1 Introduction

To measure the control of an industrial process, it is often necessary to estimate the variance in a measurement made on the parts being produced. To do this very quickly one can use a rule of thumb that requires only the range (either because it only requires two measurements if the parts are sorted or because it may be the only statistic available), as a simple and speedy estimate of the standard deviation. One such rule found in many introduction to statistics textbooks [6] is called simply the **range rule of thumb**:

$$\sigma \approx \frac{\text{sample range}}{4}.$$

We will see below that this estimate works well *only* for a normal distribution and only when the sample size is around 30. To address this, some engineering statistics texts suggest using a different value (other than 4) for each of the smaller sample sizes n. For example, in [5] it is suggested that for the normal curve we use

$$\sigma \approx \frac{\text{sample range}}{\zeta(n)},$$

where $\zeta(n)$ is as in Table 1. Our goal in this paper is to find a simple algebraic approximation for such values $\zeta(n)$, and to do this not only for the normal distribution, but also the uniform and the exponential distributions. This will give us several new rules of thumb, including one for each of the most commonly used distributions in industry.

Table 1: Constants for a rule of thumb

sample size n	**2**	**3**	**4**	**5**
$\zeta(n)$	1.128379	1.692569	2.058751	2.325929

Our first approach to finding new rules of thumb is empirical—we use Monte Carlo methods written in the statistical programming language R. We create simple functions which we can call repeatedly to (1) generate random samples from our chosen distribution, (2) calculate the statistic

$$\zeta = \frac{\text{sample range}}{\text{population standard deviation}}$$

for each sample, and then (3) return the list of these statistics ζ to the user. For example, Table 2 shows the code used to generate a list of ζ's with length `trials` from samples of size `sampleSize` from the standard normal distribution.

The idea of this function is to allow the user to input the number of trials to be performed and the sample size, and from this the function will return the statistics about the sample $\zeta(n)$'s. In particular we calculate ζ for each of the thousands of samples generated, and then

Table 2: A generic R program to compute ζ for the normal distribution

```
normalGenerator <- function(sampleSize,trials)
{
  newList <- list()
  populationSD = 1
  for(i in 1:trials)
      {
       tempStat=rnorm(sampleSize)
       tempRange=max(tempStat)-min(tempStat)
       tempRatio=tempRange/populationSD
       newList[i]=tempRatio
      }
  q<-sort(unlist(newList))
  return(...)
}
```

for the resulting sets of $\zeta(n)$ values, we calculate the mean, the 2.5% and 97.5% quantiles. These quantiles can be produced by various commands in R such as `quantile(newList)`. We manually export these data to Excel, and then visually develop empirical models.

It is important to notice that for most distributions for the original samples (even the normal distribution), *the resulting distribution for the ratios $\zeta(n)$ will not be normal*. This is most easily seen by glancing at the graphs which show the mean and middle 95% of each set of zetas (the vertical lines in Figures 1, 3, 4 and 5). If the distribution of the $\zeta(n)$ values was normal for any fixed n, the 2.5%, and 97.5% quantiles would be symmetric about the mean in these graphs, and they clearly are not. So the intervals given are prediction intervals for ζ.

In section 5 we will compare the results of the Monte Carlo method (described above) with those from a theoretical approach.

2 The Normal Distribution

Many applications are adequately modeled by a normal distribution—these include weights of babies, heights, grades, etc. For this reason, it is the first continuous distribution discussed in introductory statistics courses. Note that it is sufficient to calculate a rule of thumb for just the standard normal (mean 0, standard deviation 1) and this same rule (same values of $\zeta(n)$) will work for *any* normal distribution. To see this we prove the following. Similar results hold for our other distributions.

Theorem 2.1. *Let $X \sim N(\mu, \sigma^2)$ be a normal random variable. Then for all constants a*

and b

$$\zeta(aX + b) = \zeta(X).$$

Proof. It is well known that if $X \sim N(\mu, \sigma^2)$, then the random variable $W = aX + b$ also follows a normal distribution. The population standard deviation of the transformed data $W = aX + b$ is $|a|$ times as large as that for X. Also for any sample of X values, say $\{X_1, X_2, \ldots, X_n\}$, the range of the transformed values $\{aX_1 + b, aX_2 + b, \ldots, aX_n + b\}$ is $|a|$ times as large. Since $\zeta(aX + b)$ is the ratio of these two values (the population standard deviation and the sample range), the factors of $|a|$ cancel and we have $\zeta(aX + b) = \zeta(X)$. □

To develop our rule of thumb, for twenty-two sample sizes from 2 to 50,000, we used our program to find thousands of samples and their associated ζ values. These are summarized in Table 3.

