The interquartile range

For some purposes, it is important to measure the variability of the centrally located values. If, for example, we put the observations in ascending order, how much variability is there among the central half of the data? The last example illustrated that the sample variance can be unsatisfactory in this regard. An alternative approach, which has practical importance, is the *interquartile range*, which is just $q_2 - q_1$, the difference between the upper and lower quartiles.

Notice that the interquartile range is insensitive to the more extreme values under study. As previously noted, the upper and lower quartiles are resistant to outliers, which means that the most extreme values do not affect the values of q_1 and q_2. Consequently, the interquartile range is resistant to outliers as well.

Example 4

Consider again the 10 values 50, 50, 50, 50, 50, 50, 50, 50, 50, 50. The interquartile range is zero. If we decrease the first value to 20 and increase the last to 80, the interquartile range is still zero because it measure the variability of the central half of the data, while ignoring the upper and lower fourth of the observations. Indeed, no matter how small we make the first value, and no matter how much we increase the last value, the interquartile range remains zero.

Problems

15. The height of 10 plants is measured in inches and found to be 12, 6, 15, 3, 12, 6, 21, 15, 18 and 12. Verify that $\sum(X_i - \bar{X}) = 0$.

16. For the data in the previous problem, compute the range, variance and standard deviation.

17. Use the rules of summation notation to show that it is always the case that $\sum(X_i - \bar{X}) = 0$.

18. Seven different thermometers were used to measure the temperature of a substance. The readings in degrees Celsius are −4.10, −4.13, −5.09, −4.08, −4.10, −4.09 and −4.12. Find the variance and standard deviation.

19. A weightlifter's maximum bench press (in pounds) in each of six successive weeks was 280, 295, 275, 305, 300, 290. Find the standard deviation.

2.4 Detecting outliers

The detection of outliers is important for a variety of reasons. One rather mundane reason is that they can help identify erroneously recorded results. We have already seen that even a single outlier can grossly affect the sample mean and variance, and of course we do not want a typing error to substantially alter or color our perceptions of the data. Such errors seem to be rampant in applied work, and the subsequent cost of such errors can be enormous (De Veaux and Hand, 2005). So it can be prudent to check for outliers, and if any are found, make sure they are valid.

But even if data are recorded accurately, it cannot be stressed too strongly that modern outlier detection techniques suggest that outliers are more the rule rather than the exception. That is, unusually small or large values occur naturally in a wide range of situations. Interestingly, in 1960, the renowned statistician John Tukey (1915–2000) predicted that in general we should expect outliers. What is fascinating about his prediction is that it was made before good outlier detection techniques were available.

A simple approach to detecting outliers is to merely look at the data. And another possibility is to inspect graphs of the data described in chapter 3. But for various purposes (to be described), these two approaches are unsatisfactory. What is needed are outlier detection techniques that have certain properties, the nature of which, and why they are important, is impossible to appreciate at this point. But one basic goal is easy to understand. A fundamental requirement of any outlier detection technique is that it does not suffer from what is called masking. An outlier detection technique is said to suffer from *masking* if the very presence of outliers causes them to be missed.

A classic outlier detection method

A classic outlier detection technique illustrates the problem of masking. This classic technique declares the value X an outlier if

$$\frac{|X-\bar{X}|}{s} \geq 2. \tag{2.3}$$

(The value 2 in this last equation is motivated by results covered in chapter 4.)

Example 1

Consider the values

$$2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 1{,}000.$$

The sample mean is $\bar{X} = 65.94$, the sample standard deviation is $s = 249.1$,

$$\frac{|1{,}000 - 65.94|}{249.1} = 3.75,$$

3.75 is greater than 2, so the value 1,000 is declared an outlier. In this particular case, the classic outlier detection method is performing in a reasonable manner; it identifies what is surely an usual value.

Example 2

Now consider the values

$$2, 2, 3, 3, 3, 4, 4, 4, 100{,}000, 100{,}000.$$

The sample mean is $\bar{X} = 20{,}002.5$, the sample standard deviation is $s = 42{,}162.38$,

$$\frac{|100{,}000 - 20{,}002.5|}{42{,}162.38} = 1.897,$$

and so the classic method would not declare the value 100,000 an outlier even though certainly it is highly unusual relative to the other eight values. The problem is that both the sample mean and the sample standard deviation are

sensitive to outliers. That is, the classic method for detecting outliers suffers from masking. It is left as an exercise to show that even if the two values 100,000 in this example are increased to 10,000,000, the value 10,000,000 is not declared an outlier.

