Robust statistics for rejection-sampling timings

Daniel J. Bernstein[1,2]

[1] Department of Computer Science, University of Illinois at Chicago, USA
[2] Institute of Information Science, Academia Sinica, Taiwan
djb@cr.yp.to

Abstract. This paper gives an example of a real-world benchmarking task for which medians and quartiles are neither robust nor stable. This paper suggests using $[1/8, 3/8]$, $[3/8, 5/8]$, $[5/8, 7/8]$ means as simple, stable, robust replacements for quartiles.

1 The standard praise for medians

In his 1981 textbook "Robust statistics", Huber [10, page 107] (see also [11, page 106]) wrote that "the so-called median absolute deviation (MAD) has emerged as the single most useful ancillary estimate of scale". The statistician is given many observations $x_1, \ldots, x_n$ from some distribution; uses the median M of $x_1, \ldots, x_n$ to estimate the location (center) of the distribution; and uses the median of $|x_1 - M|, \ldots, |x_n - M|$ to estimate the scale (dispersion) of the distribution. "Ancillary" refers to the common situation that scale is "a nuisance parameter in location". The top goal is to estimate the location, but a wide distribution makes this unreliable; one estimates scale as an indication of the level of unreliability.

Why estimate location and scale using median and median absolute deviation instead of mean and standard deviation? The basic argument for medians in [10, Section 4.2] is that "the median achieves the smallest maximum bias among all translation invariant functionals". Here "bias" refers to the effect of starting from a normal distribution but corrupting a small fraction ϵ of the data. Obviously this contamination can create an arbitrarily large change in the mean no matter how small ϵ is, while the median turns out to minimize the worst-case effect of this contamination. For other robustness metrics favoring the median, see, e.g., [14, Section 2] ("the robust sample median ... can be shown to have the maximal breakdown point among all translation-invariant location estimators"; "the median was shown to be the most B-robust estimator").

Permanent ID of this document: df535ec41e9f70f55c3c1bbdf81d54bd1a6d69ee. Date: 2025-07-20. This work was funded by the Taiwan's Executive Yuan Data Safety and Talent Cultivation Project (AS-KPQ-109-DSTCP). "Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s)."

2 Median coin-flip timings

Let's see how well these supposedly robust statistics do for a simple example: the distribution of the number of times you need to flip a fair coin before it comes up heads. The population distribution is a textbook geometric distribution: 1 with probability 1/2; 2 with probability 1/4; 3 with probability 1/8; etc.

The median of the population distribution is 1.5. The median of a large sample is unstable: it might occasionally match 1.5 (the probability of this is $\Theta(1/\sqrt{n})$ if there are n observations and n is even), but it has $1/2 - O(1/\sqrt{n})$ chance of being 1 and $1/2 - O(1/\sqrt{n})$ chance of being 2.

For essentially the same reason, the median of this distribution is fragile under arbitrarily small amounts of contamination: slight corruption favoring low values will push the median from 1.5 down to 1, while slight corruption favoring high values will push the median from 1.5 up to 2. In other words, Hampel's "influence function" from [9] is unbounded. The median absolute deviation similarly lacks stability and robustness.[3]

This is a straightforward argument against using medians and median absolute deviations whenever a quantile function might be non-differentiable or might have large derivatives: in particular, for discrete distributions. This doesn't contradict the calculations in [10, Section 4.2]: those calculations consider only a limited class of starting distributions. Also, this argument for avoiding the median is not an argument for going back to the dark ages and using the mean:[4] one wants a statistic that is meaningful for discrete distributions *and* is robust against contamination.

3 A real example from software benchmarking

For many years I have been co-managing a project that collects benchmarks of cryptographic software on many computers; see [4]. Software timings are often contaminated (for a variety of reasons: e.g., the computer is distracted by an incoming network packet), sometimes spoiling the mean and variance of many timings; so, from the outset, the project instead reported medians of many timings, along with 1st and 3rd quartiles. This is a simple success story for robust statistics, right?

No, it's not that simple. More and more of the software that has been added to the project over the years relies on rejection-sampling loops. The rejection probabilities vary but sometimes, like the coin-flip example, create large jumps in the quantile function close to 25%, 50%, or 75%.