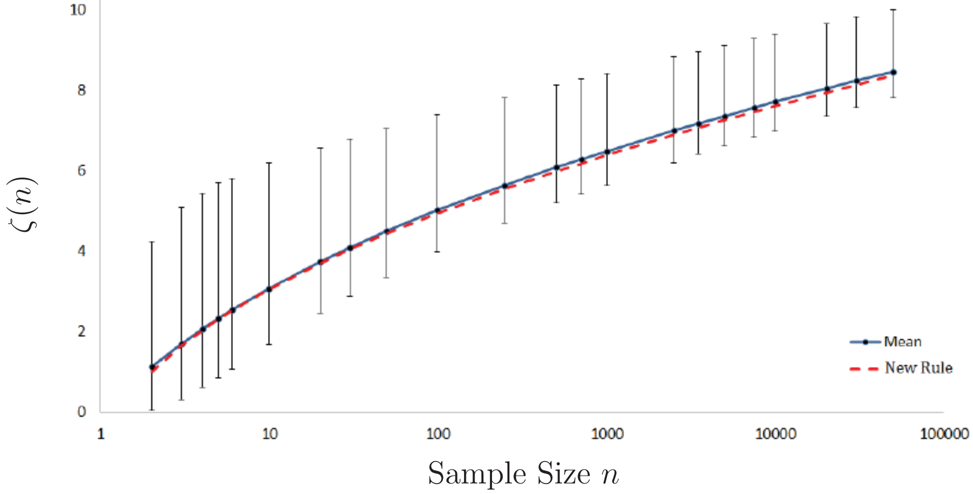
Table 3: Random sample data for ζ for a normal distribution

sample			**percentiles**		
size	**trials**	**mean**	**2.5%**	**97.5%**	**model**
2	50000	1.1255	0.0446	3.1576	0.9977
3	50000	1.6913	0.3008	3.6944	1.6444
4	50000	2.0617	0.6035	3.9655	2.0322
5	50000	2.3271	0.8485	4.2073	2.3059
6	50000	2.5340	1.0746	4.3356	2.5157
10	50000	3.0749	1.6731	4.7882	3.0523
20	50000	3.7379	2.4567	5.2884	3.6925
30	50000	4.0851	2.8642	5.5691	4.0327
50	50000	4.4981	3.3539	5.9051	4.4337
⋮	⋮	⋮	⋮	⋮	⋮
30000	10000	8.2257	7.5538	9.1444	8.1323
50000	10000	8.4549	7.8036	9.3448	8.3680

These data were transferred to Microsoft Excel to produce the graph in Figure 1. (Due to the wide range of sample sizes, we plotted the data on a logarithmic scale.) Each vertical bar represents a 95% prediction interval for $\zeta(n)$ and the dot on that bar is the point estimate of the $\zeta(n)$ values.

Looking at the data we decided to begin with a natural logarithmic model and build on it. However, since this is a logarithmic scale, the data should appear “linear” if it is to fit a logarithmic model. We noticed the curve is almost the same as a square root function, so the next approach was to take the square root of the natural log of the sample size, and from that we proceeded to find the appropriate transformations. (In other words, we considered lots of things until we got a visual fit.) The end result was that $\zeta(n) \approx 3\sqrt{\ln n} - 1.5$, which we restate in the following rule of thumb. (This rule is plotted as the dashed line in Figure 1.)

Figure 1: Mean, 2.5% and 97.5% percentiles of ζ by the size of samples which were taken from a normal distribution



Empirical Rule of Thumb 2.2. *For samples of size n from a normal distribution:*

$$\sigma \approx \frac{\text{range}}{3\sqrt{\ln n} - 1.5}.$$

To further test this model empirically, we first compared it to the old rule ($\zeta = 4$) by applying them both to *new* random samples from a normal population with a mean of 20 and standard deviation of 15. As you can see in Figure 2, neither model fared very well when the sample size was less than 10. For larger values the new rule of thumb appears to work well, but the old rule of thumb becomes increasing less accurate.

