Capstone Project

Richard Ambler

Mathematics, Thomas Edison State University

LIB-4950: Liberal Arts Capstone

Dr Denyse Lemaire-Ronveaux

September 2024

Exploring the viability of the RMSE associated with a QQ-plot as the basis for a formal normality test

ABSTRACT

The ability to test the assumption of normality for an unknown population based on information extracted from a drawn sample is important in a broad range of statistics and data science applications. This paper proposes a normality test based on the root mean squared error (RMSE) associated with a quantile-quantile (QQ) plot for a sample, and compares the test in terms of statistical power to a selection of existing normality tests over a range of contexts. The RMSE test generally compares favorably with existing normality tests and, given its natural association with the QQ-plot and simple interpretation, has the potential to augment the QQ-plot as a tool for testing the normality assumption.

KEY TERMS

Inference; Normality Test; QQ-Plot; RMSE; Sampling Distribution; Statistical Power; Monte Carlo Simulations; Log-Normal Distribution

THANKS

I would like to thank Dr Denyse Lemaire-Ronveaux for her supervision, feedback, suggestions and support during this project.

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

INTRODUCTION

A normality test uses information from a sample of data to test the hypothesis that the population from which the sample was drawn is normally distributed. Tests for normality are important in statistics; for example, many inference tests, such as the *t* -test and analysis of variance (ANOVA) tests, are based on the normality assumption, and results of such tests may be incorrect if this assumption is violated (Thode, 2002 pp. 2—3). Normality tests are also important in data science; for example, we often check errors produced by a numeric predictor for normality because if the errors are not normally distributed, it suggests an opportunity to take the distribution of errors into account to improve the predictions (Marden, 2004).

Quantile-quantile (QQ) plots are often used as stand-alone, rudimentary visual tests for normality (Thode, 2002 p. 21; Marden, 2004; Wilk & Gnanadesikan, 1968). This paper presents a formal normality test based on the sampling distribution of the *root-mean-squared-error* (RMSE) associated with a sample's QQ-plot that may augment the QQ-plot in this use case.

The first section of this report, *the RMSE normality test proposed*, discusses the terminology used, explains how an RMSE statistic can be associated with a sample's QQ-plot using an extended example of an arbitrary sample drawn from an unknown population, presents the results of Monte Carlo simulations conducted to explore the range of RMSE statistic values we might expect under the normality hypothesis, uses

simulation results to perform inference test of the normality hypothesis for the given sample, and generates results that may be used to make similar inferences for samples of up to size 50.

The second section, *fitting models to RMSE statistic sampling distributions*, discusses how we can use maximum likelihood considerations to fit simulated RMSE statistics to sampling distribution models, which eliminates the need for simulations for performing the inference test, generates the respective parameters of models for samples up to size 50, and reproduces the results from the previous section directly from the models.

The third section, *assessing the performance of the RMSE test*, discusses how we can use statistical power as a criterion to assess the performance of the RMSE test, explains how we can make use of Monte Carlo simulations to estimate the power of the test for a given sample size, non normal population and significance level.

The final section, *performance comparisons to existing tests*, compares the power of the RMSE test to that of existing normality tests by presenting estimates in contexts explored in selected published papers along with the respective power of other tests presented in those papers.

The paper concludes with a discussion of how well the proposed test performs compared to existing normality tests and potential advantages and disadvantages the test may have in practice.

Appendix A, which presents the definitions of statistics used in other normality tests considered in this paper, and appendix B, which presents a Python implementation of the RMSE test developed in conjunction with this paper, appears at the end.

