

# The median is the message: Wilson and Hilferty's reanalysis of C.S. Peirce's experiments on the law of errors

---

Roger Koenker

The Institute for Fiscal Studies  
Department of Economics, UCL

**cemmap** working paper CWP28/08

**THE MEDIAN IS THE MESSAGE:  
WILSON AND HILFERTY'S REANALYSIS OF C.S. PEIRCE'S  
EXPERIMENTS ON THE LAW OF ERRORS**

ROGER KOENKER

ABSTRACT. Data is reanalyzed from an important series of 19th century experiments conducted by C. S. Peirce and designed to study the plausibility of the Gaussian law of errors for astronomical observations. Contrary to the findings of Peirce, but in accordance with subsequent analysis by Fréchet and Wilson and Hilferty, we find normality implausible and medians an attractive alternative to means for the analysis.

One of the first necessities of the discussion of any statistical problem is to study the data themselves.

E.B. Wilson (1923)

1. INTRODUCTION

In the summer of 1872 C.S Peirce conducted a series of experiments designed to investigate the validity of the Gaussian law of errors, and thus the applicability of the method of least squares, for observational studies such as those then commonly employed in astronomy. The conclusion Peirce drew from these experiments was that faith in the “normal” law and the method of least squares was fully justified. However, subsequent analysis by Fréchet (1924) and by Wilson and Hilferty (1929) found that the normal approximation

---

Version: September 27, 2008. This research was partially supported by NSF grant SES-05-44673. The author would like to thank Marc Hallin, Steve Portnoy and Steve Stigler for helpful comments at various stages of this inquiry. Comments by an Associate Editor were also extraordinarily helpful. All data and software required to reproduce the results reported here will be made available on the web via <http://www.econ.uiuc.edu/~roger/research/frechet/frechet.html>.

was quite unsatisfactory, concluding that medians offered a preferable mode of analysis to means for Peirce’s data. The primary objective of the present note is to evaluate the methods and conclusions of Wilson and Hilferty in the light of subsequent statistical developments.

Charles Saunders Peirce (as in *purse*) (1839-1914) was a preeminent American mathematician, logician and philosopher of the 19th century. In addition to his fundamental contributions to set theory and relational logic, he is generally regarded as the progenitor of philosophical pragmatism and the field of semiotics. Stigler (1992) calls Peirce “one of the two greatest American scientific minds of [the 19th] century (the other being J. Willard Gibbs.)” Early in his prolific career Peirce was employed by the U.S. Coast and Geodetic Survey. In this capacity he undertook a series of experiments intended to investigate the “distribution of errors in observations of a phenomenon,” as Peirce expresses it, a phenomenon “not seen coming on, as in the case of a transit, but sudden, as in the case of the emersion of a star from behind the moon.” To this end, he hired a young man who had no prior experience in scientific observation whose task was to respond to a “signal consisting of a sharp sound like a rap” by depressing a telegraph operator’s key “nicely adjusted.” Response times were recorded in milliseconds with the aid of a very sophisticated instrument called a Hipp chronoscope. For 24 days in July and early August of 1872 roughly 500 such responses were collected for each day. Data was recorded as illustrated in Table 1, and published as an integral part of Peirce (1873).

## 2. WILSON AND HILFERTY’S ANALYSIS

The analysis of Wilson and Hilferty offers a revealing glimpse into an earlier era of statistical computation. Table 2 constitutes my best attempt to reproduce their main table. An encouraging feature of this exercise is that the column of means matches the original table exactly, and the medians also match to three significant digits, except for