In some cases the classic outlier detection rule will detect the largest outlier but miss other values that are clearly unusual. Consider the sexual attitude data in table 2.3. It is evident that the response 6,000 is unusually large. But even the response 150 seems very large relative to the majority of values listed, yet the classic rule does not flag it as an outlier.

The boxplot rule

One of the earliest improvements on the classic outlier detection rule is called the *boxplot rule*. It is based on the fundamental strategy of avoiding masking by replacing the mean and standard deviation with measures of location and dispersion that are relatively insensitive to outliers. In particular, the *boxplot rule* declares the value X an outlier if

$$X < q_1 - 1.5(q_2 - q_1) \tag{2.4}$$

or

$$X > q_2 + 1.5(q_2 - q_1). \tag{2.5}$$

So the rule is based on the lower and upper quartiles, as well as the interquartile range, which provide resistance to outliers.

Example 3

Consider the values

$$1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 100, 500.$$

A little arithmetic shows that the lower quartile is $q_1 = 4.417$, the upper quartile is $q_2 = 12.583$, so $q_2 + 1.5(q_2 - q_1) = 12.583 + 1.5(12.583 - 4.417) = 24.83$. That is, any value greater than 24.83 is declared an outlier. In particular, the values 100 and 500 are labeled outliers.

Example 4

For the sexual attitude data in table 2.3, the classic outlier detection rule declares only one value to be an outlier: the largest response, 6,000. In contrast, the boxplot rule labels all values 15 and larger as outliers. So of the 105 responses, the classic outlier detection rule finds only one outlier, and the boxplot rule finds 12.

Problems

20. For the values

$$20, 121, 132, 123, 145, 151, 119, 133, 134, 130,$$

use the classic outlier detection rule to determine whether any outliers exist.

21. Apply the boxplot rule for outliers to the values in the preceding problem.
22. Consider the values

$$0, 121, 132, 123, 145, 151, 119, 133, 134, 130, 250.$$

Are the values 0 and 250 declared outliers using the classic outlier detection rule?

23. Verify that for the data in the previous problem, the boxplot rule declares the values 0 and 250 outliers.
24. Consider the values

$$20, 121, 132, 123, 145, 151, 119, 133, 134, 240, 250.$$

Verify that no outliers are found using the classic outlier detection rule.

25. Verify that for the data in the previous problem, the boxplot rule declares the values 20, 240, and 250 outliers.
26. What do the last three problems suggest about the boxplot rule versus the classic rule for detecting outliers?

2.5 Some modern advances and insights

During the last half-century, and particularly during the last twenty years, there have been major advances and insights relevant to the most basic methods covered in an introductory statistics course. Most of these advances cannot be covered here, but it is very important to at least alert students to some of the more important advances and insights and to provide a glimpse of why more modern techniques have practical value. The material covered here will help achieve this goal.

Means, medians and trimming

The *mean* and *median* are the two best-known measures of location, with the mean being used in a large proportion of applied investigations. There are circumstances where using a mean gives satisfactory results. Indeed, there are conditions where it is optimal (versus any other measure of location that might be used.) But recent advances and insights have made it clear that both the mean and median can be highly unsatisfactory for a wide range of practical situations. Many new methods have been developed for dealing with known problems, some of which are based in part on using measures of location other than the mean and median. One of the simpler alternatives is introduced here.

The sample median is an example of what is called a *trimmed mean*; it trims all but one or two values. Although there are circumstances where this extreme amount of trimming can be beneficial, for various reasons covered in subsequent chapters, this extreme amount of trimming can be detrimental. The sample mean represents the other extreme: zero trimming. We have already seen that this can result in a measure of location that is a rather poor reflection of what is a typical observation. But even when it provides a good indication of the typical value, many basic methods based on the mean suffer from other fundamental concerns yet to be described. One way of reducing these problems is to use a compromise amount of trimming. That is, trim some values, but not as many

as done by the median. No specific amount of trimming is always best, but for various reasons, 20% trimming is often a good choice. This means that the smallest 20%, as well as the largest 20%, are trimmed and the average of the remaining data is computed. In symbols, first compute $.2n$, round down to the nearest integer, call this result g, in which case the 20% trimmed mean is given by