THE RMSE NORMALITY TEST PROPOSED

The *normal probability distribution* is used to model a wide range of measures and phenomena found in nature. It has two parameters, namely *mean* μ and *standard deviation* σ, and probability density function:

$$f(x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

The *standard* normal distribution is the normal distribution with parameter settings $\mu = 0$ and $\sigma = 1$, in which case the distribution becomes:

$$f(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$$

We can model any normally distributed population with the standard normal distribution if we first map the data values to z-scores; that is, if we first transform the data to the respective number of standard deviations above the mean. Specifically, we can model a normal population with mean μ and standard deviation σ using the standard normal distribution if we map each element x_k in the population to:

.

$$z_k = \frac{x_k - \mu}{\sigma}$$

Where:

$$\mu = \frac{1}{n}\sum_{k=0}^{n-1} x_k$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{k=0}^{n-1} \left(x_k - \mu\right)^2}$$

In practice, we generally do not have access to the population from which a sample is drawn. In this case, we use the sample mean $\bar{x}$ to estimate μ, and the sample standard deviation s to estimate σ, in which case the z-score for each element x_k in a sample is calculated using:

$$z_k = \frac{x_k - \bar{x}}{s}$$

Where:

$$\bar{x} = \frac{1}{n}\sum_{k=0}^{n-1} x_k$$

$$s = \sqrt{\frac{1}{n-1}\sum_{k=0}^{n-1} \left(x_k - \bar{x}\right)^2}$$

The denominator $n - 1$ is used to address bias introduced by using a sample standard deviation to estimate the standard deviation of the population the sample was drawn from.

In pseudocode, we could express the mapping of data to z-scores as follows.

```
function Z_SCORES
  input: sample, array of numerical values

  m ← MEAN(sample)
  s ← STANDARD_DEVIATION(sample)
  return SORTED(MAP(x → (x - m) / s for x in sample))
```

As an example, consider the following arbitrary sorted sample of data drawn from a population with an unknown distribution.

1.09 1.17 1.26 1.66 1.72 1.88 1.89 1.90 2.05 2.64 4.21 4.62 5.31 7.59 7.77

The mean of this sample of size $n = 15$ is $\bar{x} \approx 3.12$ and the standard deviation is $s \approx 2.26$. We can normalize the data by allocating to each datum a z-score. For example, the z-score associated with the first two elements in the sample are:

$$z_0 = \frac{x_0 - \bar{x}}{s} = \frac{1.09 - 3.12}{2.26} - 0.90$$

$$z_1 = \frac{x_1 - \bar{x}}{s} = \frac{1.17 - 3.12}{2.26} = -0.86$$

Continuing, we get the z-scores for each element in the data set shown in table 1.

.

k	x_k	$z_k = \frac{x_k - \bar{x}}{s}$
0	1.09	-0.90
1	1.17	-0.86
2	1.26	-0.82
3	1.66	-0.65
4	1.72	-0.62
5	1.88	-0.55
6	1.89	-0.54
7	1.90	-0.54
8	2.05	-0.47
9	2.64	-0.21
10	4.21	0.48
11	4.62	0.67
12	5.31	0.97
13	7.59	1.98
14	7.77	2.06

Table 1: The z-scores associated with the example sample.

Notice that these z-scores are calculated directly from the data in the sample: no assumptions about the distribution of the underlying population are necessary. If we *do* assume that the underlying population is normally distributed, however, we have an alternative way of associating a z-score with each value in the sample, which we consider next.

The *cumulative probability distribution function* Φ of the standard normal probability distribution returns the cumulative probability bound by its argument value; that is:

$$\Phi(z) = \lim_{k \to -\infty} \int_k^z f(t)\, \mathrm{d}t = \frac{1}{\sqrt{2\pi}} \lim_{k \to -\infty} \int_k^z e^{-\frac{t^2}{2}}\, \mathrm{d}t$$

