Classroom Simulation: Indications of Outliers in Boxplots of Normal Data

Jacob B. Colvin, Bruce E. Trumbo, Eric A. Suess, *California State University, East Bay (Hayward Campus)*
Department of Statistics; California State University, East Bay; Hayward, CA 94542, USA
jbcolvin@fastmail.fm, bruce.trumbo@csueastbay.edu or eric.suess@csueastbay.edu

Introduction

Many computer packages make boxplots that indicate outliers. That is, values beyond fences located a certain multiple K of the interquartile range (IQR) on either side of the box bounded by the lower and upper quartiles. When $K = 3$ the resulting outliers are sometimes called *probable* outliers and when $K = 1.5$ they are sometimes called *possible* outliers. Some textbooks recommend against the use of normality based procedures, such as t tests, when outliers are present.

This raises the question about how often boxplots give indications of outliers for normal data. Because a study of the order statistics of normal distributions is beyond the scope of undergraduate mathematics, we show how relatively straightforward simulations in R can be used to answer this and related questions.

Confusing outlier indications for normal data

Figure 1 illustrates that potentially misleading outlier indications are fairly common with normal data. It shows boxplots based on 20 simulated normal samples of size 16; five of these boxplots have indications of possible outliers ($K = 1.5$). Because we know the data are normal, these 'outliers' surely do not indicate nonnormality. Simulation shows that on average about 30% of boxplots of $n = 16$ normal observations indicate outliers. If $n = 64$, then nearly half of normal samples will yield indications of possible outliers. For probable outliers ($K = 3$) the situation is reversed; probable outliers are more often indicated for $n = 16$ (about 2.5% of samples) than for $n = 64$ (about 0.5%). See Figures 2 and 3.

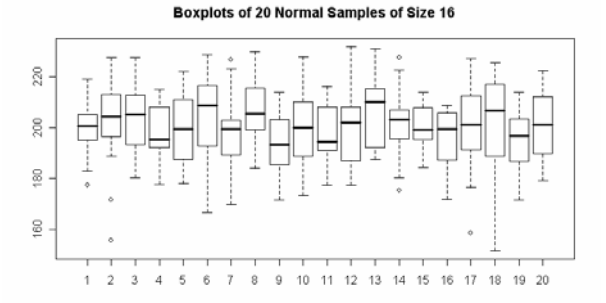

Figure 1. Five of these 20 boxplots of simulated normal data show 'outliers.' The expected number is about six.

Because boxplots give such frequent indications of possible outliers for normal data, because the probability of such outlier indications depends strongly on sample size and because boxplots carry no indication of sample size, a useful interpretation of possible outliers in boxplots is especially problematic.

This is not to deny that boxplot indications of outliers are useful. In practice—with real data—if an observation is indicated as an outlier, a responsible statistician would want to check whether this can be ascribed to a measurement or recording error. The message in our simulations is that one should not be too quick to assume data are nonnormal just because a boxplot shows an outlier.

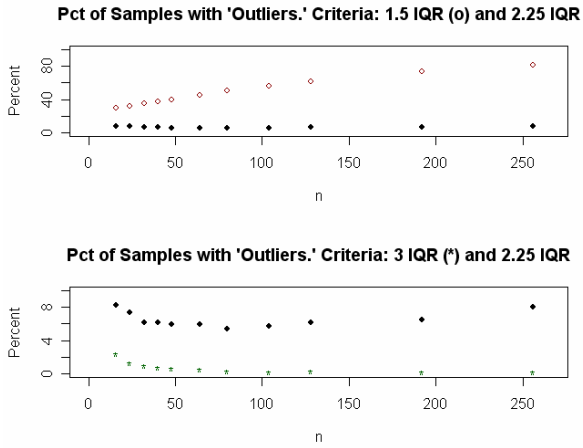

Figure 2. The upper panel compares probabilities of possible outliers ($K = 1.5$) with outliers based data points more than $K = 2.25$ IQR beyond the ends of the box. The lower panel compares $K = 2.25$ with $K = 3.0$.