$$\bar{X}_t = \frac{1}{n-2g}(X_{(g+1)} + \cdots + X_{(n-g)}). \tag{2.6}$$

Example 1

Consider the values

$$46, 12, 33, 15, 29, 19, 4, 24, 11, 31, 38, 69, 10.$$

Putting these values in ascending order yields,

$$4, 10, 11, 12, 15, 19, 24, 29, 31, 33, 38, 46, 69.$$

The number of observations is $n = 13$, $0.2(n) = 0.2(13) = 2.6$, and rounding this down to the nearest integer yields $g = 2$. That is, trim the two smallest values, 4 and 10, trim the two largest values, 46 and 69, and average the numbers that remain yielding

$$\bar{X}_t = \frac{1}{9}(11 + 12 + 15 + 19 + 24 + 29 + 31 + 33 + 38) = 23.56.$$

Example 2

Imagine a figure skating contest that uses nine judges who rate a skater on a six-point scale. Suppose the nine ratings are

$$5.1, 5.3, 5.3, 5.5, 5.0, 5.1, 5.4, 4.2, 5.2.$$

A natural concern is that some raters might not be fair under certain circumstances, or they might provide a poor reflection of how most raters would judge the skater, which in turn might make a difference in a competition. From a statistical point of view, we do not want an unusual rating to overly influence our measure of the typical rating a skater would receive. For the data at hand, the sample mean is 5.1, but notice that the rating 4.2 is unusually small compared to the remaining eight. To guard against unusually high or low ratings, it is common in skating competitions to throw out the highest and lowest scores and average those that remain. Here, $n = 9$, $0.2n = 1.8$, so $g = 1$. That is, a 20% trimmed mean corresponds to throwing out the lowest and highest scores and averaging the ratings that remain, yielding $\bar{X}_t = 5.2$.

Other measures of location

Yet another approach when measuring location is to check for outliers, remove any that are found, and then average the remaining values. There are, in fact, several variations of this strategy. There are circumstances where this approach has practical value, but the process of removing outliers creates certain technical problems that require advanced

techniques that go beyond the scope of this book.[2] Consequently, this approach to measuring location is not discussed further.

Winsorized data and the winsorized variance

When using a trimmed mean, certain types of analyses, to be covered later, are not done in an intuitively obvious manner based on standard training. To illustrate how technically correct methods are applied, we will need to know how to Winsorize data and how to compute the Winsorized variance.

The process of Winsorizing data by 20% is related to 20% trimming. When we compute a 20% trimmed mean, we compute g as previously described, remove the g smallest and largest observations, and average the remaining values. Winsorizing the data by 20% means that the g smallest values are not trimmed, but rather, they are set equal to the smallest value not trimmed. Similarly, the g largest values are set equal to the largest value not trimmed.

Example 3

Suppose the reaction times of individuals are measured yielding

$$2, 3, 4, 5, 6, 7, 8, 9, 10, 50.$$

There are $n = 10$ values, $0.2(10) = 2$, so $g = 2$. Here, 20% Winsorizing of the data means that the two smallest values are set equal to 4. Simultaneously the two largest observations, 10 and 50, are set equal to 9, the largest value not trimmed. That is, 20% Winsorizing of the data yields

$$4, 4, 4, 5, 6, 7, 8, 9, 9, 9.$$

In symbols, the observations $X_1, \ldots, X_n$ are Winsorized by first putting the observations in order yielding $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. Then the g smallest observations are replaced by $X_{(g+1)}$, and the g largest observations are replaced by $X_{(n-g)}$.

Example 4

To Winsorize the values

$$10, 8, 22, 35, 42, 2, 9, 18, 27, 1, 16, 29$$

using 20% Winsorization, first note that there are $n = 12$ observations, $.2 \times 12 = 2.4$, and rounding down gives $g = 2$. Putting the values in order yields

$$1, 2, 8, 9, 10, 16, 18, 22, 27, 29, 35, 42.$$

Then the two smallest values are replaced by $X_{(g+1)} = X_{(3)} = 8$, the two largest values are replaced by $X_{(n-g)} = X_{(10)} = 29$, and the resulting Winsorized values are

$$8, 8, 8, 9, 10, 16, 18, 22, 27, 29, 29, 29.$$

2. The technical problems are related to methods for testing hypotheses, a topic introduced in chapter 7.

The *Winsorized sample variance* is just the sample variance based on the Winsorized values and will be labeled s_w^2. In symbols, if $W_1, \ldots, W_n$ are the Winsorized values,

$$s_w^2 = \frac{1}{n-1}\sum (W_i - \overline{W})^2, \tag{2.7}$$

where

$$\overline{W} = \frac{1}{n}\sum W_i,$$

the average of the Winsorized values. The sample mean of the Winsorized values, $\overline{W}$, is called the *sample Winsorized mean*. The *Winsorized sample standard deviation* is the square root of the Winsorized sample variance, s_w.