The *inverse cumulative probability distribution* function thus maps a cumulative probability to its associated upper bound of the cumulative probability distribution; that is:

$$\Phi^{-1}(p) = z \quad \Leftrightarrow \quad \Phi(z) = p$$

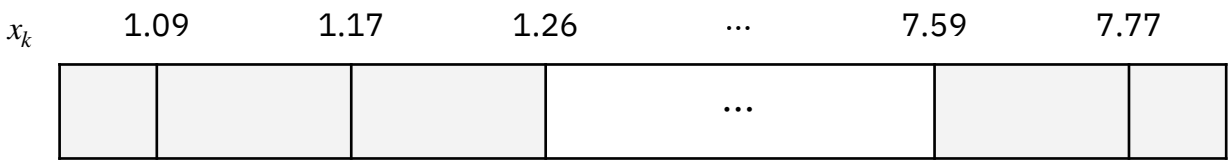
Most statistical software libraries contain a definition of f and implement algorithms to calculate, Φ and Φ^{-1}; the core Python library, *statistics*, for example contains implementations. Such algorithms are not discussed in this paper and the availability of these functions is assumed.

Φ^{-1} can be used to generate z-scores based on the position k of a datum x_k in a sorted data set under the assumption that the data was drawn from a normally distributed population; if a proportion p_k of the data in the normally distributed population lies before element x_k then the z-score associated with x_k is expected to be $\Phi^{-1}(p_k)$.

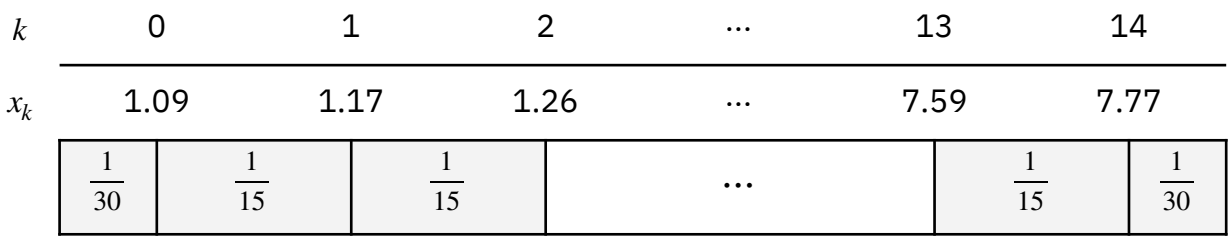
One way in which we could estimate the proportion of data in the population less than each element in the sorted sample is to spread the data evenly and center the set over

a probability of 1. This “midpoint” estimation is often used in the context of fitting parametric models, which we do shortly, to estimates obtained from data (Meeker & Escobar, 1998, pp. 129—130).

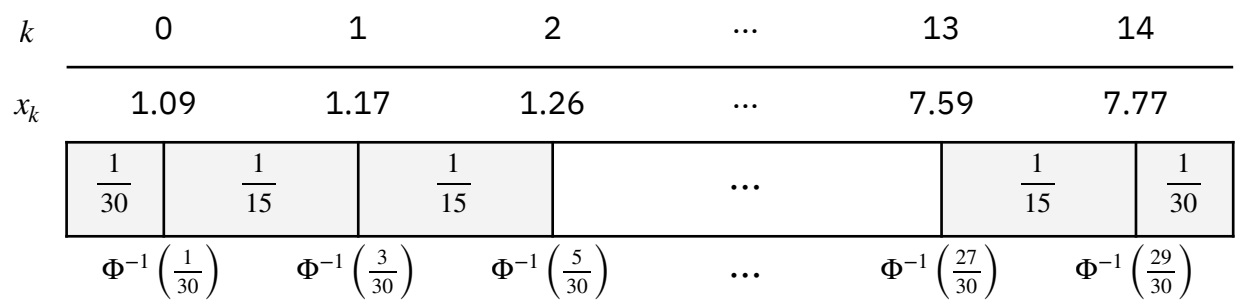
To illustrate this, consider the sorted sample presented earlier. Spreading the elements in the sample evenly and centering them over probability 1 can be illustrated as follows.



Since there are $n = 15$ elements in the sample, the “gaps” between the evenly spread out elements represent $\frac{1}{15}$ of the probability, and, since the elements are centered over the probability, half of the remaining $\frac{1}{15}$ lies before the first element and the remaining half lies after the final element, as illustrated below.