SIXTH DAY, JULY 10, 1872

66	1	117	0	137	2	157	5	177	4	197	3	217	1	237	1	257	0
72	1	8	1	8	0	8	6	8	3	8	3	8	3	8	1	8	0
75	1	9	1	9	5	9	7	9	7	9	1	9	2	9	2	9	1
87	2	120	1	140	5	160	7	180	3	200	5	220	3	240	0	260	1
88	1	1	1	1	3	1	7	1	4	1	1	1	1	1	0	1	0
101	2	2	3	2	6	2	3	2	11	2	8	2	1	2	1	2	0
2	0	3	2	3	3	3	10	3	9	3	2	3	1	3	0	3	0
3	0	4	2	4	4	4	6	4	7	4	4	4	1	4	3	4	1
4	1	5	1	5	1	5	12	5	6	5	0	5	2	5	1	272	1
5	1	6	0	6	6	6	2	6	8	6	2	6	1	6	1	277	1
6	1	7	0	7	8	7	4	7	9	7	1	7	0	7	0	280	1
7	1	8	1	8	3	8	5	8	2	8	2	8	1	8	0	285	1
8	1	9	2	9	4	9	6	9	7	9	1	9	3	9	0	287	2
9	2	130	1	150	5	170	9	190	7	210	4	230	1	250	0	290	1
110	0	1	4	1	4	1	5	1	6	1	3	1	0	1	0	316	1
1	1	2	2	2	7	2	9	2	7	2	3	2	0	2	0	327	1
2	2	3	0	3	4	3	5	3	5	3	1	3	1	3	0	367	1
3	2	4	5	4	7	4	5	4	6	4	4	4	0	4	0	376	1
4	0	5	4	5	4	5	5	5	2	5	3	5	0	5	0	392	1
5	1	136	1	156	5	176	7	196	7	216	3	236	0	256	0	411	1
116	3																

TABLE 1. Raw Data for Day 6 of the Experiment: Odd columns of the table give times in milliseconds and associated even columns report cell counts of the number of occurrences of the indicated timing. Source: Peirce(1873)

day 15, in which there appears to be a transposition of two digits. A minor mystery of the Wilson-Hilferty medians is that they are reported to four significant digits despite the fact that the original data all have unique 3-digit medians. The standard deviation of the means also agree quite well with the original table except for a few entries that could be attributed to differences in rounding conventions. I defer discussion of the entries for the “standard deviation” of the median to the next section. Since there appears to have been no generally agreed upon method for assessing the accuracy of the median at the time, these entries are one of the more intriguing aspects of the analysis.

The scale estimates in Columns 3-6 of the table are quite consistent with Wilson and Hilferty’s table. Under normality one would expect that the ratio of half the interquartile range to the scaled standard deviation would be unity, however we find instead that these

ratios are consistently smaller than one. Similarly, the number of observations lying beyond  $\pm 3.1\hat{\sigma}$  bounds is excessive, as is the number of observations lying *inside* the  $\pm 0.25\hat{\sigma}$  intervals. Thus, as emphasized by Wilson and Hilferty, Peirce's observations are both more peaked near the median, and heavier tailed than one would expect from the normal model.