As a concrete example, consider the sample quantile function displayed in Figure 3.1. This graph shows 93 measurements,[5] sorted into increasing order, of

[3] For comparison, [1, Section 2.1] gives a more complicated example of the instability of the median absolute deviation—but still claims that the median absolute deviation is a "robust measure of statistical dispersion".

[4] Regressing to the mean, one might say.

[5] The measurement program has a 32-iteration loop of checking the clock and calling the function being timed, producing 31 timings. The program was run 3 times.

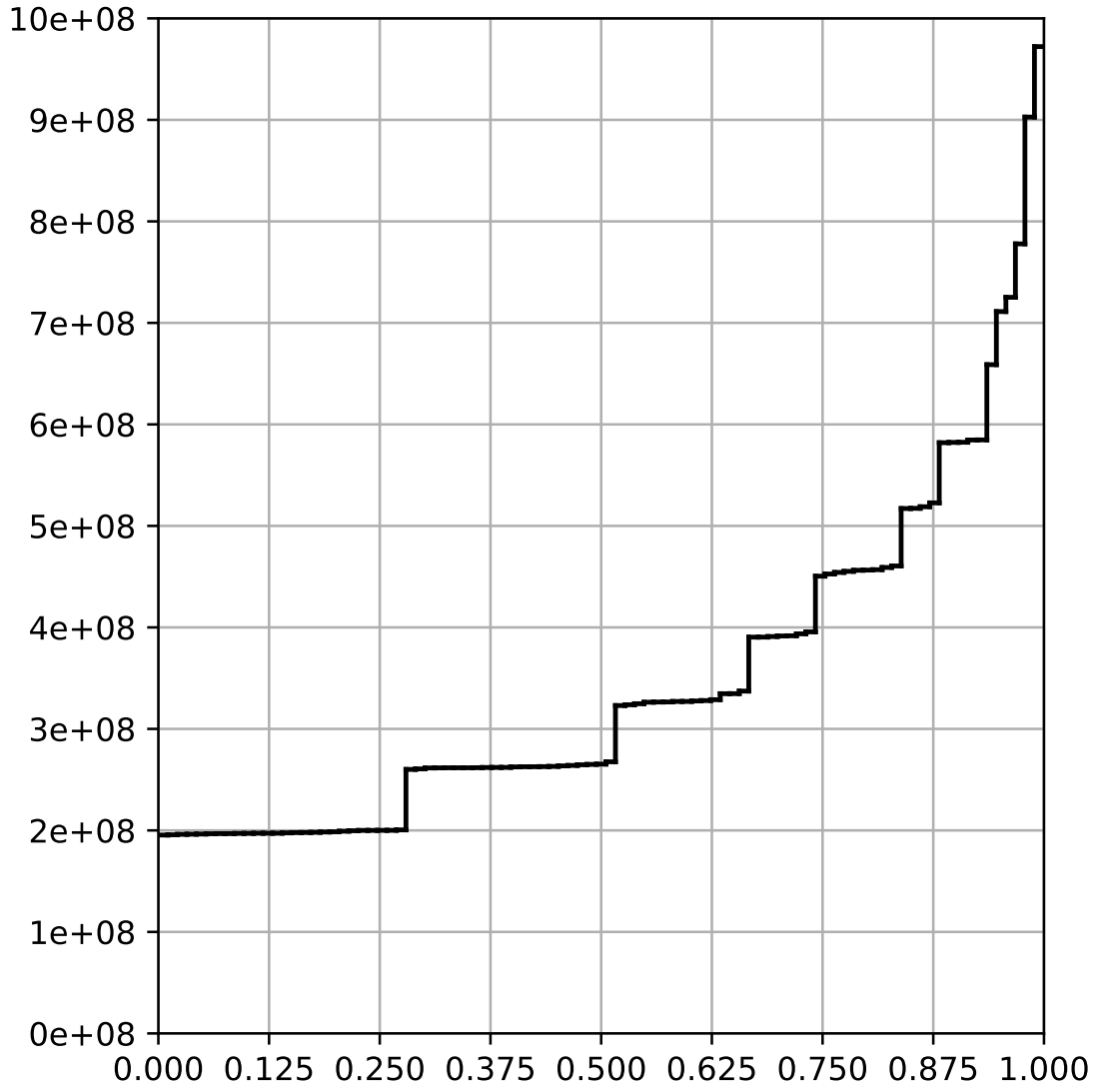

Fig. 3.1. Quantile function (transposed cdf) for 93 observations of cycle counts for `mceliece6960119` key generation on one core of a 3GHz Intel Xeon E3-1220 v5 with overclocking disabled.