In that case, the estimated proportion of the population less than the element at index k is given by $p_k = \frac{2k+1}{2n}$. Then, if the assumption that the sample has been drawn from a normally distributed population were true, we would expect a z-score of $\hat{z}_k = \Phi^{-1}(p_k)$ associated with each element x_k, as illustrated below. (The "hat" notation above the z is simply to distinguish this z-score from that calculated earlier, and to suggest that this score is a prediction based on an assumption.



In pseudocode, we could express the calculation of these theoretical z-scores under the assumption that the sample was drawn from a normally distributed population as follows.

```
function THEORETICAL_Z_SCORES
  input: sample, array of numerical values

  n ← LENGTH(sample)
  ps ← MAP(k → (2k + 1) / (2n) for k from 0 to n - 1)
  return MAP(p → INVERSE_NORMAL_CDF(p) for p in ps)
```

To continue the example, the theoretical z-score associated with the first two elements in the data set are:

.

$$\hat{z}_0 = \Phi^{-1}(p_0) = \Phi^{-1}\left(\frac{1}{30}\right) = -1.83$$

$$\hat{z}_1 = \Phi^{-1}(p_1) = \Phi^{-1}\left(\frac{1}{30} + \frac{1}{15}\right) = -1.28$$

Continuing, we get the theoretical z-scores for each element in the data set shown in table 2.

k	x_k	$p_k = \frac{2k+1}{2n}$	$\hat{z}_k = \Phi^{-1}(p_k)$
0	1.09	0.033	-1.83
1	1.17	0.100	-1.28
2	1.26	0.167	-0.97
3	1.66	0.233	-0.73
4	1.72	0.300	-0.52
5	1.88	0.367	-0.34
6	1.89	0.433	-0.17
7	1.90	0.500	0.00
8	2.05	0.567	0.17
9	2.64	0.633	0.34
10	4.21	0.700	0.52
11	4.62	0.767	0.73
12	5.31	0.833	0.97
13	7.59	0.900	1.28
14	7.77	0.967	1.83

Table 2: The theoretical z-scores under the assumption that the sample has been drawn from a normally distributed population for the example sample.

To generalize, given a sorted sample of length n, $\mathrm{X} = \left\{x_0,\ x_1,\ \ \ldots,\ x_{n-1}\right\}$, we can get two z-scores associated with each element x_k, namely the actual z-score $z_k = \frac{x_k - \bar{x}}{s}$ and an estimated, theoretical z-score $\hat{z}_k = \Phi^{-1}\left(\frac{2k+1}{2n}\right)$ that we would expect under the assumption that the sample was drawn from a normally distributed population.

In this paper, we treat the theoretical z-scores obtained through consideration of the position of each datum in the sorted sample as *predictions* of the actual z-scores obtained from the values themselves, and use the performance of these predictions to make inferences about the assumption the predictions are based on, namely that the population from which the sample has been drawn is normally distributed. Let us refer to this normality hypothesis as H_0 from here on.

A quantile-quantile (QQ) plot is obtained if we plot the z-scores against the theoretical z-scores. For example, the QQ-plot associated with the example sample is given in figure 1.