Day	n	(1)	(2)	(3)	(4)	(5)	(6)	(7)			(8)		(9)
		median	mean	IQR/2	$\kappa\sigma$	ratio	mad	N	P	T	Oin	Ein	Ex
1	495	468 ± 2.5	475.6 ± 4.1	58.0	62.1	0.934	70.1	1	3	4	111	98	13
2	490	237 ± 2.1	241.5 ± 2.1	26.5	31.6	0.838	35.6	0	0	0	115	97	19
3	489	200 ± 1.7	203.2 ± 2.1	27.0	30.6	0.883	33.8	0	7	7	118	97	22
4	499	201 ± 1.2	205.6 ± 1.8	19.5	26.6	0.734	26.9	1	7	8	141	99	42
5	490	147 ± 2.0	148.5 ± 1.6	21.5	23.5	0.916	26.7	0	4	4	112	97	15
6	489	172 ± 1.9	175.6 ± 1.8	19.5	26.9	0.724	27.9	0	6	6	116	97	20
7	496	184 ± 1.7	186.9 ± 2.2	24.5	32.9	0.745	32.3	0	6	6	137	98	40
8	490	194 ± 1.3	194.1 ± 1.4	17.5	20.5	0.853	22.0	2	4	6	115	97	19
9	495	195 ± 1.5	195.8 ± 1.6	18.0	23.8	0.758	24.0	2	4	6	136	98	39
10	498	215 ± 1.6	215.5 ± 1.3	16.0	19.0	0.842	21.0	2	1	3	113	98	15
11	499	213 ± 2.1	216.6 ± 1.7	19.8	25.2	0.783	25.8	1	5	6	130	99	31
12	396	233 ± 1.8	235.6 ± 1.7	17.0	22.4	0.757	23.4	3	5	8	102	78	31
13	489	244 ± 1.3	244.5 ± 1.2	16.5	17.7	0.930	19.9	6	1	7	109	97	12
14	500	236 ± 1.3	236.7 ± 1.9	14.6	27.9	0.524	21.6	2	3	5	193	99	95
15	498	235 ± 1.1	236.0 ± 1.5	14.0	22.4	0.625	20.8	4	4	8	166	98	69
16	498	233 ± 1.6	233.2 ± 1.7	15.5	25.5	0.609	21.7	4	2	6	167	98	70
17	507	264 ± 1.8	265.5 ± 1.7	20.5	26.0	0.788	27.6	3	5	8	128	100	28
18	495	254 ± 1.3	253.0 ± 1.1	15.8	16.5	0.954	18.7	0	4	4	121	98	23
19	500	255 ± 0.9	258.7 ± 2.0	14.0	30.6	0.458	20.4	0	3	3	208	99	110
20	494	253 ± 1.4	255.4 ± 2.0	15.4	29.3	0.525	21.7	0	3	3	181	98	85
21	502	245 ± 1.7	245.0 ± 1.2	14.5	18.3	0.790	19.2	3	4	7	114	99	15
22	499	255 ± 1.6	255.6 ± 1.4	14.8	21.6	0.683	19.2	1	4	5	145	99	46
23	498	252 ± 1.2	251.4 ± 1.4	14.9	21.5	0.691	19.7	1	3	4	140	98	43
24	497	244 ± 0.9	243.4 ± 1.1	11.5	15.9	0.723	16.7	3	3	6	117	98	19

TABLE 2. Summary Statistics for the Peirce (1872) Experiments: This table attempts to reproduce a portion of the Wilson and Hilferty (1929) analysis of the Peirce experiments. Column numbers correspond to the numbering used by Wilson and Hilferty. Following the columns with the daily sample sizes, medians and means and their associated standard errors, there are four scale statistics: half the interquartile range,  $0.6745\sigma$ , the ratio  $IQR/(2\kappa\sigma)$  and the mean absolute deviation. The next three columns contain the number of observations falling outside a  $3.1\sigma$  cutoff: numbers below, above and total. The final three columns contain the number of observations within a  $0.25\sigma$  cutoff, the expected number of such observations under the normal model, and the excess number, respectively.

Wilson and Hilferty provide a second panel of their main table containing estimates of several different measures of skewness of the observations. Unfortunately, these computations are quite inaccurate, so it seemed pointless to attempt to replicate them in detail. Instead, Table 3 reports the conventional estimates of skewness  $\gamma_1 = \mu_3/\sigma^3$  and kurtosis  $\gamma_2 = (\mu_4/\sigma^4) - 3$ . In addition, the D'Agostino (1970) and Anscombe and Glynn (1983) tests for normality based on these estimates are reported. Both of these test statistics are approximately normal under the hypothesis of normality, so neither test supports the plausibility of normality for the Peirce data.

Further visual evidence on the plausibility of normality for the Peirce data can be seen in Figure 1 where we give QQ plots of the daily samples against Gaussian quantiles. With the exception of two or three days, July 5 and 9; none of these plots supports Peirce's view that the data is Gaussian. Of course, all of the plots support the well-known Tukey maxim: All distributions are normal in the middle. But both abnormal skewness and kurtosis are clearly evident for most of the 24 days.

Why should we care whether the response times reported in the Peirce experiments are Gaussian? There are several perspectives from which to view this question. From a pragmatic perspective, we would like to know whether Peirce's original objective—to justify the use of least squares methods—is supported by the evidence of the experiments. Since Laplace's work in 1770's it had been recognized that the sample mean enjoyed a certain optimality under the so-called Gaussian law of errors, or what is sometimes called Laplace's *second* law of error. In contrast, Laplace's *first* law of error, the double exponential distribution, could be similarly invoked to justify the optimality of the sample median. We will explore in the next section the assertion of both Fréchet and Wilson and Hilferty that for the Peirce data medians offer a superior vehicle of analysis.