the time to generate keys for a cryptosystem called `mceliece6960119` described in [2]. This software actually relies on multiple rejection-sampling loops, but the most important rejection-sampling loop tries to invert a random-looking $d \times d$ matrix with entries in the field of 2 elements, specifically with $d = 1547$. The probability p_d of invertibility for a $d \times d$ matrix generated uniformly at random is $\prod_{1 \le i \le d}(1 - 1/2^i)$ by, e.g., [6, page 77, Theorem 99], and one has

$$\lim_{d \to \infty} p_d = \prod_{i \ge 1}(1 - 1/2^i) = \sum_{k \in \mathbb{Z}}(-1)^k 2^{-k(3k+1)/2} \approx 0.288788$$

by Euler's pentagonal-number theorem [7]; p_d is only negligibly different from the limit for $d \geq 20$. The population distribution thus jumps at $p_d \approx 0.288788$, at $1-(1-p_d)^2 \approx 0.494178$, at $1-(1-p_d)^3 \approx 0.640253$, at $1-(1-p_d)^4 \approx 0.744144$, etc. Sampling randomly nudges the jumps, so there is nothing surprising about the second jump in Figure 3.1 being after 0.5 instead of before 0.5, producing a sample median around $2.7 \cdot 10^8$ where the population median is around $3.2 \cdot 10^8$. (The population mean is around $3.5 \cdot 10^8$.)

In [13], as part of justifying a simplified definition of the population median, Larocque and Randles comment that "it is a rare discrete population" for which the image of the cdf includes exactly 0.5. Formally, the rejection-sampling loop described above is not an exception: 0.494178 is not 0.5. But 0.494178 is close enough to 0.5 to create a large influence function, and to create instability of the sample median until the sample size is very large. Even with 10000 observations, there will be variations on the scale of 0.01 starting from 0.494178, so both $2.7 \cdot 10^8$ and $3.2 \cdot 10^8$ have a good chance of appearing as the sample median. Similar comments apply to other quartiles, the median absolute deviation, etc.

Thousands of cryptographic functions are benchmarked in [4], and most of them do not trigger the same issue. On the other hand, this particular function has been deployed and is one of relatively few functions currently under consideration for standardization by ISO; see generally [3]. In any event, I would like *all* of the reported benchmark numbers to be reliable.

4 Solutions in the literature

The literature on robust statistics considers many location statistics other than the median. For example, Tukey wrote the following in 1960 [19, Section 17]: "In large samples the sample mean is not nearly so safe an indicator of location as is the mean of the observations which remain after a small percentage of the highest, and an equal percentage of the lowest, have been set aside (use of a lightly truncated mean)." An example is the interquartile mean, although Tukey's examples of "light" truncation were at most 6% rather than 25%.

Daniell [5] had already considered "discard averages" in 1920, as noted in, e.g., [18].[6] Today "trimmed means" are well known, even if not as well known as medians.[7] An interquartile mean is easy to compute, easy to explain, and

[6] Daniell cites Poincaré [16, page 211] for discarding outliers ("rejeter une observation qui présente avec toutes les autres une divergence exagérée"). Poincaré in [16, pages 212ff] considers different models for a contaminated distribution and obtains different rules for which outliers to discard; Daniell starts with more general approaches such as "quartile-discard averages" (interquartile means), and analyzes how well those approaches apply to various distributions.

[7] I tried Google Scholar searches on 19 July 2025 for articles since 2024, putting search phrases into quotes. Google Scholar reported ("about") 11600 results for "trimmed mean"; 477 results for "truncated mean"; 319 results for "interquartile mean"; and 63500 results for "median", although a skim rapidly found that some of these were for, e.g., "median sternotomy".

not afraid of discrete distributions. For the coin-flip example from Section 2, the interquartile mean of a large sample is consistently close to 1.5.