Simulating the probability of false outliers

Figure 2 shows the probability of at least one outlier for normal samples of various sizes and for $K = 1.5$, 2.25, and 3.0. The value $K = 2.25$ is not a standard one. We show it because it has a relatively stable probability of outlier indications across samples of small to moderate size. Over the range of sample sizes shown, the 2.25 IQR criterion yields outlier indications for about 5-8% of normal samples. The 1.5 IQR criterion is increasingly likely to show (possible) outliers as sample size increases, and the (probable) outliers from the 3 IQR criterion become less likely as sample size increases. Some of the numbers used to make Figure 2 are shown in Figure 3. Each point is based on 10,000 samples.

n	1.5 IQR	2.25 IQR	3 IQR
16	**0.2950**	0.0823	**0.0244**
32	0.3475	0.0618	0.0102
48	0.3976	0.0590	0.0061
64	**0.4462**	0.0598	**0.0053**
104	0.5582	0.0566	0.0026
128	0.6153	0.0618	0.0032
192	0.7310	0.0649	0.0020

Figure 3. Probabilities of outlier indications as in Figure 2.

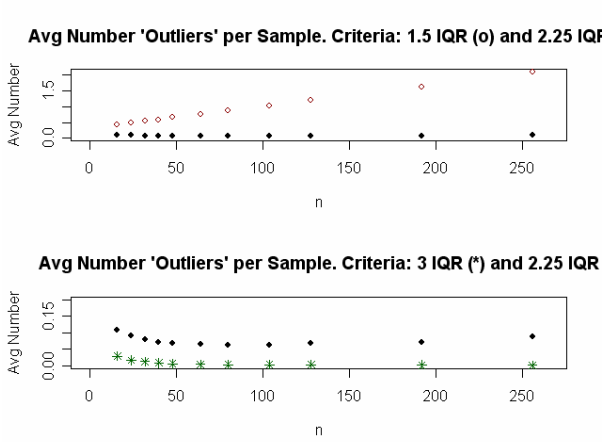

Figure 4. Average numbers of outlier indications for normal data. Each point is based on 10,000 samples.

It is not surprising that, as the probability of at least one outlier in a sample increases, so does the average number of outliers in a sample. However, multiple outliers are relatively rare for small normal samples. Figure 4 is similar to Figure 2, but it shows the average number of outliers per sample.

Controlling the probability of outlier indications

As the value of *K* increases, the probability decreases that a normal sample will show one or more outliers. For several sample sizes, the R code in Figure 5 finds values of *K* that restrict the probability of outliers to 1%, 7% and 20. (Code for some labels and other embellishments is omitted.) Results are plotted in Figure 6. The horizontal line shows that the value *K* = 2.25 we used above is somewhat arbitrary. Among possible values in the vicinity we chose it because it is midway between 1.5 and 3.0, and easy to remember.

```
m = 10000
n = c(24, 32, 40, 48, 64, 80, 104, 128,
   192, 256, 320)
k.99 = k.93 = k.80 = numeric(length(n))
for (i in 1:length(n))
{
  x = rnorm(m*n[i])
  DTA = matrix(x, nrow=m)
  q = t(apply(DTA, 1, quantile))
  iqr = q[,4] - q[,2]
  wskr = pmax(q[,2]-q[,1], q[,5]-q[,4])
  cnst = wskr/iqr
  k.93[i] = quantile(cnst, .93)
  k.99[i] = quantile(cnst, .99)
  k.80[i] = quantile(cnst, .80)
}
plot(n, k.93, ylim=c(1.5,3.5))
points(n, k.99, pch="*")
points(n, k.80, pch="x")
abline(h=2.25, lty=2)
```

Figure 5. R code used to produce Figure 6.