From a deeper ontological perspective we might be tempted to ask: Is Gaussian random sampling *ever* a plausible model for scientific observation? Wilson and Hilferty (1929)

Day	Skewness	D'Agostino	Kurtosis	Anscombe
1	0.881	4.663	6.113	6.300
2	0.437	2.535	3.870	3.012
3	1.082	5.423	6.775	6.828
4	1.843	7.803	13.234	9.722
5	0.387	2.260	4.401	4.096
6	1.472	6.715	9.430	8.365
7	2.880	9.906	27.392	11.814
8	0.481	2.764	7.068	7.051
9	1.673	7.328	16.633	10.435
10	0.516	2.970	8.728	8.095
11	1.634	7.251	12.722	9.584
12	0.634	3.192	7.511	6.751
13	-0.214	-1.279	5.522	5.667
14	5.693	13.340	66.242	13.688
15	1.518	6.911	17.542	10.622
16	5.851	13.452	93.550	14.233
17	0.244	1.479	7.257	7.284
18	0.257	1.536	4.746	4.681
19	8.162	15.161	99.494	14.349
20	7.685	14.771	93.784	14.192
21	0.217	1.315	11.556	9.259
22	5.305	12.972	71.157	13.801
23	4.293	11.899	57.699	13.419
24	0.064	0.386	7.931	7.654

TABLE 3. Skewness and Kurtosis Estimates for Peirce Experiment: Columns one and three of the table report daily estimates of the skewness and kurtosis coefficients,  $\gamma_1$  and  $\gamma_2$ . Columns two and four report the corresponding D'Agostino (1970) and Anscombe-Glynn (1983) test statistics for normality based upon the respective coefficients. Both test statistics are approximately standard Gaussian under the null hypothesis, so at conventional significance levels Gaussian skewness is rejected in 19 of the 24 days, while Gaussian kurtosis is rejected for all 24 days.

cite Poincaré's famous dictum that everybody believes in the Gaussian law of errors: the mathematicians because they think it has been empirically demonstrated by experimenters, and the experimenters because they think the mathematicians have proven it *a priori*. What accounts for such an extraordinary suspension of skepticism? No one would be so surprised to learn from Wilson (1923) that Laplace's first law sometimes fits *economic* observations better than his second law. Wilson's source for this example, Crum (1923) also argues for the advantages of medians over means for seasonal adjustment of his interest