Example 5

To compute the 20% Winsorized mean and variance for the observations

$$1, 2, 8, 9, 10, 16, 18, 22, 27, 29, 35, 42,$$

first Winsorize these values yielding

$$8, 8, 8, 9, 10, 16, 18, 22, 27, 29, 29, 29.$$

The mean of these Winsorized values is the Winsorized mean given by

$$\overline{X}_w = \frac{8+8+8+9+10+16+18+22+27+29+29+29}{12} = 17.75.$$

The Winsorized sample variance is

$$s_w^2 = \frac{(8^*-17.75)^2 + (8-17.75)^2 + \cdots + (29-17.75)^2}{12-1} = 82.57.$$

The Winsorized sample standard deviation is $s_w = \sqrt{82.57} = 9.1$.

For the observations in the last example, the sample mean is $\bar{X} = 18.25$ and the sample variance is $s^2 = 170.57$, which is about twice as large as the sample Winsorized variance, $s_w^2 = 82.57$. Notice that the Winsorized variance is less sensitive to extreme observations and roughly reflects the variation for the middle portion of your data. In contrast, the sample variance, s^2, is highly sensitive to extreme values. This difference between the sample variance and the Winsorized sample variance will be seen to be important.

Example 6

For the data in the last example, suppose we increase the largest value, 42, to 60. Then the sample mean increases from 18.25 to 19.75 and the sample variance, s^2, increases from 170.57 to 275.3. In contrast, the Winsorized sample mean and variance do not increase at all, they are still equal to 17.75 and 82.57, respectively. The sample Winsorized variance provides resistance to outliers because its value does not increase as we increase the largest observation, a property that will turn out to have great practical value.

A Summary of Some Key Points

- Several measures of location were introduced. How and when should one measure of location be preferred over another? It is much too soon to discuss this issue in a satisfactory manner. An adequate answer depends in part on concepts yet to be described. For now, the main point is that different measures of location vary in how sensitive they are to outliers.
- The sample mean can be highly sensitive to outliers. For some purposes, this is desirable, but in many situations this creates practical problems, as will be demonstrated in subsequent chapters.
- The median is highly insensitive to outliers. This plays an important role in some situations, but the median has some negative characteristics yet to be described.
- In terms of sensitivity to outliers, the 20% trimmed mean lies between two extremes: no trimming (the mean) and the maximum amount of trimming (the median).
- The sample variance also is highly sensitive to outliers. We saw that this property creates difficulties when checking for outliers (it results in masking), and some additional concerns will become evident later in this book.
- The interquartile range measures variability without being sensitive to the more extreme values. This property makes it well suited to detecting outliers.
- The 20% Winsorized variance also measures variation without being sensitive to the more extreme values. But it is too soon to explain why it has practical importance.

Problems

27. What is the typical pulse rate (beats per minute) among adults? Imagine that you sample 21 adults, measure their pulse rate and get

$$80, 85, 81, 75, 77, 79, 74, 86, 79, 55$$

$$82, 89, 73, 79, 83, 82, 88, 79, 77, 81, 82.$$

Compute the 20% trimmed mean.

28. For the observations

$$21, 36, 42, 24, 25, 36, 35, 49, 32$$

verify that the sample mean, trimmed mean and median are $\bar{X} = 33.33$, $\bar{X}_t = 32.9$ and $M = 35$.

29. The largest observation in the last problem is 49. If 49 is replaced by the value 200, verify that the sample mean is now $\bar{X} = 50.1$ but the trimmed mean and median are not changed.

30. For the last problem, what is the minimum number of observations that must be altered so that the trimmed mean is greater than 1,000?

31. Repeat the previous problem but use the median instead. What does this illustrate about the resistance of the mean, median and trimmed mean?

32. For the observations

$$6, 3, 2, 7, 6, 5, 8, 9, 8, 11$$

verify that the sample mean, trimmed mean and median are $\bar{X} = 6.5$, $\bar{X}_t = 6.7$ and $M = 6.5$.