Another way to test this model is to compare our model for ζ to the values in Table 4 which are the correct theoretical values from [2, 3] and [5]. Our rule produces estimates of σ that are slightly too large; we decided to leave it like this for "safety" (it may be more conservative to assume less control in the industrial process).

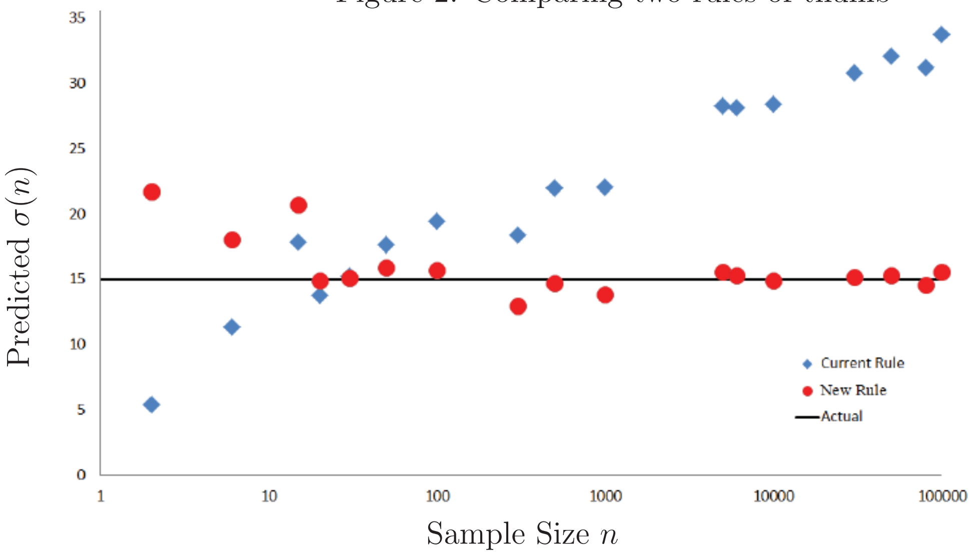
Table 4: Constants ζ for the rule of thumb: $\sigma \approx \text{range}/\zeta$

sample size n	**2**	**3**	**4**	**5**	**10**	**30**	**60**	**100**
Theoretical	1.12838	1.69257	2.05875	2.32593	3.07751	4.08552	4.63856	5.01519
Our model	0.99766	1.64444	2.03223	2.30591	3.05228	4.03270	4.57035	4.93789

3 The Exponential Distribution

Classical examples that lead to the use of the exponential distribution (defined by the p.d.f $p(x) = \frac{1}{\theta}e^{-x/\theta}$ for $x \geq 0, p(x) = 0$ otherwise) include the time it will take for radioactive

Figure 2: Comparing two rules of thumb



material to decay, and the lifetime of components like light bulbs. When determining the rule of thumb for the exponential distribution the same process used to find the estimate for the normal distribution was applied. We tried several values of θ and the distribution of ζ appears independent of θ, so we used $\theta = 3$ in our calculations. Naturally, we had to change the R code slightly to have it generate random data from the exponential distribution (see Table 5). From here we once again transferred the data into Excel and graphed the data with the statistic $\zeta(n)$ dependent upon the sample size n (see Figure 3)—so the linear appearance makes sense.

We were originally surprised by the linear appearance the data took and were expecting something more unusual (as we saw with the normal distribution). We first tried to fit a linear model to the data by using Excel's `Trendline` command. The results it gave us did not fit the data well enough. We next tried a model of the form $\zeta = a\ln(n) + b$. This gave us a model ($\zeta = \ln n + \frac{4}{9}$), that fit the data extremely well (see Empirical Rule of Thumb 3.1). We will later show this rule is asymptotically correct.