There are other choices in the literature, including choices that have advantages over trimmed means. For example, Stigler [**17**] showed in 1974 that samples of a smoothed trimmed mean such as $16 \int_{0.25}^{0.75} (0.25 - |x - 0.5|) Q(x)\, dx$ are asymptotically normal for a wide range of quantile functions Q, while a similar result for samples of the interquartile mean $2 \int_{0.25}^{0.75} Q(x)\, dx$ requires more stringent assumptions about the behavior of Q at 0.25 and at 0.75.

On the other hand, trimmed means are easier to explain, even for an audience that hasn't seen them before. Explainability is an important feature for someone like me choosing statistics to use in an application, whereas I don't see why this application needs the asymptotic normality result in [**17**] (even though I'm looking at Q in Figure 3.1 that really does jump close to 0.25 and close to 0.75!). None of the other robust-even-for-the-discrete-case location statistics seem to be as popular as trimmed means.

What about scale statistics? Again the literature provides various options—for example, Tukey writes that "the use of truncated variances is likely to be quite satisfactory"—but these seem far less widely used than standard deviations, quartiles, and the interquartile range,[8] all of which are unsatisfactory.

In short, the literature has solutions to the problem at hand, but it seems that the solutions are *popular* only for location, not for scale. This makes it even more important for me to be able to explain whichever solution I end up using. From this perspective, I'm not satisfied saying "you know about mean and standard deviation already; I'll use trimmed mean and trimmed standard deviation". Even without trimming, standard deviation is more complicated than mean: consider the squares, the square roots, the common variants that differ noticeably from each other when the sample size is small, the tricky visual interpretation. This is all to make some formulas work out nicely, but trimming breaks those formulas. What do I say to a reader who complains that a trimmed standard deviation is neither comprehensible nor standard?

Furthermore, I want to have three statistics rather than two, to see not just location and scale but also skewness—the asymmetry of Figure 3.1, for example. Can I really explain replacing the 1st quartile with, say, the interquartile mean of the variable number of observations below the interquartile mean of the sample? I'd rather have something simpler.

[8] More Google Scholar searches on 19 July 2025 for articles since 2024: 49800 results for "standard deviation"; 22600 results for "interquartile range"; 18900 results for "quartile"; 45 results for "trimmed standard deviation"; 12 results for "truncated standard deviation"; 73 results for "truncated variance"; 15 results for "trimmed variance". Note that "interquartile range" typically refers to the difference between the 3rd quartile and the 1st quartile, rather than the two statistics separately—this is confusing terminology and should be fixed; see generally [**8**]—but, either way, just a few quantiles are being used.

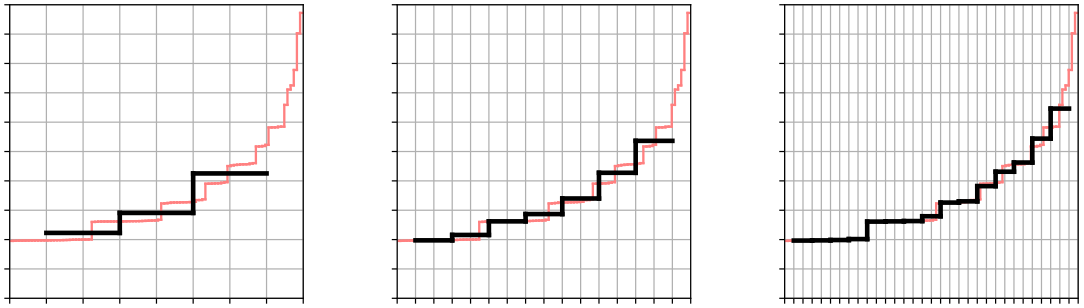

Fig. 5.1. First graph, in black: averages on $[1/8, 3/8]$ and $[3/8, 5/8]$ and $[5/8, 7/8]$ of the quantile function from Figure 3.1. Second graph, in black: averages on $[1/16, 3/16]$ etc. Third graph, in black: averages on $[1/32, 3/32]$ etc. Original quantile function is shown in light red.