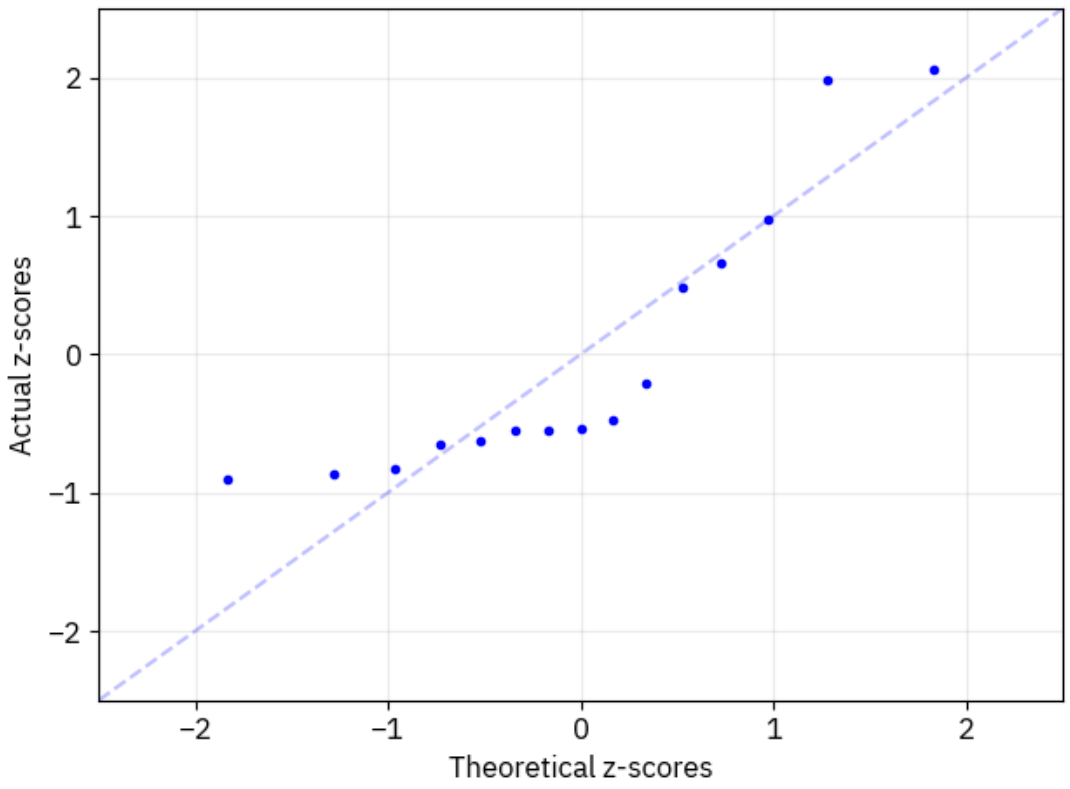


Figure 1: The QQ-plot associated with the example sample.

If H_0 were true, we would expect the predicted z-scores to be similar to the actual z -scores obtained from the data in the sample; in our worked example, we would expect the points in figure 1 to approximately align with the diagonal. This is, indeed, a common use case of the QQ-plot — that is, as an informal, visual test for the normality assumption (Thode, 2002 p. 21; Marden, 2004). Further, if we treat $\hat{z}_k$ as a prediction of z_k for each element in the data set, there exists an implied set of errors, $\varepsilon_k = \hat{z}_k - z_k$ associated with the points in the QQ-plot.

The *root mean squared error* (RMSE) is a measure of how poorly a set of numerical predictions are. Conceptually, the RMSE is analogous to an average of the errors

produced; to ensure the errors do not cancel out, however, we calculate the average of the squared errors and then take the square root of that average to return the units to that of the original data. That is:

$$\text{RMSE} = \sqrt{\frac{1}{n-1}\sum_{k=0}^{n-1}\varepsilon_k^2} = \sqrt{\frac{1}{n-1}\sum_{k=0}^{n-1}\left(\hat{z}_k - z_k\right)^2}$$

Notice, however, that a denominator $n-1$ has been used; this is to avoid bias in the estimate of the population RMSE based on a sample, analogous to how a denominator of $n-1$ is used to estimate a populations's standard deviation. After all, the standard deviation is ultimately an RMSE measure in which the sample mean is used as the prediction model.