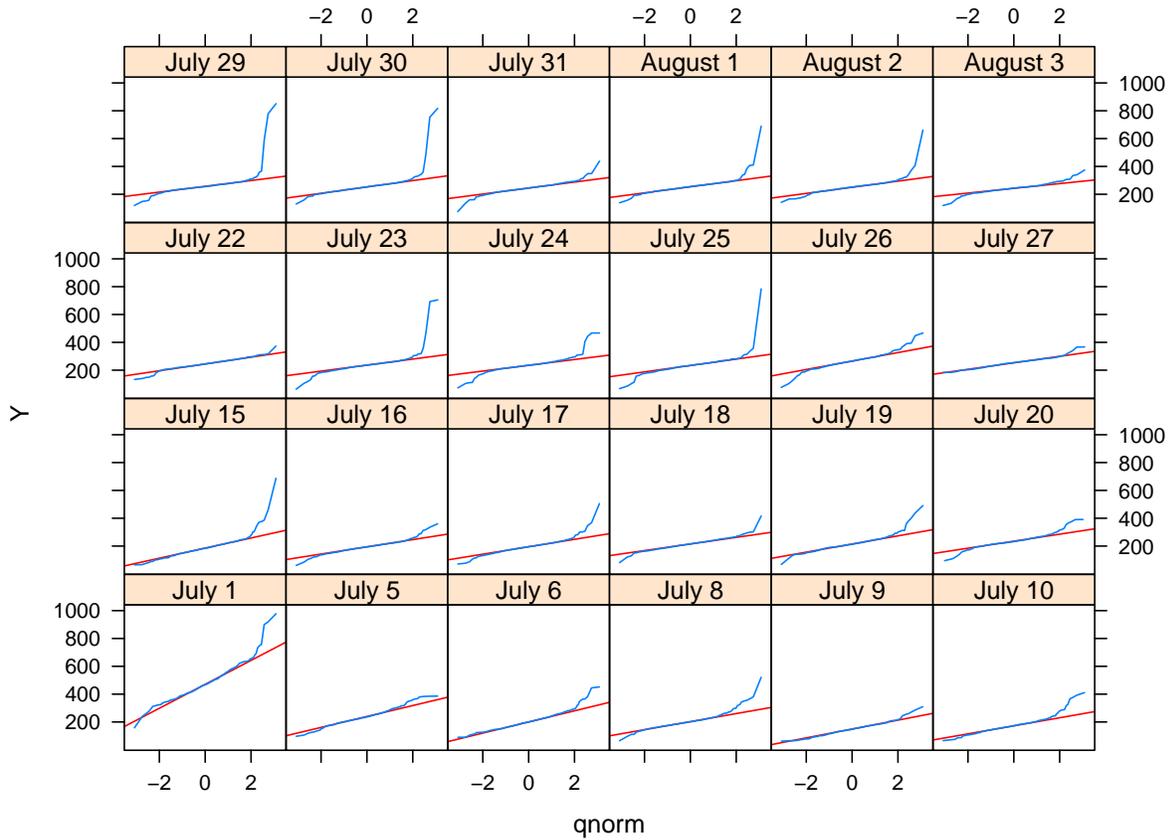


FIGURE 1. Normal QQ Plots for each of the 24 days of the Peirce Experiments:

rate series. Crum fits a contaminated normal mixture model obtaining a contamination proportion  $1/4$  and relative scale of  $4.8$ . Following Yule (1917), Crum then solves a quartic equation to find that for this contamination proportion the median is more efficient than the mean if relative scale exceeds  $2.6$ . But Peirce’s experiments seem to be the very model of a reproducible scientific experiment; if we do not see Gaussian behavior here, then where could we expect to find it? In fairness, we should recall that Stigler (1977) considers several historical instances of estimating a well-defined location parameter—all cases in which the refinement of modern measurement provides a considerably more accurate estimate of

the quantity in question—and concludes that while some modest amount of trimming is preferable to the untrimmed sample mean, the sample median is “too inefficient.”

### 3. THE STANDARD DEVIATION OF THE MEDIAN

The most puzzling aspect of the Wilson and Hilferty (1929) analysis for me involves their reported “standard deviations of the medians.” These values are crucial to their argument:

A comparison of the standard deviations of the median and mean in columns (1) and (2) shows that for these observations the median is better determined than the mean on 13 days, worse determined on 9 days, and equally well determined on 2 days. Roughly speaking this means that mean and median are on the whole about equally well determined.

But what *is* the standard deviation of the median? Even today there is more than a little ambiguity about the phrase, how are we to interpret it in 1929? One possibility, suggested by the approach taken in Wilson (1927), is that Wilson and Hilferty adopted Laplace’s first law of error wholeheartedly and used the mean absolute deviation, the maximum likelihood estimator of the Laplace scale parameter, to estimate the density and thereby the standard deviation. This approach yields the standard deviation estimates in the “Laplace” column of Table 4 and we see that they correspond rather poorly to the estimates reported by Wilson and Hilferty. The Laplacian estimates are substantially smaller for most of the days with a mean of only 1.155 versus the mean of 1.538 for the Wilson and Hilferty estimates. So, if not the leap of Laplacian faith, what else?

