CHAPTER 3

Boxplots and Batch Comparison

John D. Emerson
Middlebury College

Judith Strenio
Westat Inc.

A graphical display of the five-number summary of a batch of numbers—the boxplot—shows much of the structure of the batch. From a boxplot we can pick out the following features of a batch:

Location
Spread
Skewness
Tail length
Outlying data points.

Thus the boxplot provides a visual impression of several important aspects of the empirical distribution of a batch of data.

This compact visual display is especially useful for comparing several batches of data. By drawing a boxplot for each batch and arranging them in parallel, we can compare the batches with respect to location and spread, and perhaps also skewness and tail heaviness. In this comparison, we may find that the data from the different batches do not all fit well into the same scale. In particular, batches located far from the origin may be much more spread out than batches located near the origin. Thus if the batches are plotted on a common scale, the details of batches close to the origin will be harder to see.

An appropriate transformation can often alleviate this difficulty by making the variability of the batches more nearly comparable. In drawing some guidance from the data as to what transformations may achieve these objectives, a spread-versus-level plot may suggest a power transformation that tends to equalize spread across different levels, or locations, of the batches.

Throughout this chapter, we restrict our examples to batches of measured or counted data. We assume that observations are nonnegative and possibly quite large. The origin then provides a lower bound, but there is no upper bound. Thus we do not consider the special features of such types of data as fractions bounded above by 1, and percents bounded above by 100%. (Such assumptions are not needed in Sections 3A and 3B.)

3A. THE BOXPLOT FOR A SINGLE BATCH

We introduce the boxplot for a single batch of data, using an example from the 1960 Census.

EXAMPLE:

The World Almanac (1967) reported the populations of United States cities; Table 3-1 gives the populations (to the nearest 10,000) of the 15 largest cities. We form the boxplot in order to pick out major features of the batch. In particular, we ask whether the batch of 15 cities is skewed and whether it has outlying data points.

As a first step in analysis, we construct the 5-number summary (see Section 2D), and we calculate the fourth-spread and the cutoffs for outliers based on the fourth-spread. The 5-number summary displays the median, fourths, and extremes of the batch:

#	15	U.S. Cities	
M	8	88	
F	4.5	74	184
	1	63	778

The *fourth-spread* is the range of the data defined by the upper fourth and lower fourth. It is closely related to the *interquartile range*, although technical differences between quartiles and fourths distinguish the two concepts.

Data values that are far enough beyond the fourths are considered as potential *outliers*. We use the fourth-spread to make this vague concept precise and give technical meaning to the term "outlier." Specifically, we

TABLE 3-1. Populations of the 15 largest U.S. cities in 1960.

City	Population (10,000s)
New York	778
Chicago	355
Los Angeles	248
Philadelphia	200
Detroit	167
Baltimore	94
Houston	94
Cleveland	88
Washington, DC	76
St. Louis	75
Milwaukee	74
San Francisco	74
Boston	70
Dallas	68
New Orleans	63

Source: *The World Almanac*, 1967 edition. New York: Newspaper Enterprise Association, Inc. (Data from p. 323).

define $F_L - \frac{3}{2}d_F$ and $F_U + \frac{3}{2}d_F$ as the *outlier cutoffs*, where F_L and F_U denote the fourths and d_F is $F_U - F_L$, the fourth-spread. Data values that are smaller than $F_L - \frac{3}{2}d_F$ or larger than $F_U + \frac{3}{2}d_F$ are called outliers and will receive special attention.

For the 15 cities,

$$d_F = 184 - 74 = 110,$$

$$F_L - \tfrac{3}{2}d_F = 74 - \tfrac{3}{2} \times 110 = -91,$$

and

$$F_U + \tfrac{3}{2}d_F = 184 + \tfrac{3}{2} \times 110 = 349.$$

Thus the outlier cutoffs are -91 and 349, and so cities with populations over 3,490,000 (namely, New York and Chicago) are classified as outliers.

To construct the boxplot, we first draw a box with ends at the lower fourth and upper fourth and a crossbar at the median. Next, we draw a line from each end of the box to the most remote point that is not an outlier.

The resulting figure
outliers. The outliers
outlier cutoffs.

Figure 3-1 shows
plot is drawn horizo
some settings.

