and half of the time spent on CRs fall into the two shortest-duration bins, indicating that maintenance activities are dominated by small changes measured either by number or total time.

ESTIMATION ACCURACY OF MAINTENANCE DURATION

Because company personnel did not change during the period the maintenance records cover, we might expect them to have become progressively more accurate in assessing the time taken to perform a CR. Estimation accuracy of maintenance duration can be determined by looking for any change in bias in estimating duration and in the variance in estimating duration.

Figure 3 shows the difference between estimated and actual CR durations for the full set of records in increasing chronological order. Points above the zero line correspond to CRs performed more quickly than expected. Note that the frequency with which CRs were resolved did not change significantly throughout the measurement period.

Changes in bias

Figure 3 clearly indicates a systematic pessimistic bias. We can quantify this using the data in Figure 2, which reveals that the average duration of a CR was 5.17 hours and the average bias 1.83 hours: On average, engineers overestimated across all CRs by around 35 percent. What is not clear, however, is whether this behavior changed over time as the developers gained more experience with these products.

Splitting the data into two halves reveals that the average bias in the first half is 2.45 hours and the average bias in the second half is 1.2 hours. We can therefore hypothesize that the average bias in each half either came from the same population (null hypothesis) or did not come from the same population.

We can analyze the data using the z-test for the difference of means in a population,[12] which states that the following statistic is approximately distributed as $N(0,1)$:

$$z = \frac{\overline{X_1} - \overline{X_2}}{\left(\frac{(s_1)^2}{N_1} + \frac{(s_2)^2}{N_2} \right)^{\frac{1}{2}}}$$

where $\overline{X}_i$, s_i, and N_i are respectively the sample means, standard deviations, and number of samples for each half sample.

Substituting the appropriate values yields

$$z = \frac{2.45 - 1.20}{\left(\frac{44.95}{477} + \frac{29.78}{478} \right)^{\frac{1}{2}}} \simeq 3.16$$

This corresponds to a probability of less than 0.001 that the change in bias occurred by chance. Thus, the null hypothesis can be comprehensively rejected, and it must be concluded that there is a highly significant drop in the average bias in the two half samples.

Repeating this analysis solely on the product that consumed most of the company's maintenance effort also yields a highly significant result of $z = 2.45$, which corresponds to a probability of only around 0.003 that the change in bias occurred by chance.

Spread of estimates

The variances in the two half-samples can be compared using the F-test:[12]

$$F = \frac{S_1^2}{S_2^2} \simeq 1.51$$

The number of degrees of freedom for the numerator and denominator is 476 and 477, respectively. At the 5 percent significance level, the corresponding value of the F statistic is less than 1.25. It can therefore be deduced that the drop in variance is significant at the 5 percent level.

Relationship with CR duration

The distribution of CR durations over the maintenance period appears random, so it is useful to see if there is any relationship between estimation accuracy and the actual duration of the CR—for example, it might reasonably be expected that the developers predicted shorter-duration CRs more accurately.

Figure 4 shows a clear anticorrelation: In general, the developers pessimistically pre-

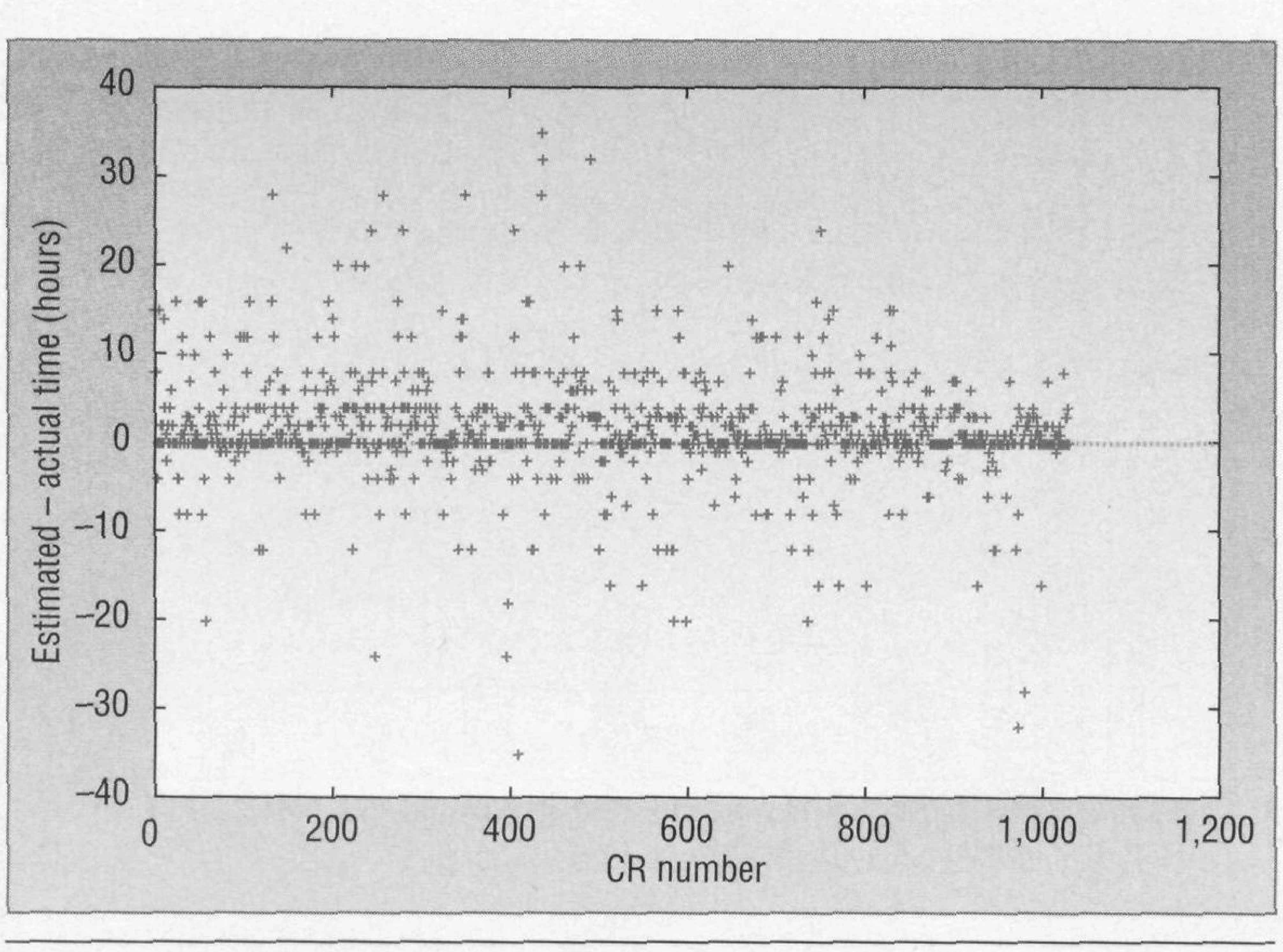

Figure 3. Estimated minus actual CR duration in chronological order. The frequency with which CRs were resolved did not change significantly throughout the measurement period.