In pseudocode, an RMSE statistic associated with the data in a sample could thus be calculated as follows.

```
function RMSE_STATISTIC
  input: sample, array of numerical values

  n ← LENGTH(sample)
  z_scores ← Z_SCORES(sample)
  theoretical_z_scores ← THEORETICAL_Z_SCORES(sample)

  return SQRT(
    SUM(
      MAP(
        k → (theoretical_z_scores[k] - z_scores[k])^2)
        for k from 0 to n - 1
      )
    ) / (n - 1)
  )
```

To continue with the previous example, table 3 shows the squared errors associated with each point in the QQ-plot; based on these values, the RMSE associated with this QQ-plot turns out to be around 0.449.

k	x_k	z_k	$\hat{z}_k$	$\varepsilon_k^2 = \left(\hat{z}_k - z_k\right)^2$
0	1.09	-0.90	-1.83	0.874
1	1.17	-0.86	-1.28	0.175
2	1.26	-0.82	-0.97	0.021
3	1.66	-0.65	-0.73	0.007
4	1.72	-0.62	-0.52	0.009
5	1.88	-0.55	-0.34	0.043
6	1.89	-0.54	-0.17	0.142
7	1.90	-0.54	0.00	0.291
8	2.05	-0.47	0.17	0.411
9	2.64	-0.21	0.34	0.305
10	4.21	0.48	0.52	0.002
11	4.62	0.67	0.73	0.004
12	5.31	0.97	0.97	0.000
13	7.59	1.98	1.28	0.492
14	7.77	2.06	1.83	0.052

Table 3: The squared errors associated with each point in the example sample.

Of course, even if the null hypothesis H_0 were true, we would not expect a zero RMSE statistic value because of random effects: especially for small samples, a sample is never expected to be perfectly representative of the population it is drawn from unless it comprises the entire population, which is not possible for a finite sample drawn from

a theoretical normal distribution. For the RMSE statistic to be helpful, then, we need to know how the RMSE statistic is distributed for a given sample size when H_0 is true, which would provide a context in which we could make judgements about the value obtained for a specific sample.

In this paper, we explore such sampling distributions using Monte Carlo simulations. Specifically, for a given sample size n, we:

1. Draw a random sample from a population we know to be normally distributed, that is, within a context in which H_0 is true.
2. Calculate and record the RMSE statistic for the sample.
3. Repeat from step 1 a large number of times.

This provides us with a large set of RMSE statistic values that we would expect under the hypothesis H_0, and this sampling distribution provides us with a context in which to assess the RMSE statistic generated from the actual sample.

The pseudocode for generating a random sample of data could be expressed as follows.

```
function RANDOM_SAMPLE
  input: sample_size, a sample size
         POPULATION,   a random value generator

  return MAP(_ → POPULATION() for _ from 1 to sample_size)
```

Here, POPULATION is a pseudorandom function that returns a random element drawn from a population of interest.

Then, the Monte Carlo simulations could be described in pseudocode as follows.

```
function MONTE_CARLO_DISTRIBUTION
  input: n,           sample size
         simulations, number of simulations to perform

  return SORTED(
    MAP(
      _ → RMSE_STATISTIC(RANDOM_SAMPLE(n, NORMAL(mu: 0, sigma: 1)))
      for _ from 1 to simulations
    )
```

Here, NORMAL(mu: 0, sigma: 1) is a pseudorandom function that draws a random element from a normally distributed population with mean 0 and standard deviation 1, that is, from the standard normal distribution, and *simulations* is a large number. Most statistics software packages provide an implementation of NORMAL(mu: 0, sigma: 1), but, historically, investigators have used published lists of normally distributed random numbers; for example, Shapiro, Wilk & Chen (1968) and Filliben (1975) used normally distributed data published in RAND (1955).

To continue the example, figure 2 shows a histogram of 100,000 simulated RMSE values calculated for random samples of size 15 drawn from a normally distributed population.