The boxplot show
and outlying data p
median, the crossbar
the *spread*, using th
median, the upper f
skewness; the median
fourth, indicating tha
with unbounded pos
extended to New Orl
and New York).

The message of th
batch is heavily skew
greatest value of the b
information about the

Resistance of the Box

We have seen that th
needs the median and
fourths are resistant t
also resistant to gross
of the data values ca
disturbing the median

The "tails" of the b
data values that are
undisturbed by gross
only modestly affecte
outlier cutoffs. Of cou

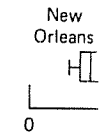


Figure 3-1.

The resulting figure schematically represents the body of data minus the outliers. The outliers are represented individually by xs situated beyond the outlier cutoffs.

Figure 3-1 shows the boxplot for the U.S. cities. To conserve space, this plot is drawn horizontally, but vertical plots may be more appropriate in some settings.

The boxplot shows at a glance the location, spread, skewness, tail length, and outlying data points. The *location* of the batch is summarized by the median, the crossbar in the interior of the box. The length of the box shows the *spread*, using the fourth-spread. From the relative positions of the median, the upper fourth, and the lower fourth, we also see some of the *skewness*; the median is much closer to the lower fourth than to the upper fourth, indicating that the batch is positively skewed—a common situation with unbounded positive data. The plot indicates *tail length* by the lines extended to New Orleans and to Los Angeles, and by the *outliers* (Chicago and New York).

The message of the boxplot for the 15 largest U.S. cities is strong: the batch is heavily skewed, and there are two outlying data points. But the greatest value of the boxplot is its ability to convey visually some important information about the shape of this batch of data.

Resistance of the Boxplot

We have seen that the construction of the rectangular box in the boxplot needs the median and the fourths of a data set. Because the median and the fourths are resistant to the impact of a few wild data values, the boxplot is also resistant to gross influence by these values. More specifically, up to 25% of the data values can be made arbitrarily large ("wild") without greatly disturbing the median, the fourths, or the rectangular box in the boxplot.

The "tails" of the boxplot are determined primarily by the most extreme data values that are within the outlier cutoffs. Thus they are relatively undisturbed by gross changes in the values of any outliers, and they can be only modestly affected by gross changes of values originally within the outlier cutoffs. Of course, the outlier cutoffs themselves are defined using

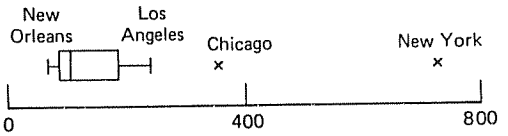


Figure 3-1. Boxplot for the 15 largest U.S. cities in 1960.

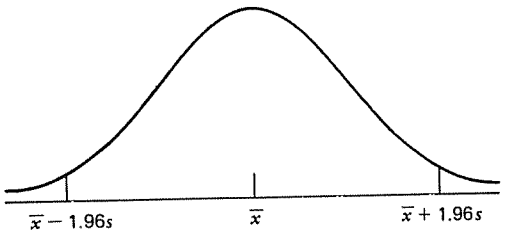
only the fourths of a batch. Thus they can resist gross disturbances in up to 25% of the data.

The resistance properties of the boxplot make it attractive for use in exploratory data analysis. Although an analogous plot could be based on the sample mean and sample standard deviation, such a plot would necessarily lack resistance to the influence of even a single wild data value.

Our definition of outliers as data values that are smaller than $F_L - \frac{3}{2}d_F$ or larger than $F_U + \frac{3}{2}d_F$ is somewhat arbitrary, but experience with many data sets indicates that this definition serves well in identifying values that may require special attention. Section 2C describes the relationship between fourth-spread and standard deviation for a Gaussian distribution. We show below that if the cutoffs are applied to a Gaussian distribution, then .7% of the population is outside the outlier cutoffs; this figure provides a standard of comparison for judging the placement of the outlier cutoffs (cp. p. 40).

Comparison with Classical Methods

The boxplot shows characteristics that derive from the actual data, not from an assumed distributional form. It is helpful to contrast the boxplot with the familiar visual display of the sample mean, $\bar{x}$, plus and minus 1.96 times the sample standard deviation, s. The latter sketch is often used when a batch resembles a random sample from a population believed to be single-humped, perhaps vaguely like a Gaussian distribution:



The interval mentioned would contain roughly 95% of the population if the true mean and standard deviation of the population were $\bar{x}$ and s. When we cannot or do not assume a distributional form for our data, we can use the boxplot to show analogous features of location and width.