Yule (1917) seems to be the first to systematically treat the problem of estimating the precision of the sample median. Yule provides two suggestions: the first is to simply assume normality and inflate the usual estimate of the standard deviation of the mean by the infamous factor, known already to Laplace,  $\sqrt{\pi/2} \approx 1.253$ . This suggestion is clearly not very helpful in the present instance, since it begs the question of whether

Day	WH	Laplace	Yule	Siddiqui	Exact I	Exact II	Jeffreys	Boot
1	3.0	3.1	2.5	3.3	3.8	3.8	3.8	3.6
2	2.4	1.6	2.1	2.6	2.3	2.1	2.3	2.4
3	1.8	1.5	1.7	1.5	2.3	2.2	2.6	2.1
4	1.2	1.2	1.2	1.3	1.3	1.3	1.3	1.2
5	2.1	1.2	2.0	2.3	2.0	2.0	2.0	2.1
6	1.8	1.3	1.9	1.8	1.5	1.5	1.8	1.9
7	1.6	1.4	1.7	1.8	1.3	1.3	1.3	1.5
8	1.2	1.0	1.3	0.9	1.5	1.3	1.5	1.3
9	1.4	1.1	1.5	1.3	1.3	1.3	1.5	1.4
10	1.5	0.9	1.6	1.3	1.5	1.5	1.5	1.5
11	1.8	1.1	2.1	1.8	1.8	1.8	1.8	1.7
12	1.9	1.2	1.8	2.0	1.8	1.8	1.8	1.7
13	1.3	0.9	1.3	1.4	1.3	1.3	1.3	1.3
14	1.2	1.0	1.3	0.9	1.3	1.3	1.3	1.2
15	1.0	0.9	1.1	0.9	1.0	1.0	1.0	1.2
16	1.4	1.0	1.6	1.3	1.3	1.3	1.3	1.4
17	1.7	1.2	1.8	1.8	1.8	1.8	1.8	1.8
18	1.3	0.8	1.3	1.3	1.0	1.0	1.0	1.2
19	0.9	0.9	0.9	0.9	1.0	1.0	1.0	1.0
20	1.3	1.0	1.4	1.3	1.5	1.3	1.5	1.3
21	1.6	0.9	1.7	1.8	1.5	1.4	1.3	1.5
22	1.5	0.9	1.6	1.8	1.3	1.3	1.3	1.5
23	1.1	0.9	1.2	0.9	1.3	1.1	1.3	1.2
24	0.9	0.8	0.9	0.9	1.0	1.0	1.0	0.9
Mean	1.538	1.155	1.567	1.549	1.573	1.531	1.594	1.584
MAE	0.000	0.393	0.129	0.135	0.180	0.166	0.191	0.103
MSE	0.000	0.219	0.027	0.029	0.064	0.056	0.079	0.025
MXE	0.000	0.896	0.457	0.306	0.827	0.777	0.827	0.553

TABLE 4. Standard Deviations for the Medians: The table reports Wilson and Hilferty’s estimates of the standard deviation of the median and seven attempts to reproduce their estimates as described in the text. Column means and three measures of the discrepancy between the original estimates and the new ones are given: mean absolute error, mean squared error, and maximal absolute error.

the median is more precisely estimated than the mean. Yule’s second suggestion is to estimate the asymptotic standard deviation of the sample median,  $\omega = 1/(2f_0\sqrt{n})$ , where  $f_0$  denotes the density function of the observations evaluated at the median. He offers one numerical example of how to compute this estimate, an example that is reproduced in all his subsequent editions up to and including those coauthored with Maurice Kendall in the

1930's and 1940's. Based on the heights of 8585 adult males in the United Kingdom, Yule's estimate of  $f_0$  took the frequency of subjects in the cell containing the median, 1329 men 67 inches tall, and divided by the sample size. This can be interpreted as an estimate,

$$\hat{f}(x) = \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h}$$

and happens to simplify by taking  $h = 1/2$ , because the cell boundaries were defined in inches. The appeal of this calculation to Yule was—one can't resist speculating—that the result yielded an estimate of 0.0349, essentially identical to that obtained by his normal theory approach, 0.0348. In the present circumstance it is difficult to judge *a priori* how to choose the bandwidth  $h$  for an analogous calculation for the Peirce data, so in the spirit of “reverse engineering,” we tried a grid of values seeking one that most closely reproduces Wilson and Hilferty's results. Unfortunately, although we can come close, we are unable, with a fixed bandwidth, to match their results to the two significant figures they report. The best we can do is reported in the “Yule” column of Table 4. This strategy yields  $h = 1.6$ , which is quite small by modern standards. The textbook of Kelly (1923) describes a procedure essentially similar to Siddiqui's and illustrates its use on a sample of 62 daily temperature observations, considering bandwidths ranging from 0.065 to 0.25.

