1.5 | Distance and Outliers

After quantifying the central tendency and dispersion of each dimension separately and of each possible pairwise combination of dimensions, a common final step in exploratory data analysis is to investigate the influence of each data point separately. These analyses typically focus on providing a standardized measure of both the one- and two-dimensional distance of outlying points from the corresponding central tendency. The resulting distances are then compared to critical values derived from theoretical distributions to determine if the data points should be considered as significantly outlying and subsequently checked for influence.

1.5.1 | Quantifying Distance in One Dimension

As we have seen earlier, we can measure the one-dimensional distance of each data point $x_{i,1}$ from the mean $\bar{x}_1$ using the *deviation*:

$$\left(x_{i,1} - \bar{x}_1\right) \tag{1.21}$$

Unfortunately, the units of measurement for this metric are the same as those of the original variable, which makes comparisons between different dimensions challenging. How do you compare a distance of three feet to a distance of seven degrees? Which is more extreme?

The *standard score* or *z-score* provides a standardized and unitless measure of distance from the central tendency that can be easily compared across dimensions and different data sets. By dividing the deviation by the standard deviation s_1 of the corresponding dimension, this metric measures how many standard deviations the data point is from the mean:

$$z = \frac{\left(x_{i,1} - \bar{x}_1\right)}{s_1} \tag{1.22}$$

Note that this *standardization* by dividing some value by the standard deviation is very similar to that used to find the correlation (a standardized measure of covariability) from the covariance (an unstandardized measure of covariability).

We can easily use R to calculate this score for any data point within a dimension.

```
(MG[1,1] - mean(MG[,1])) / sd(MG[,1])
[1] 0.6132387
```

The sign of the score indicates the direction of the distance, with positive scores corresponding to data points that are above average and negative scores to points that are below average.

The magnitude, on the other hand, indicates how far the data point is from the mean, with higher absolute scores denoting more extreme values. The first data point in the first dimension of our example data, for instance, is almost two-thirds of a standard deviation above the mean (of the amount of area removed via timber harvest).

Because of the ease with which standardized scores can be used to make comparisons both within and between dimensions or even across different data sets, it is quite common to standardize entire dimensions or data sets prior to many data analyses.

In R, we can standardize large sets of data by using the `scale()` function.

```
scale(MG[,1])
         [,1]
1  0.6132387
2  0.6132387
3  0.6132387
4 -0.7918519
5  0.6370538
6 -1.6849179
attr(,"scaled:center")
[1] 7.325
attr(,"scaled:scale")
[1] 4.199018
```

Note that in addition to providing standardized scores for each data point in the dimension, this function also outputs the corresponding measures of central tendency and dispersion as the `scaled:center` and `scaled:scale` attributes, respectively.

When the `scale()` function is applied to an entire data frame or matrix, the output is standardized separately for each column, ensuring that each data point is compared to the mean of its specific dimension rather than the mean of the data as a whole.

```
scale(MG)
         Area    Patches   Decrease
1  0.6132387  1.9181752 -0.9120837
2  0.6132387 -0.3926973 -0.4721108
3  0.6132387 -0.7551871  1.3882683
4 -0.7918519 -0.1208299 -0.6242002
5  0.6370538  0.1057262 -0.5372920
6 -1.6849179 -0.7551871  1.1574184
attr(,"scaled:center")
    Area  Patches Decrease
 7.32500 17.66667 37.48333
attr(,"scaled:scale")
     Area   Patches  Decrease
 4.199018 22.069587 36.820452
```

This standardization produces dimensions that have a mean of zero and a standard deviation of one, a fact that we will see below is very useful for quantitatively identifying outliers.

1.5.2 | Identifying Outliers in One Dimension

Data points with extreme standardized (or unstandardized) values on a single dimension are referred to as univariate or one-dimensional *outliers*. Because standardizing produces dimensions with a mean of zero and a standard deviation of one, these scores can be compared to a *standard normal distribution* with a mean of zero and a standard deviation of one.