We know what a boxplot shows about the data in hand, regardless of their origin. So in the distribution-free setting, we can more easily interpret a boxplot. But we also may wonder what the boxplot represents for a random sample from a Gaussian population. We next explore this question for very large samples.

EXAMPLE: APPLICATION TO A GAUSSIAN POPULATION

Consider the standard Gaussian distribution, with mean 0 and variance 1. We look for population values of this distribution that are analogous to the sample values used in the boxplot.

For a symmetric distribution, the median equals the mean, so the population median of the standard Gaussian distribution is 0. The population fourths are -0.6745 and 0.6745, so the population fourth-spread is 1.349, or about $\frac{4}{3}$. Thus $\frac{3}{2}$ times the fourth-spread is 2.0235 (about 2). The population outlier cutoffs are ± 2.698 (about $2\frac{2}{3}$), and they contain 99.3% of the distribution.

We can gain some further understanding of the values chosen to define the boxplot by considering the population values of the median, fourths, and outlier cutoffs for several familiar distributions. Table 3-2 shows these values, together with the probabilities beyond the outlier cutoffs.

In this table we see that for large samples from extremely short-tailed symmetric distributions, exemplified here by the uniform distribution, all the data values tend to fall within the outlier cutoffs. For the standard Gaussian distribution just discussed for very large samples, we expect only seven-tenths of one percent of the values to be outliers. We choose the t-distributions with various numbers of degrees of freedom to represent symmetric, heavy-tailed distributions. As the tails become heavier, we have a greater probability of observing outliers. Thus we can judge whether our data seem heavier-tailed than Gaussian by how many points fall beyond the outlier cutoffs.

Table 3-2 also includes the chi-squared distributions as examples of asymmetric distributions. These range from the extremely skewed χ^2_1 to the more nearly symmetric χ^2_5 and χ^2_{20}. We find one trait that often occurs in skewed situations (it happened in our example of U.S. cities): the lower outlier cutoff is below the smallest possible data value. Thus the probability of an outlier on this side is 0, and so we get an indication of skewness not only from the relative position of median and fourths, but also from whether all outliers are on one side of the box.

The preceding discussion describes what happens for boxplots of very large samples from some familiar distributions. What happens for smaller samples? The smallest sample size that allows comparison is five, for which the fourths are given by the second smallest and second largest data points. In a sample of size four, the largest data value cannot be outside the outlier cutoffs. This follows from the fact that the upper fourth is the average of the two largest data values and thus involves the extreme value in a direct way.

To examine the question of sampling behavior for small samples, we carried out a simulation study for samples of size five from a standard Gaussian distribution. The results of this experiment suggest that 67% of the

TABLE 3-2. Population values of median, fourths, and outlier cutoffs, and percent outliers for various distributions.

Distribution	M^a	Upper[b] Fourth	Outlier[c] Cutoffs	Total[d] % Out	Value[e] of 1.96σ	% Outside $\mu \pm 1.96\sigma$
			Symmetric			
$U(-1, 1)$	0	0.500	±2.000	none	1.132	none
$N(0, 1)$	0	0.674	±2.698	0.70	1.960	5.00
t_{20}	0	0.687	±2.748	1.24	2.066	5.20
t_{10}	0	0.700	±2.800	1.88	2.191	5.32
t_5	0	0.727	±2.908	3.35	2.530	5.25
t_1	0	1.000	±4.000	15.59	—	—
			Nonsymmetric			
χ^2_1	0.45	0.102	−1.730[f]	7.58	−1.772	5.22
		1.323	3.155		3.772	
χ^2_5	4.35	2.675	−3.252[f]	2.80	−1.198	4.78
		6.626	12.552		11.198	
χ^2_{20}	19.34	15.452	2.888	1.39	7.604	4.53
		23.828	36.392		32.396	

[a] M = median of distribution. Defined so that $F(M) = .5$, where F is the cumulative distribution function.