Equally simple and appealing is the suggestion of Siddiqui (1960) to base the estimate of the standard deviation of the median on an estimate of the reciprocal of the density, or sparsity function,

$$\hat{s}(\tau) = \frac{\hat{F}_n^{-1}(\tau+h) - \hat{F}_n^{-1}(\tau-h)}{2h}.$$

Now,  $\hat{\omega} = \hat{s}(.5)/(2\sqrt{n})$ , and we can again optimize over the bandwidth to find the best fit to the Wilson and Hilferty estimates. This time  $h = 0.025$ , which is again quite small.

A third option that may have been open to Wilson and Hilferty involves estimating the standard deviations from the width of direct confidence intervals based on the binomial

theory of the order statistics. Lehmann (1959) attributes this technique to Thompson (1936), but it is not impossible that something similar occurred to Wilson. Two versions of this “exact” method are explored here: the first involves constructing the conventional “conservative” interval and dividing its length by the factor 3.92, the second involves interpolating the limits of the conservative interval and the interval formed by its adjacent order statistics and again rescaling the length to obtain a standard deviation estimate. Both of the foregoing estimates are reported in Table 4 together with their means over the 24 days, and their deviation from the Wilson and Hilferty estimates measured by mean squared error. Another variant of the exact methods, suggested by Jeffreys (1939), is to use the normal approximation to the binomial. Results for this option are reported as well.

For the sake of further comparison, we have also considered bootstrap estimates of the standard deviation, even though this is certainly not a plausible candidate for Wilson and Hilferty’s method. For this purpose we have done 500 bootstrap replications. Ironically, this method comes closest to the results of Wilson and Hilferty in the sense of mean squared error.

All of the methods investigated, with the possible exception of the Laplace method, yield standard deviations quite close to those reported by Wilson and Hilferty, and all support their conclusion that the mean and median are about equally well determined. Unfortunately, however, none of the methods successfully “reproduce” the Wilson and Hilferty results, so the question remains open: How did they do it?

#### 4. QUOTIDIAN FLUCTUATIONS

A striking feature of the Peirce experiments is the degree of daily variation when compared to the intra-day variability we have just considered. This is illustrated graphically in Figure 2 where we plot daily means and medians with associated pointwise error bands. The initial drop in average response times over the first five days can be attributed to

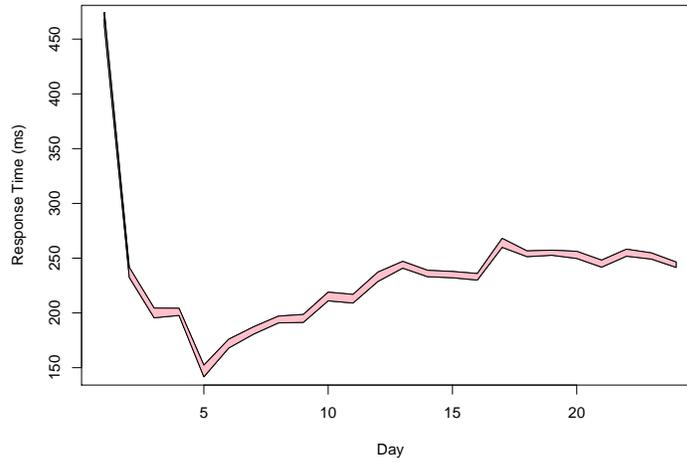


FIGURE 2. Daily Medians and Means with Confidence Band

“learning by doing,” but the gradual increase and possible leveling off is more difficult to explain. Perhaps, after becoming proficient, some element of *ennui* sets in. In any case, the daily variability is so large that it swamps the intra-day variation and the confidence bands for the means and medians are indistinguishable.