Empirical Rule of Thumb 3.1. *For samples of size n from an exponential distribution:*

$$\sigma \approx \frac{\text{range}}{\ln n + \frac{4}{9}}.$$

4 The Uniform Distribution

Programming languages today such as C++ have built in pseudo-random number generators. These generators will follow closely a uniform distribution. That is, given an interval from which a number may be chosen at "random", each number in that interval is equally likely to be picked. To find the rule of thumb for the uniform distribution we began by looking at

Table 5: Random sample data for ζ for an exponential distribution

sample			**percentiles**		
size	**trials**	**mean**	**2.5%**	**97.5%**	**model**
2	50000	0.9985	0.0254	3.6844	1.1376
3	50000	1.4980	0.1706	4.3976	1.5431
4	50000	1.8379	0.3473	4.7698	1.8307
5	50000	2.0723	0.5051	5.0492	2.0539
6	50000	2.2810	0.6548	5.2813	2.2362
10	50000	2.8209	1.0793	5.9044	2.7470
20	50000	3.5472	1.7315	6.5945	3.4402
30	50000	3.9685	2.1297	7.0401	3.8456
50	50000	4.4806	2.6352	7.5609	4.3565
⋮	⋮	⋮	⋮	⋮	⋮
30000	10000	10.8960	9.0045	13.9092	10.7534
50000	10000	11.4186	9.5015	14.4478	11.2642

$U[0,1]$. Now it should be noted that for any uniform distribution $U[a,b]$, the endpoints a and b can be linearly transformed into 0 and 1, and so we know the result of Theorem 1.1 also applies to the uniform distribution as well. The results of the data we generated are located in Table 6. The graph in Figure 4 clearly has an asymptote as $n \to \infty$, because the expected range will approach 1, so we have $\zeta \to \frac{1}{\sigma} = \sqrt{12}$. We tried models which were a rational function times $\sqrt{12}$ and settled on $\zeta(n) \approx \frac{n-1}{n+1}\sqrt{12}$. (We will prove this is the correct expected value of $\zeta(n)$ in section 5.1.) This is expressed as Empirical Rule of Thumb 4.1.

Empirical Rule of Thumb 4.1. *For samples of size n from a uniform distribution:*

$$\sigma \approx \frac{n+1}{n-1} \cdot \frac{\text{range}}{\sqrt{12}}.$$

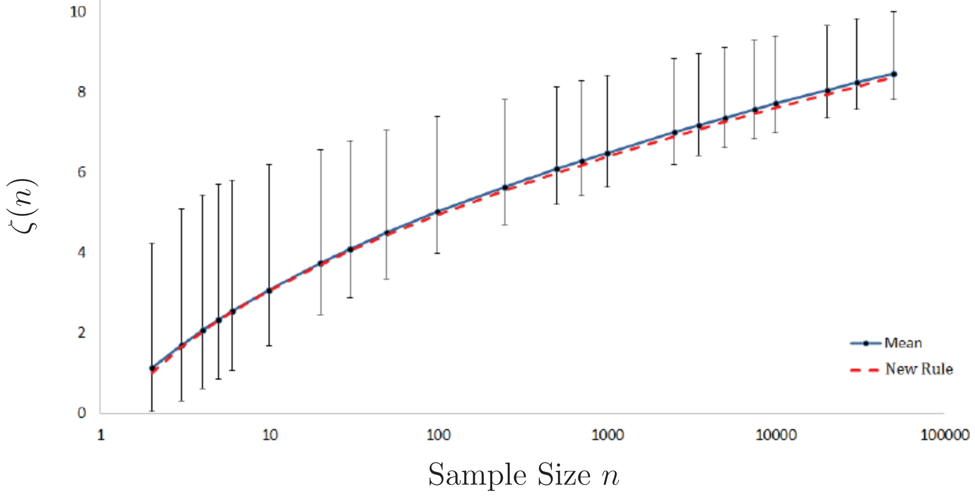
5 A Theoretical Approach

In this section we will attempt to verify our empirical models using theoretical methods. Suppose the random variables $X_1, X_2, \ldots, X_n$ are a sample of size n from a distribution with cumulative distribution function $P(x)$ (and p.d.f. $p(x) = P'(x)$). When we arrange the sample in ascending order

$$X_{(1)}, X_{(2)}, \ldots, X_{(n)},$$

we call $X_{(i)}$ the ith order statistic ($i = 1, 2, \ldots, n$). The difference $W_i = X_{(n-i)} - X_{(i+1)}$ is called the ith quasi-range and of course w_0 is just the range. In his text "Order Statistics" [2, p. 10], H. A. David gives the following joint distribution for the quasi-ranges.