5 Stabilized quartiles

The first black graph in Figure 5.1 is a graph of the following three numbers derived from the sample in Figure 3.1:

- "$[1/8, 3/8]$ mean" or "StQ_1": the mean between the 1st and 3rd octiles, as a stabilized substitute for the 1st quartile.
- "$[3/8, 5/8]$ mean" or "StQ_2": the mean between the 3rd and 5th octiles, as a stabilized substitute for the median.
- "$[5/8, 7/8]$ mean" or "StQ_3": the mean between the 5th and 7th octiles, as a stabilized substitute for the 3rd quartile.

Figure 5.2 shows how easy it is to compute these three statistics in Python. One can use these statistics in the same way as quartiles: for example, computing the difference of StQ_3 and StQ_1 as an overall estimate of scale, or computing appropriate ratios as an estimate of skewness. As a concrete example, where [20] uses Bowley's skewness coefficient $(Q_3 + Q_1 - 2Q_2)/(Q_3 - Q_1)$, one can instead use $(\mathrm{StQ}_3 + \mathrm{StQ}_1 - 2 \cdot \mathrm{StQ}_2)/(\mathrm{StQ}_3 - \mathrm{StQ}_1)$.

For a normal distribution, these stabilized quartiles are numerically close to quartiles: an integration exercise concludes that $\mathrm{StQ}_3 - \mathrm{StQ}_1$ is

$$2^{5/2}\pi^{-1/2}\left(\exp\left(-\operatorname{erf}^{-1}(1/4)^2\right) - \exp\left(-\operatorname{erf}^{-1}(3/4)^2\right)\right) \approx 1.3867336971$$

times the standard deviation, while the difference of third and first quartiles is $2^{3/2}\operatorname{erf}^{-1}(0.5) \approx 1.3489795004$ times the standard deviation.

These stabilized quartiles are robust against contamination. Their breakdown points (1/8 for the 1st and 3rd, 3/8 for the 2nd) are not as high as for quartiles (1/4, 1/2) or median absolute deviation (1/2); but, again, influence functions and sampling show that quartiles and median absolute deviation are fragile and unstable starting from the distribution in Section 2.

Asymmetrically trimmed means aren't new (see, e.g., [12]), but I haven't found literature proposing these statistics as an easy replacement for quartiles. I

```
def mean(S):
  S = list(S)
  return sum(S)/len(S)

def stq(S):
  S = sorted(8*list(S))
  n = len(S)//8
  return mean(S[n:3*n]),mean(S[3*n:5*n]),mean(S[5*n:7*n])
```

Fig. 5.2. Python 3 function `stq` to compute the three stabilized quartiles of a sample. In applications where the sample already has length divisible by 8, one can skip the initial multiplication by 8; one can also, by adding code to handle edges, avoid the initial multiplication in all cases.

think the description above in terms of octiles is simpler than a description as interquartile means of (1) the observations below the median, (2) the interquartile observations, and (3) the observations above the median.

The second and third graphs in Figure 5.1 are similar to the first but show stabilized octiles and stabilized hexadeciles. The narrower spacing makes each statistic less stable, and (because the spacing comes closer to the edges) makes the outer statistics more vulnerable to contamination; but using more statistics provides more information, coming closer and closer to showing the full sample distribution, which in turn says something useful about the population distribution if there are enough samples.

One can instead work directly with the full sample distribution. I normally graph the distribution; sometimes I compare it directly to a model distribution. But I also want to compress this information to a few statistics. For numerical tables summarizing benchmark results, I continue to think that 3 statistics are the right level of detail, so I'm planning to switch to stabilized quartiles.

A histogram also compresses a sample distribution to fewer numbers, but a histogram chooses an equal (often artificial) spacing of *values*. The numbers here, like conventional quartiles, choose an equal spacing of *probabilities*, with values naturally dictated by the sample provided as input. A separate issue is that a histogram traditionally displays density, which is less stable than distribution.

6 Experimental stability evaluation

To close, I'll experimentally evaluate the stability of StQ_1, StQ_2, and StQ_3 on a specific distribution for which medians are unstable. For simplicity and reproducibility, I'll again take the coin-flip distribution from Section 2.