We can use this theoretical distribution to identify a *critical value* beyond which a data point is so far removed from the mean that it can be considered to be a significant outlier. Common thresholds for delineating outliers include critical values corresponding to the most extreme 1% or 0.1% of the standard normal distribution. Because normal distributions include both positive and negative values, these correspond to both the lower and upper 0.5% (for a total of 1%) and to both the lower and upper 0.05% (for a total of 0.1%) of the distribution, respectively.

In R, we can use the `qnorm()` function to quickly identify the threshold corresponding to these or any other percentile.

```
qnorm(1 - 0.01 / 2)
[1] 2.575829
```

Note that, much like with the `quantile()` function, the percentiles must be expressed as a proportion. Although not strictly necessary, it is common to evaluate such thresholds using the complement of the desired level of significance, hence the `1 - 0.01/2` rather than just `0.01/2`.

Based on this output, any data point with a standard score in excess of either positive or negative 2.576 would be considered an outlier at a significance level of 1%. Note again that because the normal distribution and the standard scores that are compared to it can be both positive and negative, both directions should be considered when investigating outlier.

In R, we can identify one-dimensional outliers by comparing the absolute value of the standardized data points in a dimension to a desired critical value using basic operations.

```
abs(scale(D[,1])) > 2.575829
   [,1]
1 FALSE
2 FALSE
3 FALSE
4 FALSE
5 FALSE
6 FALSE
```

Using this output, we can see that there are no univariate outliers in the first dimension of our data frame. A similar analysis could then be performed for each of the remaining dimensions.

It is important to recognize, however, that outliers in and of themselves are not necessarily problematic to data analyses. Rather, we must determine if the inclusion or removal of an outlier would substantially alter the results of any further analyses.

Outliers that substantially alter the final results of an analysis are referred to as *influential* and should be considered for removal or transformation to limit their effect on the analysis. Outliers that are not influential, on the other hand, pose very little threat and can be retained in the data as they are.

For one-dimensional analyses, a common method of investigating influence is to calculate the standard measures of central tendency and dispersion both with and without outlying values. If, for example the skewness or kurtosis of a dimension changes substantially when an outlier is removed, this would indicate that the outlier is influential.

Another means of investigating influence is to determine if removing an outlier changes the skewness from positive (i.e., right-skewed) to negative (i.e., left-skewed), the kurtosis from less than 3 (i.e., platykurtic) to greater than 3 (i.e., leptokurtic), or vice versa. This change in interpretation is generally considered to be much more concerning than a change in magnitude that still leads to the same interpretation or conclusion.

1.5.3 | Quantifying Distance in Two Dimensions

The idea of measuring distance from the mean in one dimension can be easily expanded into a two-dimensional space. As we have seen before, data points measured on two dimensions often form a cloud of points when visualized. The central tendency or centroid of this cloud of points can be expressed using cartesian coordinates as $(\bar{x}_1, \bar{x}_2)$.

Expanding upon the one-dimensional deviation defined earlier, we can express the two-dimensional deviation of a data point from the centroid as a vector composite of the corresponding one-dimensional deviations:

$$x_i - \bar{x} = [x_{i,1} - \bar{x}_1 \quad x_{i,2} - \bar{x}_2] \tag{1.23}$$

As with one-dimensional deviations, this approach does not provide a standardized measure that can be easily compared between different pairs of dimensions. This also makes it difficult to identify a critical two-dimensional distance beyond which a point would be considered an outlier.

The *mahalanobis distance*, on the other hand, provides a standardized and unitless measure of two-dimensional distance from the centroid. It is also a natural extension of the one-dimensional standardized score introduced earlier, in that it also divides a deviation by a measure of variability. However, now that we must consider the interaction between the two

dimensions in addition to each of the dimensions themselves, this measure must include the covariability between dimensions as well as the variabilities of the dimensions themselves.

