1,x_2) & SS(x_2) \end{bmatrix} \end{array} \tag{1.14}$$

Each row and column represent a different dimension (or variable), and the numerical quantity at the intersection of each row and column is the sum of products or the sum of squares between the two variables.

By dividing the sum of products by $n - 1$, which again averages both one-dimensional squared deviations and the two-dimensional unsquared deviations, we can arrive at the *variance-covariance matrix*, $\boldsymbol{\Sigma}$, commonly referred to as just the *covariance matrix*:

$$\boldsymbol{\Sigma} = \frac{\boldsymbol{S}}{n-1} = \begin{array}{cc} & \begin{array}{cc} x_1 & x_2 \end{array} \\ \begin{array}{c} x_1 \\ x_2 \end{array} & \begin{bmatrix} Var(x_1) & Cov(x_1,x_2) \\ Cov(x_1,x_2) & Var(x_2) \end{bmatrix} \end{array} \tag{1.15}$$

We can use R to quickly calculate this matrix, but only if the two dimensions are referenced as a single concatenated object:

```
SIGMA <- cov(MG[,2:3])
SIGMA
           Patches  Decrease
Patches   487.0667 -563.6267
Decrease -563.6267 1355.7457
```

Finally, with a little bit of matrix algebra, we can standardize the covariance matrix to produce the *correlation matrix*, $\boldsymbol{R}$:

$$\boldsymbol{R} = \left(\left(\sqrt{diag(\boldsymbol{\Sigma})}\right)^{-1}\right) \times \boldsymbol{\Sigma} \times \left(\left(\sqrt{diag(\boldsymbol{\Sigma})}\right)^{-1}\right)^{T} \qquad (1.16)$$

Matrix algebra such as these operations are very easy to implement in R, or we can always use built-in functions like `cor()` to achieve the same result:

```
R <-   solve(sqrt(diag(diag(SIGMA)))) %*% SIGMA %*%
     t(solve(sqrt(diag(diag(SIGMA)))))
R
           [,1]       [,2]
[1,]  1.0000000 -0.6935986
[2,] -0.6935986  1.0000000

cor(MG[,2:3])
            Patches   Decrease
Patches   1.0000000 -0.6935986
Decrease -0.6935986  1.0000000
```

If you're unsure how these operations yield the result above, either mathematically or in R, a review of matrix algebra is included in Section 1.3.4 below.

As expected, the correlation of a variable with itself, which is denoted on the *main diagonal* of the correlation matrix, is always 1. Note also that the correlation matrix, as well as the covariance and sum of squares and products matrices, are symmetric about the main diagonal, a characteristic that we will make use of in future analyses.

1.3.4 | A Review of Matrix Algebra

The most common matrices in data analysis are square matrices that have the same number of rows as columns. Although sum of squares and products, variance-covariance, and correlation matrices are all square and symmetric, the calculations illustrated in this review will instead use 3x3 matrices that are not symmetric to better illustrate the underlying tenets of matrix algebra:

$$\boldsymbol{A} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = \begin{bmatrix} 8 & 1 & 6 \\ 3 & 5 & 7 \\ 4 & 9 & 2 \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} r & s & t \\ u & v & w \\ x & y & z \end{bmatrix} = \begin{bmatrix} 16 & 9 & 14 \\ 11 & 13 & 15 \\ 12 & 17 & 10 \end{bmatrix} \tag{1.17}$$

As we've done before, we can construct these matrices in R using the `matrix()` function.

```
A <- matrix(c(8,3,4,1,5,9,6,7,2), nrow = 3, ncol = 3)
A
     [,1] [,2] [,3]
[1,]    8    1    6
[2,]    3    5    7
[3,]    4    9    2

B <- matrix(c(16,11,12,9,13,17,14,15,10), nrow = 3, ncol = 3)
B
     [,1] [,2] [,3]
[1,]   16    9   14
[2,]   11   13   15
[3,]   12   17   10
```

The *main diagonal* of a matrix is the sequence of numbers that runs from the upper left to the lower right. In R, we can use the `diag()` to find this sequence.

```
diag(A)
[1] 8 5 2
```

The *trace* of a matrix is the sum of all the numbers in this main diagonal. In R, we can find this quantity by simply summing the values produced by the `diag()` function.

```
sum(diag(A))
[1] 15
```

For a sum of squares and products matrix, the trace is the total sum of squares for the data. For a correlation matrix, the trace is simply the dimensionality of the data.