Figure 3: Mean, 2.5% and 97.5% percentiles of ζ by the size of samples which were taken from an exponential distribution



$$f(w_r) = \frac{n!}{r!^2(n-2r-2)!}\int_{-\infty}^{\infty}(P(x)-P(x)P(x+w))^r \cdot \tag{5.1}$$
$$(P(x+w)-P(x))^{n-2r-2}\,p(x)p(x+w)dx.$$

In the following subsections we integrate $w_0 f(w_0)$ to find the expected value $E(w_0)$ of the the range, use this to compute a theoretical expected value for $\zeta(n)$, and then compare it to our Monte Carlo estimates.

5.1 Uniform Distribution

First, recall that it is sufficient to consider a uniform distribution on the interval $[0,1]$, so we begin with the following probability density function:

$$p(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

The cumulative density function, $P(x)$, is the anti-derivative of $p(x)$:

$$P(x) = \begin{cases} 0 & x < 0 \\ x & 0 \le x \le 1 \\ 1 & x < 1. \end{cases}$$

For $r = 0$, the probability density function for the range $w = w_0$ can be computed from Equation (5.1) as follows. Note that $p(x)p(x+w) = 0$ if $x < 0$ or $x + w > 1$, so our limits

Table 6: Random sample data for ζ for an uniform distribution

sample size	**trials**	**mean**	**percentiles 2.5%**	**97.5%**	**model**
2	50000	1.1532	0.0446	2.9100	1.1547
3	50000	1.7364	0.3344	3.1353	1.7321
4	50000	2.0772	0.6812	3.2280	2.0785
5	50000	2.3107	0.9873	3.2789	2.3094
6	50000	2.4736	1.2402	3.3174	2.4744
10	50000	2.8342	1.9713	3.3802	2.8343
20	50000	3.1331	2.6032	3.4219	3.1342
30	50000	3.2415	2.8711	3.4357	3.2406
50	50000	3.3284	3.0953	3.4472	3.3283
⋮	⋮	⋮	⋮	⋮	⋮
30000	10000	3.4639	3.4635	3.4641	3.4639
50000	10000	3.4640	3.4637	3.4641	3.4640

of integration run from 0 to $1-w$:

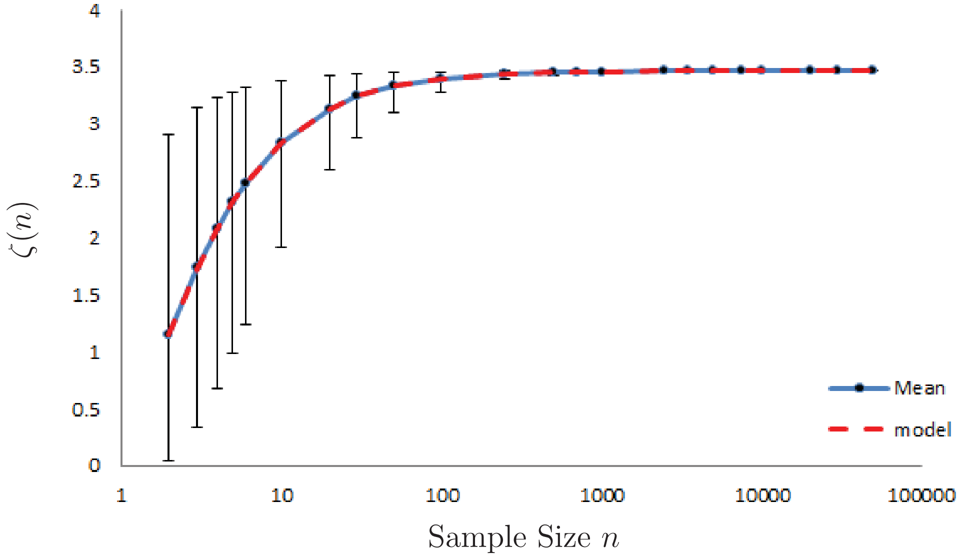
$$\begin{aligned} p(w) &= \frac{n!}{0!^2(n-2(0)-2)!}\int_0^{1-w}(x-x(x+w))^0((x+w)-x)^{n-2(0)-2}(1)(1)dx \\ &= \frac{n!}{(n-2)!}\int_0^{1-w} w^{n-2}dx \\ &= n(n-1)\left. xw^{n-2}\right|_0^{1-w} \\ &= n(n-1)(1-w)w^{n-2}. \end{aligned}$$

