OUTLIERS

1. Introduction

In a data set, an outlier is a point far from the bulk of the data. Rice has a discussion of outliers on pp. 393–395. Some attempts have been made to give a precise definition of outlier, but Rice doesn't and we won't in this course. Sometimes outliers are produced by gross errors of some kind. In economic data there can be outliers not necessarily due to errors. For example, a billionaire's wealth is an outlier compared to the wealths of individuals even moderately high on the wealth scale.

2. An example: heat of sublimation of platinum

Fig. 10.10 on p. 394 of Rice shows 26 measurements of the heat of sublimation of platinum. Of the measurements, 21 are less than 137 (in fact ≤ 136.6, the second measurement). All measurements are ≥ 133.7 (the 20th measurement), so the bulk of the data are between 133.7 and 136.6. The other 5 are ≥ 141.2 (the 15th measurement) and range up to 148.8 (the 10th measurement). These 5 measurements are examples of outliers. There are no outliers in the first 7 measurements or the last 11, so the outliers are somewhat concentrated in the horizontal direction, with the largest two coming on the 9th and 10th measurements and the 4th-largest on the 8th. So it seems something was amiss during part of the experiment; Rice mentions possible "equipment malfunctions."

Rice notes that the sample mean of all 26 of the measured values is 137.05 (checked in R), which is larger than all values other than the outliers, recalling that the largest non-outlier is 136.6. Excluding the 5 outliers, the sample mean is smaller (135.03 from R). So the outliers make a difference of about 2.0 in the sample mean which in this case is large enough to matter. (The authors of the study Rice quotes actually excluded the largest 7 measurements and got 134.9, but the rationale for excluding the additional two largest measurements is not clear.)

The sample median of all 26 values is 135.1, which is well within the bulk of the data. The sample median of the non-outliers is 134.9, smaller by just 0.2, so we see that the sample mean is much more influenced by outliers than the sample mean is. On p. 397, Rice finds,

---

Date: 18.650, Nov. 13, 2015.

assuming that the 26 observations are i.i.d. from some continuous distribution, that a 97% confidence interval for the true median m of the distribution is $[X_{(8)}, X_{(19)}] = [134.8, 135.8]$. This is true because the probability that $m < X_{(8)}$ is the probability of 7 or less successes in $n = 26$ independent trials with probability $p = 1/2$ of success on each trial ("success" meaning in this case $X_j \leq m$), and $B(7, 26, 1/2) = \text{pbinom}(7, 26, 1/2) = .01447964$ (from R) $\doteq .0145$ as Rice gives (Example A, p. 396). The probability that $m > X_{(19)}$ is the same by symmetry. So the probability that m is outside the confidence interval is $2(.0145) = 0.029 < .03$ as stated. Both sample medians, 135.1 for the full data and 134.9 for the non-outliers, are within the confidence interval for the true median. As Rice says, the independence assumption is questionable (with e.g. the consecutive 9th and 10th measurements being largest).

The main point is that considering medians (true or sample) gives values within the bulk of the data (not at all surprisingly for the sample median of non-outliers), which the sample mean isn't necessarily, as in this example.

3. The effect of outliers on sample variances

We already saw that outliers, even not terribly extreme ones, can influence sample means. Outliers affect sample variances even more, because if an observation is an outlier, being large in absolute value, its square will be much larger still.

For the heat of sublimation of platinum data, the sample variance of the full data set of 26 points is 19.79, giving a sample standard deviation of 4.49, and an approximate 95% confidence interval for the true mean, based on the $t(25)$ distribution, of width about 3.63, which is much wider than the width about 1.0 for the exact 97% confidence interval for the true median, also based on the full data. About the same width was estimated in a 1970 study from other data (see Experimental Results, below). The excessive width of 3.63 shows the bad effect of contributions of outliers to the sample variance. For the 21 non-outliers, the sample variance is 0.487, leading to an approximate 95% confidence interval based on the $t(20)$ distribution with width about 0.651 which is if anything too small. This shows that simply discarding points that appear to be outliers is not necessarily a good way to proceed, as it may lead to underestimating the amount of variability in the true distribution.

3.1. **A heavy-tailed distribution: the Cauchy distribution.**

The standard Cauchy distribution is the distribution on the real line with

density function

$$f(x) = \frac{1}{\pi(1+x^2)}, \qquad -\infty < x < +\infty.$$

This is a t distribution with 1 degree of freedom. From Rice's t table one can see that $t_{.975}(1) = 12.706$, $t_{.99}(1) = 31.821$, and $t_{.995}(1) = 63.657$. For a $N(0,1)$ distribution the corresponding quantiles are 1.960, 2.326 and 2.576, much smaller. The Cauchy distribution has a tendency to produce values which, as compared with N(0,1) distributed data, are outliers.

3.1.1. *Experiments with Cauchy-distributed variables.* I did some experiments in R, generating 20 i.i.d. Cauchy variables with the command x = rcauchy(20), then finding their sample variance var(x). I soon found a sample with variance of 521.7, although in 7 tries, I also found one variance as small as 3.26. In the sample with variance 521.7, the largest X_j in absolute value was -85.95, the second-largest, 49.6, third-largest, -10.42, and fourth-largest, 4.75. So we can say that -85.95 and 49.6 are outliers, and one can easily see how they would contribute to a large sample variance.

If X has a standard Cauchy distribution, we have $\Pr(X \le -85.95) \doteq 0.00370$ and so $\Pr(|X| \ge 85.95) \doteq 0.00740$. Considering that we had a sample of size 20, by a Bonferroni correction, multiplying by 20, one gets 0.148 which is not significantly small.

The sample medians of the seven Cauchy samples were much better behaved than the sample variances. The largest in absolute value was -1.78, next-largest -0.33.

3.2. **Experimental results.** Rice's data came from a paper by Hampson and Walker (1961). Later came a paper by Plante, Sessoms and Fitch (1970) estimating the heat of sublimation of platinum as 134.92 ± 0.5 kcal/mol (kilocalories per mole). That interval has the same length 1 kcal/mol as for the confidence interval for the median from the 1961 data, and the two intervals have an overlap of width about 0.6. The 1970 interval is shifted to the left by about 0.4 compared to the one from the 1961 data.

REFERENCE

Plante, E. R., Sessoms, A. B., and Fitch, K. R. (1970). Vapor pressure and heat of sublimation of platinum. *J. Research Nat. Bureau of Standards* **74A**.