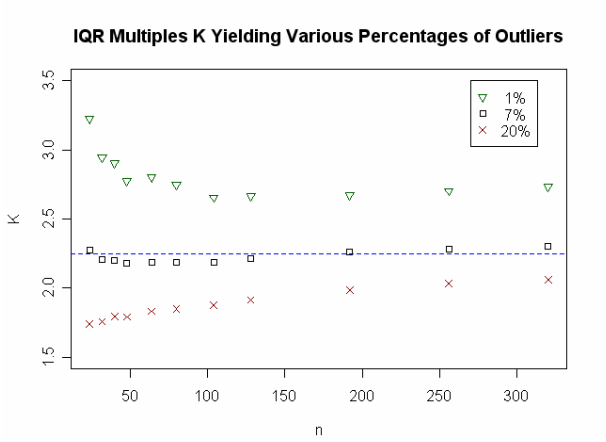

Figure 6. Values of *K* that yield various percentages of outlier indications in simulated normal samples

What is an outlier?

Vaguely speaking, an outlier is a value in a dataset that is, according to some criterion, located at an unusually large distance from the rest of the data. While boxplots are probably the most familiar way to look for outliers, there are many other methods of doing so. Another criterion, often used in regression and not explored here, is for a value to lie more than two standard deviations away from the mean of the data, where the value in question is omitted when the standard deviation is computed.

Particularly with supposedly normal data, an outlier may be traceable to equipment failure or a recording error—an observation that did not arise from the assumed normal process, but from some other kind of process. But for many skewed or long-tailed distributions (for example, exponential and other members of the gamma family, lognormal, Pareto, Cauchy, Laplace, and so on) outliers are an expected feature of the data.

Great care should be taken in omitting an outlier before data analysis, unless one can be sure the observation arose from a mechanism that is not a valid feature of the process under study. For example, in studying earthquakes, there are hundreds of low-magnitude seismic events every day that can be detected with special equipment., but it is *only* the extreme outliers that are of practical importance to the general public. In the next section we discuss some properties of boxplot outliers for a few nonnormal distributions.

Outliers in nonnormal distributions

As one would expect, simulation shows that a sample from a long-tailed distribution is more likely to yield boxplot outlier indications than is a normal sample of the same size. Figures 7 and 8 show results for t-distributed data (df = 3) and GAMMA(10, 1) data, respectively. They are similar to Figure 2, except that here it is feasible to plot results for all three values of *K* on the same scale.

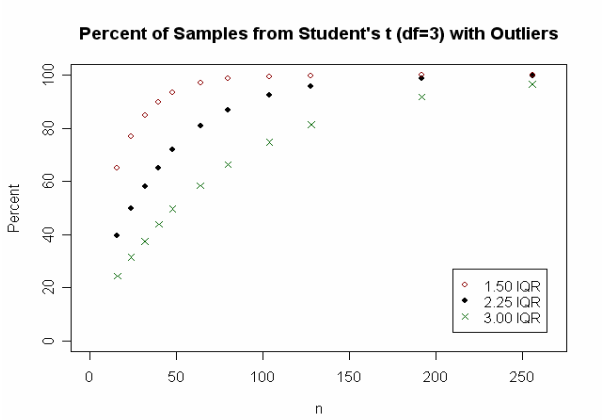

Figure 7. Percent of samples from a Student's t distribution with 3 degrees of freedom that show outliers.

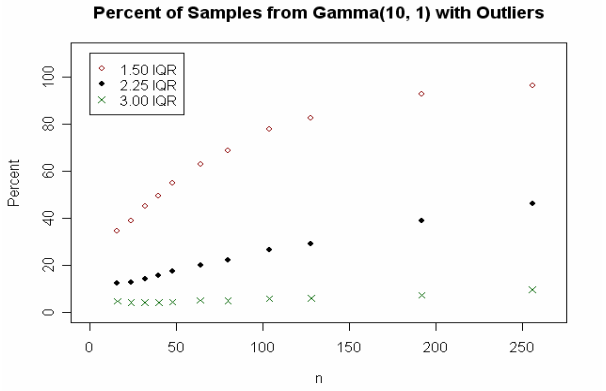

Figure 8. Percent of samples from a gamma distribution with shape parameter 10 that show outliers.