The most common way to achieve such a division is to use a variance-covariance matrix and some matrix algebra:

$$MD = (x_i - \bar{x}) \times \boldsymbol{\Sigma^{-1}} \times (x_i - \bar{x})^T \tag{1.24}$$

Once we know the variance-covariance matrix, we can use R to manually calculate this distance for any data point using some matrix algebra. First, however, we must explicitly convert the two-dimensional deviation into a matrix.

```
DEV <- matrix(MG[1,1:2] - c(mean(MG[,1]), mean(MG[,2])), nrow = 1)
SIGMA <- cov(MG[,1:2])
DEV %*% solve(SIGMA) %*% t(DEV)
          [,1]
[1,] 3.695879
```

We can also use the `mahalanobis()` function to more quickly arrive at this same value by using the `x`, `center`, and `cov` arguments to specify the two-dimensional data point, the means of each dimension, and the corresponding covariance matrix between the two dimensions.

```
mahalanobis(x = MG[1,1:2],
            center = c(mean(MG[,1]), mean(MG[,2])),
            cov = cov(MG[,1:2]))
[1] 3.695879
```

Similar to one-dimensional standardized scores, the magnitude of the mahalanobis distance indicates how far the data point is from the centroid, with higher measures denoting more extreme values. A distance of zero, for example, would again indicate that a data point is exactly average on both the first and the second dimension being measured. Unlike the standardized score, however, the mahalanobis distance is always positive.

Using our example data, we can see that the first data point is approximately 3.70 standardized distances from the centroid on the first and second dimensions. Note that unlike standard scores, this value is not commonly interpreted as the number of some measures of variability away from the centroid.

The `mahalanobis()` function can also technically be used to find a standardized distance in a one-dimensional space.

```
mahalanobis(x = MG[1,1], center = mean(MG[,1]), cov = var(MG[,1]))
[1] 0.3760617
```

Note that this value is exactly the square of the standardized score calculated earlier for this same data point.

```
(MG[1,1] - mean(MG[,1])) / sd(MG[,1])
[1] 0.6132387

((MG[1,1] - mean(MG[,1])) / sd(MG[,1])) ^ 2
[1] 0.3760617
```

This is due in part to the fact that standardized scores use the standard deviation, whereas the mahalanobis distance uses the variance, which is equal to the square of the standard deviation.

1.5.4 | Identifying Outliers in Two Dimensions

Data points with unusual combinations of values on two dimensions are referred to as *bivariate* or two-dimensional *outliers*. It is interesting to note that a data point that is not a univariate outlier on either of two dimensions can still be a bivariate outlier when the dimensions are considered together. For example, it is perfectly normal to find a gray wolf in Minnesota that weighs 42 kg (approximately 93 lb.). Similarly, finding a wolf that is 11 months old would not be unusual in terms of age. However, a wolf that weights 42 kg but is only 11 months old would be extremely unusual and would likely constitute a bivariate outlier.

To quantify how unusual a combination of variables is on two dimensions, we use the mahalanobis distance in much the same way as we did the standardized score. Unlike the standardized score, however, standardizing data to the mahalanobis distance does not produce distributions with a mean of zero and a standard deviation. As such, we cannot compare them to the standard normal distribution or otherwise use that distribution to identify critical values. Instead, mahalanobis distances are compared to what is known as a *chi-square distribution*.

There are two main things that differentiate this comparison from the one used previously for univariate outliers. First, the chi-square distribution does not contain any negative values; as such, a significance level of 1% is denoted as the 0.01 percentile without any additional division. Second, the overall shape and the values in a chi-square distribution actually change based on the *degrees of freedom* in the data, which must be identified prior to finding a critical value.

For a two-dimensional mahalanobis distance, the degrees of freedom are simply equal to two, one for each of the dimensions used in the original calculation of distance. With that information, we can use the `qchisq()` function in R, with the `df` argument set to the number of dimensions, to identify the critical threshold for any desired level of significance.

```
qchisq(p = 1 - 0.01, df = 2)
[1] 9.21034
```

Based on this output, any data point with a two-dimensional mahalanobis distance in excess of 9.21 would be considered a bivariate outlier at a significance level of 1%.