Now we use this to calculate the expected value of the range $w = w_0$ as follows:

$$\begin{aligned} E(w) &= \int_0^1 xp(x)dx \\ &= n(n-1)\int_0^1 x(x^{n-2}-x^{n-1})dx \\ &= n(n-1)\left.\left[\frac{x^n}{n}-\frac{x^{n+1}}{n+1}\right]\right|_0^1 \\ &= n(n-1)\frac{(n+1)-n}{n(n+1)} \\ &= \frac{n-1}{n+1}. \end{aligned}$$

Finally, recall the standard deviation of the uniform distribution on $[0,1]$ is $1/\sqrt{12}$, so the

Figure 4: Mean, 2.5% and 97.5% percentiles of ζ by the size of samples which were taken from a uniform distribution



expected value of $\zeta(n)$ is $\frac{n-1}{n+1}\sqrt{12}$; this is exactly what we computed empirically (Empirical Rule of Thumb 4.1).

Looking at the other quasi-ranges (r values greater than zero) we found

$$E(w) = \frac{n-(2r+1)}{n+1}. \tag{5.2}$$

This gives an additional (proven, not just empirical) rule of thumb:

Rule of Thumb 5.1. *For samples of size n from a uniform distribution:*

$$\sigma \approx \frac{n+1}{n-2r-1} \cdot \frac{(r\text{th quasi-range})}{\sqrt{12}} \qquad (r \geq 0,\ n \geq 2r+2).$$

5.2 Exponential Distribution

For the exponential distribution (with $\theta > 0$) we started with the probability density function

$$p(x) = \begin{cases} \frac{1}{\theta}e^{-x/\theta} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

so

$$P(x) = \begin{cases} 1 - e^{-x/\theta} & x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Table 7: Expected values of $\zeta = \frac{W_i}{\sigma}$ $(i = 0, 1)$ for the Exponential Distribution

n	$E\left(\frac{W_0}{\sigma}\right)$	$E\left(\frac{W_1}{\sigma}\right)$	n	$E\left(\frac{W_0}{\sigma}\right)$	$E\left(\frac{W_1}{\sigma}\right)$
3	$\frac{3}{2}$	-	9	$\frac{761}{280}$	$\frac{223}{140}$
4	$\frac{11}{6}$	$\frac{1}{2}$	10	$\frac{7129}{2520}$	$\frac{481}{280}$
5	$\frac{50}{24}$	$\frac{5}{6}$	11	$\frac{7381}{2520}$	$\frac{4609}{2520}$
6	$\frac{274}{120}$	$\frac{26}{24}$	12	$\frac{83711}{27720}$	$\frac{4861}{2520}$
7	$\frac{1764}{720}$	$\frac{154}{120}$	15	$\frac{1171733}{360360}$	$\frac{785633}{360360}$
8	$\frac{13068}{5040}$	$\frac{1044}{720}$	20	$\frac{275295799}{77597520}$	$\frac{10190221}{4084080}$

Because of the complexity of computing the integrals in Equation 5.1, and then integrating the result to find the expected value, we used the computational software package Maple [8]. We computed $E(w_r)$ with $r = 0$ and 1 to find the results in Table 7.

Our advisor helped us discover that for $r = 0$, these values are $H(n-1)$, where $H(n)$ is the nth harmonic number $H(n) = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}$; and for $r = 1$, these values are $H(n-2)-1$. The harmonic numbers have the asymptotic expansion

$$H(n) = \ln n + \gamma + \frac{1}{2n} - \frac{1}{12n^2} + \frac{1}{120n^4} + \mathcal{O}\left(\frac{1}{n^6}\right),$$

where γ is the Euler-Mascheroni constant $0.5772156649\ldots$ (see [7, p. 971, 1307]). It seems reasonable to conjecture

$$\begin{aligned} \frac{E(W_0)}{\sigma} &\approx \ln(n-1) + \gamma + \frac{1}{2n-2} \\ \frac{E(W_1)}{\sigma} &\approx \ln(n-2) + (\gamma-1) + \frac{1}{2n-4} \end{aligned}$$

We tested these against the values in Table 7 and with many other sample sizes n up to $n = 10{,}000$.

Recall that Empirical Rule of Thumb 3.1 is equivalent to $E(W_0)/\sigma \approx \ln n + \frac{4}{9}$ which does match the above asymptotically. But a better model might be the following.