I tried 100 observations of the number of coin flips—simulated as the number of `random.randrange(2)`—required to obtain heads, and calculated stabilized quartiles for that sample. The hope, of course, is to obtain not just something stable but something unbiased, close to the stabilized population quartiles 1,

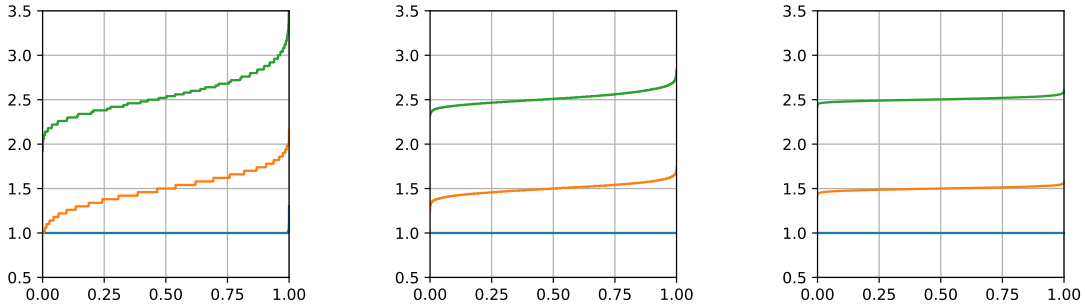

Fig. 6.1. Experimental distribution of StQ_1 (bottom, blue), StQ_2 (middle, orange), and StQ_3 (top, green) for n observations of the number of coin flips to obtain heads, where $n = 100$ (left), $n = 1000$ (middle), or $n = 10000$ (right). Each distribution is displayed as a quantile function for 10000 size-n samples.

1.5, and 2.5; but with sample size 100 one also expects deviations on the scale of 0.1.

I repeated the computation for 10000 samples, and plotted the resulting experimental distributions of StQ_1, StQ_2, and StQ_3, obtaining the three curves in the left graph in Figure 6.1. One sees from the graph that StQ_1 is occasionally (right side of the graph) above 1 (for the occasional samples where ≥ 2 appeared before the 3rd octile), but is almost always 1. StQ_2 is occasionally 1, and occasionally 2 or higher, but between 1.4 and 1.6 about half the time. StQ_3 has a broader range and an evident asymmetry (the right side is pushed up by the frequent cases where ≥ 4 appears before the 7th octile) but is between 2.4 and 2.7 about half the time.

The other two graphs in Figure 6.1 come from taking 10000 size-1000 samples and 10000 size-10000 samples. Unsurprisingly, the experimental distributions of the stabilized sample quartiles tighten around the stabilized population quartiles as the sample size increases.

Finally, as a closer look at the shrinkage, Figure 6.2 plots the experimental distributions of $\sqrt{n}(\mathrm{StQ}_3 - 2.5)$ for $n \in \{100, 1000, 10000\}$. These distributions are very close to each other, as one would hope. Given that the jump in the population distribution at the 7th octile is at the edge of StQ_3, and given the aforementioned asymptotic results from Stigler [**17**], it is unsurprising that there is a persistent asymmetry in Figure 6.2 as n increases.

References

[1] Andrey Akinshin, *Quantile absolute deviation* (2022). URL: https://arxiv.org/abs/2208.13459. Citations in this document: §2.

[2] Martin R. Albrecht, Daniel J. Bernstein, Tung Chou, Carlos Cid, Jan Gilcher, Tanja Lange, Varun Maram, Ingo von Maurich, Rafael Misoczki, Ruben Niederhagen, Kenneth G. Paterson, Edoardo Persichetti, Christiane Peters, Peter Schwabe, Nicolas Sendrier, Jakub Szefer, Cen Jung Tjhai, Martin