The *diagonalization* of a matrix is achieved by setting all elements outside the main diagonal to zero, which we can achieve by using the `diag()` function in R twice.

```
diag(diag(A))
     [,1] [,2] [,3]
[1,]    8    0    0
[2,]    0    5    0
[3,]    0    0    2
```

The *transpose* of a matrix is a reflection of the values about the main diagonal, such that the first row becomes the first column, the second row becomes the second column, and so on:

$$\boldsymbol{A}^T = \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & i \end{bmatrix} = \begin{bmatrix} 8 & 3 & 4 \\ 1 & 5 & 9 \\ 6 & 7 & 2 \end{bmatrix} \qquad (1.18)$$

In R, this is achieved using the `t()` function.

```
t(A)
     [,1] [,2] [,3]
[1,]    8    3    4
[2,]    1    5    9
[3,]    6    7    2
```

The *inverse* useful property is the *inverse* of a matrix, but its calculation is a bit beyond the scope of this text. Instead, we will rely on the `solve()` function in R to find this quantity.

```
solve(A)
            [,1]        [,2]        [,3]
[1,]  0.14722222 -0.14444444  0.06388889
[2,] -0.06111111  0.02222222  0.10555556
[3,] -0.01944444  0.18888889 -0.10277778
```

For those who are interested, the inverse is found by solving the following equation, where $\boldsymbol{I}$ is an *identity matrix* that has ones along the main diagonal, zeroes everywhere else, and the same dimensions as matrix $\boldsymbol{A}$:

$$\boldsymbol{A} \times \boldsymbol{A}^{-1} = \boldsymbol{I} \qquad (1.19)$$

To multiply two matrices together, we multiply the elements in the first row of $\boldsymbol{A}$ by those in the first column of $\boldsymbol{B}$, add the resulting values together, and use this as the new value in the first row, first column. Similarly, we multiply the elements in the first row of $\boldsymbol{A}$ by those in the second column of $\boldsymbol{B}$ and add them together to find the value in the first row, second column:

$$\boldsymbol{A} \times \boldsymbol{B} = \begin{bmatrix} ar + bu + cx & as + bv + cy & at + bw + cz \\ rd + eu + fx & ds + ev + fy & dt + ew + fz \\ gr + hu + ix & gs + hv + iy & gt + hw + iz \end{bmatrix} \qquad (1.20)$$

In R, we can multiply two matrices together using the `%*%` operator.

```
A %*% B
     [,1] [,2] [,3]
[1,]  211  187  187
[2,]  187  211  187
[3,]  187  187  211
```

We will make extensive use of these matrix operations in later analyses, so keep them in mind.

Exercises for Section 1.3

1.09 Kaprekar’s Constant. In number theory, Kaprekar's routine is an iterative algorithm that, with each iteration, takes a natural number, creates two new numbers by sorting the digits of its number by descending and ascending order, and subtracts the second from the first to yield the natural number for the next iteration. For example, starting with 1234:

$$\begin{aligned}
&\text{Step 1:} \quad 4321 - 1234 = 3087\\
&\text{Step 2:} \quad 8730 - 0378 = 8352\\
&\text{Step 3:} \quad 8532 - 2358 = 6174\\
&\text{Step 4:} \quad 7641 - 1467 = 6174
\end{aligned}$$

It is named after its inventor, the Indian mathematician Dattatreya Ramchandra Kaprekar, who also showed that in the case of 4-digit numbers in base 10, if the initial number has at least two distinct digits, after at most seven iterations, this process always yields the number 6174.

a. Find the mean of the ascending (i.e., 1, 4, 6, and 7) and descending digits (i.e., 7, 6, 4, and 1) of Kaprekar’s Constant by using R as a calculator (without any build-in functions such as `mean()`, `sum()`, `length()`, or others).
b. Find standard deviation of the ascending and descending digits by using R as a calculator (which will be the same because the digits are identical, just reordered)
c. Find the covariance between the ascending and descending digits by using R as a calculator
d. Find the corresponding correlation by using R as a calculator

1.10 Presidential Approval Ratings and Margin of Victory. Exercise 1.04 introduced data on a sitting president’s prospects of winning a second term in office and their job approval rating:

	Year	Approval	Margin
Eisenhower	1956	68	15.4
Johnson	1964	74	22.6
Nixon	1972	59	23.2
Ford	1976	45	-2.1
Carter	1980	37	-0.7
Reagan	1984	58	18.2
BushSr	1992	35	-5.5
Clinton	1996	54	8.5
BushJr	2004	48	2.4
Obama	2012	52	3.9
Trump	2020	45	-4.4

The data above include the margin of victory (or defeat) in the popular vote (which often, but not always, matches the results of the electoral vote).

a. Construct this data matrix in R with the proper row and column names
b. Create a scatter plot of each president’s `Approval` rating and `Margin` of victory
c. Without doing any calculations, indicate whether the covariability appears to be positive or negative and strong or weak

d. Find the correlation between the two dimensions

1.11 Wolves and Moose on Isle Royale (continued). Exercise 1.01 introduced an excerpt of the data collected on the number of wolves and moose on Isle Royale, which is recreated below:

	Wolves	Moose	Kill Rate	Predation Rate	Moose Recruitment Rate
1959	20	538	NA	NA	20
1960	22	564	NA	NA	14.3
1961	22	572	NA	NA	19.5
1962	23	579	NA	NA	16.5
1963	20	596	NA	NA	21.2
1964	26	620	NA	NA	15.9
1965	28	634	NA	NA	13.2
1966	26	661	NA	NA	18
1967	22	766	NA	NA	21
1968	22	848	NA	NA	20
1969	17	1041	NA	NA	16.5
1970	18	1045	NA	NA	16.1
1971	20	1183	0.615	0.062	11.4
1972	23	1243	0.819	0.091	10.7
1973	24	1215	0.760	0.090	15.1
1974	31	1203	0.599	0.093	14.7
1975	41	1139	0.645	0.139	12.3
1976	44	1070	0.563	0.139	10.3
1977	34	949	0.298	0.064	6.1
1978	40	845	0.507	0.144	10.2
1979	43	857	0.387	0.117	13
1980	50	788	0.330	0.126	12
1981	30	767	0.217	0.051	23.7
1982	14	780	0.869	0.094	20.7
1983	23	830	0.394	0.065	16.4
1984	24	927	0.440	0.068	14.5
1985	22	976	0.457	0.062	13.1
1986	20	1014	0.670	0.079	16
1987	16	1046	0.549	0.050	16
1988	12	1116	0.864	0.056	15.2
1989	12	1260	0.810	0.046	13.4
1990	15	1315	0.857	0.059	14.8
1991	12	1496	1.029	0.050	13.9
1992	12	1697	1.428	0.061	10.9
1993	13	1784	0.877	0.038	14
1994	17	2017	0.792	0.040	12
1995	16	2117	1.391	0.063	7
1996	22	2398	1.274	0.070	3

The data above include the number of Wolves and Moose, as well as the number of moose killed per wolf (Kill Rate), the proportion of moose being killed by wolves (Predation Rate), and the proportion of moose that are calves (Moose Recruitment Rate) over a 38-year period.

a. Construct this data matrix in R with the proper row and column names

b. Create a scatter plot of the covariability between the number of Wolves and Moose and separate it into four quadrants using the centroid
c. Use a scatter plot quadrant analysis to determine if the covariability between these two dimensions is likely to be positive, negative, or close to zero
d. Find the covariance and the correlation between these two dimensions
e. Use these summary statistics to describe if the covariability is positive or negative
f. Use these summary statistics to describe if the covariability is strong, weak, or moderate

1.12 Anscombe's Quartet. Anscombe's Quartet comprises four data sets that have nearly identical descriptive statistics yet have very different distributions and appear very different when graphed.

	X1	Y1	X2	Y2	X3	Y3	X4	Y4
1	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
2	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
3	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
4	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
5	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
6	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
7	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
8	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
9	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
10	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
11	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

These sets of values were constructed in 1973 by the English statistician Francis Anscombe to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."

a. Construct this data matrix in R with the proper row and column names
b. Find the covariance between each pair of quartets (i.e., X1 and Y1, X2 and Y2, etc.)
c. Find the correlation between each pair of quartets
d. Describe which pair of quartets, if any, appears to have the strongest covariability (ignoring rounding error)
e. Construct a scatter plot for each pair of quartets
f. Use the answers above to describe why only presenting summary statistics such as the correlation is inadequate to properly describe the covariability within data sets