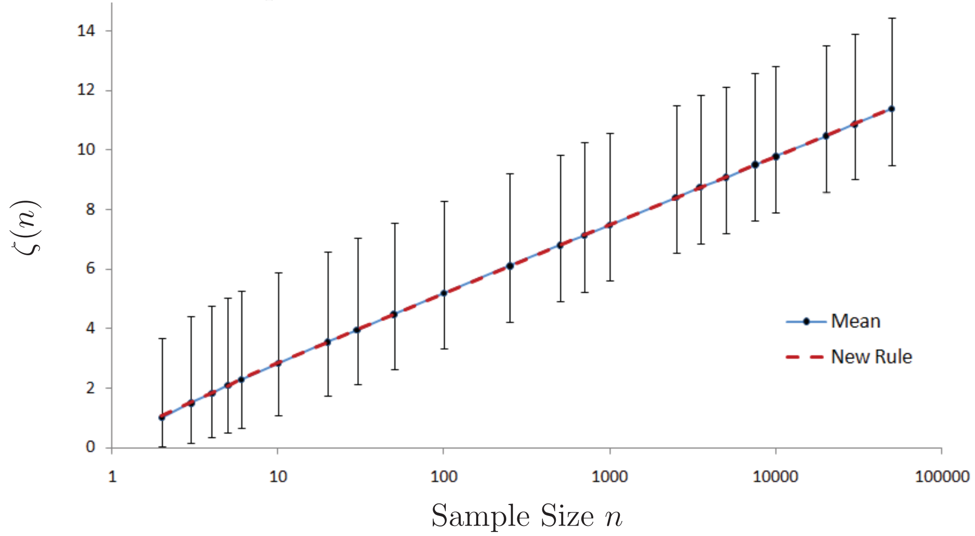
Empirical Rule of Thumb 5.2. *For samples of size n from an exponential distribution:*

$$\sigma \approx \frac{\text{range}}{\ln(n-1) + \gamma + \frac{1}{2n-2}}.$$

The term $\frac{1}{2n-2}$ could be left off for all but the smallest values of n (say $n \leq 10$). We had been very pleased with the fit of our previous rule of thumb (see Figure 3) until we graphed this one (see Figure 5).

Finally, using the first quasi-range could be more resistant to outliers (which is especially important for distributions with infinite support like the exponential), so we suggest:

Figure 5: Mean, 2.5% and 97.5% percentiles of ζ by the size of samples which were taken from an exponential distribution



Empirical Rule of Thumb 5.3. *For samples of size n from an exponential distribution:*

$$\sigma \approx \frac{\text{first quasi-range}}{\ln(n{-}2) + (\gamma{-}1)}.$$

5.3 Normal Distribution

The standard normal distribution is defined by:

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right),$$

so

$$P(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-w^2/2} dw.$$

To analyze the normal distribution as we did the other distributions above we would need to integrate $wf(w)$ using the p.d.f. $f(w)$ from Equation 5.1. Finding $f(w)$ alone requires us to integrate powers of the integral for $P(x)$ above. At the time of writing this paper we are still working with Maple to try and approximate these values of $E(W_r)$ using Maple's built-in error function.

6 Conclusion

[redacted]

[redacted]

References

[1] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, Dover Publications, Inc., New York, 1965.

[2] H. A. David, *Order Statistics*, John Wiley & Sons, New York, 1970.

[3] H. L. Harter and N. Balakrishnan, *CRC Handbook of Tables for the Use of Order Statistics in Estimation*, CRC Press, New York, 1996. ISBN 0-8493-9452-X.

[4] R. V. Hogg and E. A. Tanis, *Probablilty and Statistical Inference*, 8th ed., Prentice Hall, New Jersey, 2010. ISBN 0-321-58475-9-X.

[5] R. A. Johnson, *Probability and Statistics for Engineers*, 7th edition, Pearson Prentice Hall, Upper Saddle River, NJ, 2005. ISBN 0-13-143745-X.

[6] M. F. Triola, *Elementary Statistics*, 11 ed., Pearson Education, Boston, 2010. ISBN 0-321-50024-5.

[7] E. W. Weisstein, *CRC Concise Encyclopedia of Mathematics*, 2nd ed., CRC Press, Boco Raton, Florida, 2002.

[8] "Maple 14" (program), `http://www.maplesoft.com`, Waterloo Maple Inc., Waterloo ON, Canada,