The qchisq() function can also technically be used to identify critical values for outliers in one dimension by setting the df argument equal to 1 to indicate a single dimension.

```
qchisq(p = 1 - 0.01, df = 1)
[1] 6.634897
```

However, much like the one-dimensional distance derived using the mahalanobis() function, this one-dimensional critical value is actually the square of that produced by the qnorm() function earlier.

```
qnorm(1 - 0.01 / 2)
[1] 2.575829

(qnorm(1 - 0.01 / 2)) ^ 2
[1] 6.634897
```

Note again the need to account for both positive and negative values in the normal distribution.

As before, we can also use the mahalanobis() function to calculate the mahalanobis distance for each case in a data matrix simultaneously.

```
m <- mahalanobis(x = MG[,1:2],
                 center = c(mean(MG[,1]), mean(MG[,2])),
                 cov = cov(MG[,1:2]))
m
        1         2         3         4         5         6
3.6958790 0.8357240 1.5212701 0.6655062 0.4279179 2.8537029
```

Note, however, that such an approach can only be applied to a single pair of dimensions at a time. If we desire to calculate mahalanobis distances for every pairwise combination of dimensions, this approach will need to be repeated multiple times.

Once we have these distances, we can again use basic operators in R to quickly identify which ones constitute significant outliers.

```
m > 9.21034
    1     2     3     4     5     6
FALSE FALSE FALSE FALSE FALSE FALSE
```

Unlike the standard scores used in one-dimension, mahalanobis distances cannot be negative, so we do not need to check any negative values. Using this output, we can see that there are no

two-dimensional outliers on the first two dimensions of our example data, at least not at a 1% level of significance. A similar analysis should then be performed on all other pairwise combinations of dimensions in the data matrix to identify any and all significant outliers.

One final note is that, as with one-dimensional or univariate outliers, the presence of two-dimensional or bivariate outliers isn't necessarily cause for concern. Instead, care should be taken to determine if any identified outliers are influential in further analyses.

For two-dimensional analyses, a common metric that forms the basis for many further analyses (including exploratory factor analysis, which we will discuss in Chapter 2) is the correlation, making it an excellent choice for investigating the influence of any potential outliers. As with one-dimensional outliers, any substantial change in value or (more concerningly) a change from positive to negative or vice versa would be indicative of an influential outlier.

```
cor(MG[,1:2])
              Area   Patches
Area    1.0000000 0.3815666
Patches 0.3815666 1.0000000
```

Although the first and second dimensions in our example data did not have any significant two-dimensional outliers, we can still see the effect of removing the most outlying point, which happens to be the first data point with a mahalanobis distance of approximately 3.70.

```
mean(MG[-6,1])
              Area   Patches
Area    1.0000000 0.3042992
Patches 0.3042992 1.0000000
```

Removing this data point did not substantially change the measure of covariability; which is to be expected given that it did not yield a significantly outlying measure of mahalanobis distance.

1.5.5 | Distance and Outliers in Higher Dimensions

Much like the bivariate scatter plots discussed earlier, these notions of centroids, mahalanobis distances, and critical values can all technically be expanded into three or more dimensions. However, just as with three-dimensional scatter plots, doing so does not substantially enhance our understanding of how extreme data points are when compared to measures of central tendency and dispersion.

Instead, a pairwise approach can be used to explore bivariate distance and corresponding outliers for every pair of available dimensions. If no outliers are detected in any combination of two dimensions, then it is very unlikely that any exist in three or more dimensions.