The corresponding graph for samples from the severely right-skewed exponential distribution looks very similar to Figure 7. In contrast, samples from a uniform distribution, which has no tails, typically show outliers only for very small sample sizes.

Again for nonnormal data, as the probability of outliers in a sample increases, so does its expected number of outliers. For samples of various sizes from an exponential distribution, Figure 9 shows the average number of outliers. (Here, each point is based on 1000 samples.) It appears that the number of outliers increases approximately linearly with sample size. Across the sample sizes we investigated, the criteria $K = 1.5$, 2.25, and 3.0 consistently designate as outliers about 5%, 2%, and 1% of observations, respectively.

Comments

We close with some comments on boxplots, outliers and the use of simulation in undergraduate classes.

- We have avoided samples of size $n = 15$ and smaller. It does not make sense to use boxplots for samples of size smaller than 10 or 15, because boxplots are based on the five-number summary, elements of which may behave erratically for small sample sizes.
- One should take care in using boxplot indications of outliers (especially from the criterion $K = 1.5$) to declare data to be nonnormal. In R, the default value is $K = 1.5$, but other values can be easily substituted. In Minitab, the professional graphics boxplots are based on $K = 1.5$, and the boxplots made with character graphics essentially use both $K = 3$ and $K = 1.5$. When normal data are anticipated, we think $K = 2.25$ might be a better choice.
- Simulations at the level used in this paper make intellectually stimulating projects for undergraduate and first year MS students. The number of R commands required can be kept within reasonable bounds, and program structure can be kept relatively simple. Routine simulations can illustrate distributional information that is widely used in practice, but beyond the level of such students to derive analytically. More ambitious simulation projects may reveal facts not found in standard texts.

Annotated References

Davies, Laurie and Gather, Ursula: "The identification of multiple outliers," *JASA,* Vol. 88 (1993), No. 423, pp 782-792, A view of outlier rules based on ideas of robustness.

Hoaglin, David C.; Iglewicz, Boris; Tukey, John W.: Performance of some resistant rules for outlier labeling, *JASA*, Vol. 81 (1986), No. 396, 991-999. An early simulation study, to which ours is somewhat similar.

Moore, David S.: *The Basic Practice of Statistics*, 3rd ed. (2004). Chapter 16 has an excellent discussion of outliers and the robustness of t-procedures.

Trumbo, Bruce E.: *Learning Statistics with Real Data,* Duxbury (2002). Unit 1 has several examples of outliers in real data arising through various mechanisms.

Tukey, John W.: *Exploratory Data Analysis,* Addison-Wesley (1977). Chapter 1 has one of the earlier published discussions of boxplots.

`www.sci.csueastbay.edu/~btrumbo/JSM2006/Outliers` has posters from JSM 2006 (with additional simulations by J. Colvin) and complete R code.

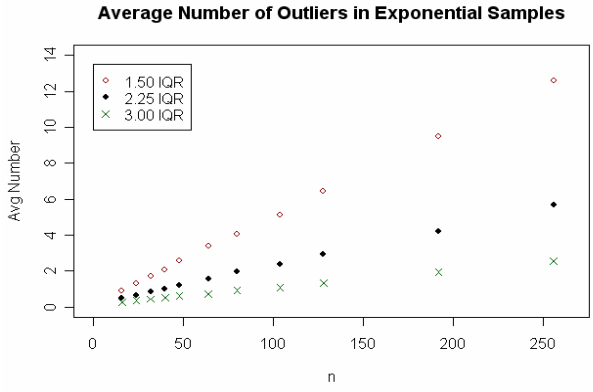

Figure 9. Average number of outlier indications in samples from an exponential distribution. Numbers of outliers increase approximately linearly with sample size. Each point is based on 1000 samples.