[b] Upper fourth is the value above which .25 of the probability lies. (Lower fourth has .25 of probability below it.) For the nonsymmetric distributions, the entries in this column are the lower fourth and the upper fourth.

[c] Upper outlier cutoff = upper fourth + $\frac{3}{2}$ × (upper fourth − lower fourth). (Lower outlier cutoff = lower fourth − same quantity.)

[d] % Out = percent of probability below the lower outlier cutoff or above the upper outlier cutoff.

[e] For $U(-1, 1)$, $\sigma = \sqrt{1/3} \approx .58$ and $\mu = 0$. For t_ν, $\sigma = \sqrt{\nu/(\nu - 2)}$ for $\nu > 2$ and $\mu = 0$. For χ^2_ν, $\mu = \nu$ and $\sigma = \sqrt{2\nu}$.

[f] For skewed distributions, one of the pair of cutoffs often falls beyond the range of possible values.

samples had no valu
samples with values l
outlier cutoffs and 9%
Thus we can expect

About a tenth of G
the outlier cutoffs,

About a quarter o
cutoffs, and

About two-thirds o

How should we comp
must examine the det
have parallel structur
values are outliers wit
outliers with probabili
with probability .09.
(Table 3-2) is that .7%
time. (The approximat

The directions and,
Table 3-2 do apply ir
occasionally be helpf
samples.

With Table 3-2 and
tion to the boxplot. W
of batches.

3B. COMPARING I

A display of parallel
batches of data. Fron
among the batches w
cussed.

EXAMPLE: LARGEST C

The 1967 *World Almar*
among these, we selec
populations of these c
summaries, the outlier

samples had no values beyond the outlier cutoffs. For the remaining samples with values beyond the cutoffs, 24% had one value outside the outlier cutoffs and 9% had two values that were outliers, a total of 33%. Thus we can expect

About a tenth of Gaussian samples of five to have both extremes outside the outlier cutoffs,

About a quarter of them to have just *one* extreme outside the outlier cutoffs, and

About two-thirds of them to have no outliers.

How should we compare these results with the large-sample results? We must examine the details carefully because the two sets of results do not have parallel structure. For samples of size five, 0% (none) of the data values are outliers with probability about .67, 20% (one value in five) are outliers with probability about .24, and 40% (two values in five) are outliers with probability .09. For "infinitely large samples," the analogous result (Table 3-2) is that .7% of the data lie beyond the outlier cutoffs *all* of the time. (The approximate formulas on p. 40 help to fill the gap.)

The directions and, probably, something of the amounts of difference in Table 3-2 do apply in a qualitative way to small samples. Table 3-2 can occasionally be helpful, but it is far from the whole story about such samples.

With Table 3-2 and the sampling experiment, we conclude our introduction to the boxplot. We are now ready to consider its use in the comparison of batches.

3B. COMPARING BATCHES USING BOXPLOTS

A display of parallel boxplots can facilitate the comparison of several batches of data. From the display we can see similarities and differences among the batches with respect to each of the five features already discussed.

EXAMPLE: LARGEST CITIES IN 16 COUNTRIES

The 1967 *World Almanac* lists 16 countries that have 10 or more large cities; among these, we selected the 10 most populous cities. Table 3-3 gives the populations of these cities, in 100,000s. Table 3-4 provides the 16 5-number summaries, the outlier cutoffs, and the outliers for these batches.

Understanding Robust and Exploratory Data Analysis

Edited by

DAVID C. HOAGLIN
Harvard University and Abt Associates Inc.

FREDERICK MOSTELLER
Harvard University

JOHN W. TUKEY
Princeton University and Bell Laboratories

John Wiley & Sons, Inc.
New York • Chichester • Brisbane • Toronto • Singapore

Library of Congress Cataloging in Publication Data:
Main entry under title:

Understanding robust and exploratory data anlaysis.

(Wiley series in probability and mathematical statistics. Applied probability and statistics, ISSN 0271-6356)
Bibliography: p.
Includes index.
1. Mathematical statistics. I. Hoaglin, David Caster, 1944– . II. Mosteller, Frederick, 1916– . III. Tukey, John Wilder, 1915– . IV. Series.

QA276.U5 1982 519.5 82-8528
ISBN 0-471-09777-2 AACR2

Printed in the United States of America

10 9 8 7 6 5 4 3 2