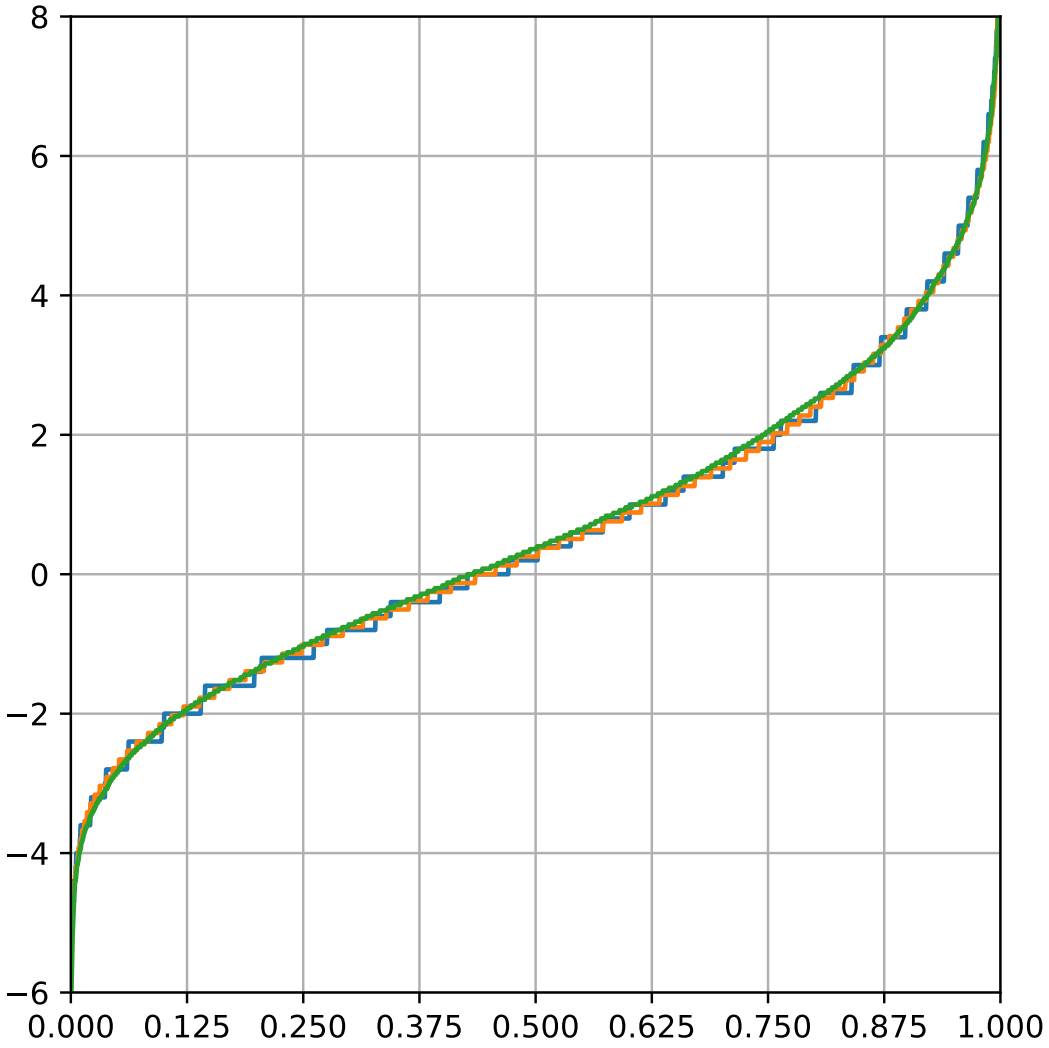

Fig. 6.2. Experimental distribution of $\sqrt{n}(\mathrm{StQ}_3 - 2.5)$ for n observations of the number of coin flips to obtain heads, where $n = 100$ (blue), $n = 1000$ (orange), $n = 10000$ (green). Each distribution is displayed as a quantile function for 10000 size-n samples.

Tomlinson, Wen Wang, *Classic McEliece: conservative code-based cryptography: cryptosystem specification* (2022). URL: https://classic.mceliece.org/mceliece-spec-20221023.pdf. Citations in this document: §3.

[3] Daniel J. Bernstein, *McEliece standardization* (2025). URL: https://blog.cr.yp.to/20250423-mceliece.html. Citations in this document: §3.