Exercises for Section 1.5

1.17 Multidimensional Poverty Measure (continued). Exercise 1.02 introduced a sample of data collected by the World Bank Group on different South American countries' access to education, basic infrastructure, and other dimensions of poverty, which is recreated below:

	Water	Electricity	Sanitation	Education
Argentina	0.3	0.0	0.4	1.5
Bolivia	7.4	4.9	16.3	13.2
Brazil	1.7	0.2	34.2	16.0
Chile	0.1	0.3	0.6	4.0
Colombia	2.4	1.3	8.2	5.1
Ecuador	4.3	1.4	3.6	3.9
Paraguay	2.1	0.3	9.0	6.3
Peru	6.2	4.1	12.1	5.4
Uruguay	0.5	0.1	1.0	2.0

The values above denote what percent of each country's population did not have access to the designated dimension of poverty in 2019.

a. Construct this data matrix in R with the proper row and column names
b. Without doing any calculations, indicate which countries are likely to have significant univariate outliers in each of the four dimensions
c. Find the standardized score for each data point in each of the four dimensions
d. Find the corresponding critical threshold for these standardized scores that would indicate outliers at the 1% level of significance
e. Determine which countries contain significant outliers at this level of significance
f. Describe any difference between this list of outliers and those indicated above

1.18 Stock Portfolios (continued). Exercise 1.13 introduced summary data on the performance of a diverse stock portfolio that spanned five different crypto currency stocks, which are recreated below:

	SOL	BNB	ETH	BTC	DOGE
SOL	3805	4980	42330	468665	0.354
BNB	4980	8293	64715	700917	1.401
ETH	42330	64715	559182	6294123	14.990
BTC	468665	700917	6294123	80834339	249.131
DOGE	0.354	1.401	14.990	249.131	0.005

The covariance matrix above was calculated using the daily adjusted closing values of five different crypto currency stocks over the course of a single year.

a. Construct this covariance matrix in R with the proper row and column names
b. Use matrix algebra to find the two-dimensional mahalanobis distance for a day when the daily return of BTC (Bitcoin) was 8211.057 below average and that of DOGE (Dogecoin) was 0.105 below average

c. Find the corresponding critical thresholds for this distance at the 0.1%, 1%, and 5% levels of significance
d. Use the values above to determine if this day constituted a significant two-dimensional outlier in terms of the joint performance of these two stocks at any of the commonly used levels of significance

1.19 Natural Gas Prices (continued). Exercise 1.07 introduced an excerpt of the data collected on the monthly price of natural gas since 1997, which is recreated below:

	Price
2017-05	3.15
2017-06	2.98
2017-07	2.98
2017-08	2.9
2017-09	2.98
2017-10	2.88
2017-11	3.01
2017-12	2.82
2018-01	3.87
2018-02	2.67
2018-03	2.69
2018-04	2.8
2018-05	2.8
2018-06	2.97
2018-07	2.83
2018-08	2.96
2018-09	3
2018-10	3.28
2018-11	4.09
2018-12	4.04
2019-01	3.11
2019-02	2.69
2019-03	2.95
2019-04	2.65
2019-05	2.64
2019-06	2.4
2019-07	2.37
2019-08	2.22
2019-09	2.56
2019-10	2.33
2019-11	2.65
2019-12	2.22
2020-01	2.02
2020-02	1.91
2020-03	1.79
2020-04	1.74
2020-05	1.75
2020-06	1.63
2020-07	1.77
2020-08	2.3

The values above denote the average monthly price of natural gas over a period of 40 months.

a. Construct this data matrix in R with the proper row and column names
b. Use the `which()` function to determine which months were significant outliers at the 5% level of significance
c. Find the skewness for these data both with and without any significant outliers
d. Describe any differences you see between the two measures of skewness above and use this information to determine if the outliers identified earlier were influential

1.20 Anscombe's Quartet (continued). Exercise 1.12 introduced four data sets that have nearly identical simple descriptive statistics but very different distributions, which are recreated below:

	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 mahalanobis distance for every data point in the third quartet of dimensions (i.e., `X3` and `Y3`)
c. Find the critical value beyond which these mahalanobis distances would indicate that a data point is an outlier at a 5% level of significance
d. Determine which data points are outliers at this level of significance
e. Find the correlation for this quartet both with and without any significant outliers
f. Describe any differences you see between the two correlations above and use this information to determine if the outliers identified earlier were influential at the 5% level of significance