Peirce concludes from this that “transit observers be kept in constant training by means of some observations of an artificial event which can be repeated with rapidity, ... as it is the general condition of the nerves which it is important to keep in training more than anything peculiar to this or that kind of observation.” This conclusion seems to rest on the fact that the average response times tend to level off in the last few days of the experiment, and the standard deviations are also lowest in this period. Stigler (1992) provides a comprehensive review of Pierce’s substantial influence on subsequent developments in experimental psychology.

## 5. CONCLUSION

Peirce's experiments and their subsequent analysis constitute an object lesson in the sophistication of early data collection and analysis. Given the computational difficulties faced by Wilson and Hilferty it is remarkable how much of their analysis *was* reproducible. Their work sets an enviable standard for contemporary computational research. The primary conclusion drawn by their analysis—that medians were competitive with means for the very carefully designed experiments of Peirce—is entirely vindicated.

Such findings naturally raise new questions about extending median methods to more complex statistical settings. There is a long tradition of such inquiry beginning in the 18th century with the “method of situation” of Boscovich and Laplace, and continuing with Edgeworth's (1988) “plural median” estimator for regression. More recent developments like Tukey's median polish for two-way anova and quantile regression have addressed some of these questions. But it is still not uncommon to encounter the attitude: “Sure, median methods work well in practice, but how do they work in theory?” The struggle continues.

## REFERENCES

- ANSCOMBE, F., AND W. GLYNN (1983): “Distribution of kurtosis statistic for normal statistics,” *Biometrika*, 70, 227–234.
- CRUM, W. L. (1923): “The Use of the Median in Determining Seasonal Variation,” *Journal of the American Statistical Association*, 18, 607–614.
- D'AGOSTINO, R. (1970): “Transformation to Normality of the Null Distribution of  $G_1$ ,” *Biometrika*, 57, 679–681.
- FRÉCHET, M. (1924): “Sur la loi des erreurs d'observation,” *Matematicheskii Sbornik*, 32, 5–8.
- JEFFREYS, H. (1939): *Theory of Probability*. Oxford.
- KELLY, T. L. (1923): *Statistical Method*. MacMillan: New York.
- LEHMANN, E. L. (1959): *Testing Statistical Hypotheses*. Wiley: New York.

- PEIRCE, C. S. (1873): "On the Theory of Errors of Observation," *Report of the Superintendent of the U.S. Coast Survey*, pp. 200–224., Reprinted in *The New Elements of Mathematics*, (1976) collected papers of C.S. Peirce, ed. by C. Eisele, Humanities Press: Atlantic Highlands, N.J., vol. 3, part 1, 639–676.
- SIDDIQUI, M. (1960): "Distribution of Quantiles from a Bivariate Population," *Journal of Research of the National Bureau of Standards*, 64, 145–150.
- STIGLER, S. M. (1977): "Do Robust Estimators Work with Real Data?," *The Annals of Statistics*, 5, 1055–1077.
- STIGLER, S. M. (1992): "A Historical View of Statistical Concepts in Psychology and Educational Research," *American Journal of Education*, 101, 60–70.
- THOMPSON, W. R. (1936): "On Confidence Ranges for the Median and Other Expectation Distributions for Populations of Unknown Form," *The Annals of Statistics*, 7, 122–128.
- WILSON, E. B. (1923): "First and Second Laws of Errors," *Journal of the American Statistical Association*, 18, 841–851.
- (1927): "Probable Inference, and the Law of Succession and Statistical Inference," *Journal of the American Statistical Association*, 22, 209–211.
- WILSON, E. B., AND M. M. HILFERTY (1929): "Note on C.S. Peirce's Experimental Discussion of the Law of Errors," *Proceedings of the National Academy of Sciences of the U.S.A.*, 15, 120–125.
- YULE, G. U. (1917): *An Introduction to the Theory of Statistics*. Charles Griffen: London, 4th edn.