[4] Daniel J. Bernstein, Tanja Lange (editors), *eBACS: ECRYPT Benchmarking of Cryptographic Systems* (2025), Accessed 19 July 2025. URL: https://bench.cr.yp.to. Citations in this document: §3, §3.

[5] Percy John Daniell, *Observations weighted according to order*, American Journal of Mathematics **42** (1920), 222–236. URL: https://www.jstor.org/stable/pdf/2370465.pdf. Citations in this document: §4.

[6] Leonard Eugene Dickson, *Linear groups with an exposition of the*

Galois field theory, Teubner, 1901. URL: https://archive.org/details/lineargroupswith00ledi/. Citations in this document: §3.

[7] Leonhard Euler, *De mirabilis proprietatibus numerorum pentagonalium*, Acta Academiae Scientiarum Imperialis Petropolitanae **1780:I** (1783), 56–75. URL: https://scholarlycommons.pacific.edu/euler-works/542/. Citations in this document: §3.

[8] Cynthia P. Haanappel, Anne F. Voor in 't holt, *Using the interquartile range in infection prevention and control research*, Infection Prevention in Practice **6** (2024), 100337. URL: https://pure.eur.nl/en/publications/using-the-interquartile-range-in-infection-prevention-and-control. Citations in this document: §4.

[9] Frank Rudolf Hampel, *Contributions to the theory of robust estimation*, Ph.D. thesis, University of California, Berkeley, 1968. Citations in this document: §2.

[10] Peter J. Huber, *Robust statistics*, Wiley, 1981; see also newer version [**11**]. ISBN 978-0-47141805-4. DOI: 10.1002/0471725250. Citations in this document: §1, §1, §2.

[11] Peter J. Huber, Elvezio M. Ronchetti, *Robust statistics*, 2nd edition, Wiley, 2009; see also older version [**10**]. ISBN 978-0-47012990-6. DOI: 10.1002/9780470434697. Citations in this document: §1.

[12] Jonathan Kearns, *The distribution and measurement of inflation* (1998). URL: https://www.rba.gov.au/publications/rdp/1998/pdf/rdp9810.pdf. Citations in this document: §5.

[13] Denis Larocque, Ronald H. Randles, *Confidence intervals for a discrete population median*, The American Statistician **62** (2008), 32–39. Citations in this document: §3.

[14] Po-Ling Loh, *A theoretical review of modern robust statistics*, Annual Review of Statistics and Its Application **12** (2025), 477–496. DOI: 10.1146/annurev-statistics-112723-034446. Citations in this document: §1.

[15] Ingram Olkin, Sudhist G. Ghurye, Wassily Hoeffding, William G. Madow, Henry B. Mann (editors), *Contributions to probability and statistics: Essays in honor of Harold Hotelling*, Stanford Studies in Mathematics and Statistics, 2, Oxford University Press, 1960. See [19].

[16] Henri Poincaré, *Calcul des probabilités*, 2nd edition, Gauthier-Villars, 1912. URL: https://archive.org/details/calculdeprobabil00poinrich. Citations in this document: §4, §4.

[17] Stephen M. Stigler, *Linear functions of order statistics with smooth weight functions*, The Annals of Statistics **2** (1974), 676–693. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-2/issue-4/Linear-Functions-of-Order-Statistics-with-Smooth-Weight-Functions/10.1214/aos/1176342756.pdf. Citations in this document: §4, §4, §6.

[18] Stephen M. Stigler, *The changing history of robustness*, The American Statistician **64:4** (2010), 277–281. DOI: 10.1198/tast.2010.10159. Citations in this document: §4.

[19] John W. Tukey, *A survey of sampling from contaminated distributions*, in [**15**] (1960), 448–485. URL: https://jugander.github.io/rare/Tukey1960_ContaminatedDistributions.pdf. Citations in this document: §4.

[20] Randal Verbrugge, Saeed Zaman, *Improving inflation forecasts using robust measures*, International Journal of Forecasting **40** (2024), 735–745. URL: https://www.clevelandfed.org/-/media/project/clevelandfedtenant/clevelandfedsite/publications/working-papers/2023/wp2223r.pdf. DOI: 10.1016/j.ijforecast.2023.05.003. Citations in this document: §5.