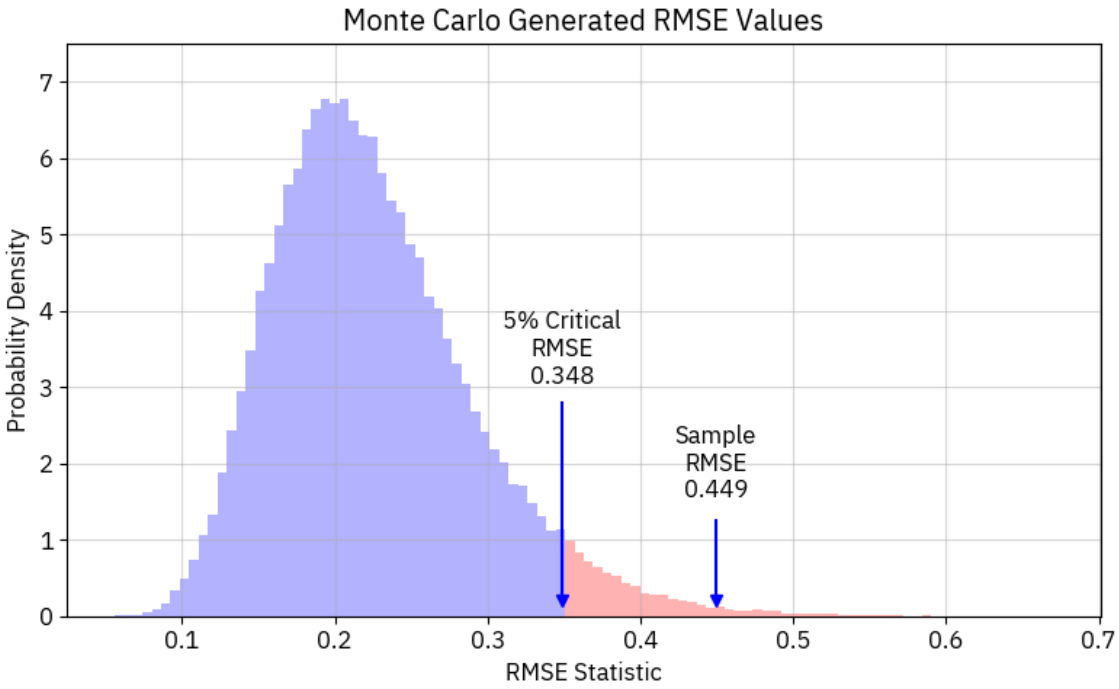


Figure 2: A histogram of 100,000 RMSE statistic values generated by repeatedly drawing samples of size $n = 15$ from a normally distributed population, with the 5% critical RMSE value and the sample RMSE statistic value shown.

We notice from figure 2 that the RMSE value we obtained for the example sample, namely 0.449, falls far to the right of the sampling distribution of RMSE values we would expect under H_0.

When performing inference, we effectively use considerations of how rare the observed statistic value is under the normality hypothesis. One way in which we can do this is by considering *the proportion of the sampling distribution RMSE values under* H_0 *that are at least as extreme as the observed RMSE value*, which is referred to as the p-value.

In pseudocode, this could be calculated as follows.

```
function P-VALUE
  input: distribution, ordered set of RMSE statistics
         statistic,    RMSE statistic of sample under question

  n ← LENGTH(distribution)

  return LENGTH(FILTER(x → x > statistic) for x in distribution) / n
```

If we perform this calculation for the RMSE value we obtained from the example sample, namely 0.449, we get a p-value of less that 1%, which suggests that an RMSE value this extreme is very unlikely if H_0 is true.

Alternatively, we could choose a significance level α, say 5%, and find the cut-off RMSE value, referred to as a *critical value*, for that proportion of the most extreme RMSE values under H_0. This is equivalent to finding the $1 - \alpha$ quantile in the sampling distribution, which we can get using simple linear interpolation.

In pseudocode, the calculation might be expressed as follows.

```
function QUANTILE
  input: distribution, ordered set of RMSE statistics
         proportion,   desired bound proportion

  n ← LENGTH(distribution)
  last_index ← n - 1
  d ← proportion * last_index
  index ← FLOOR(d)

  if index = last_index then
    return distribution[last_index]

  run ← d - index
  rise ← distribution[index + 1] - distribution[index]

  return distribution[index] + run * rise
```

To find the critical value that cuts off the most extreme 5% of the RMSE values under H_0, for example, we would find the 95% quantile of the sampling distribution, which in this case turns out to be around 0.348, as shown in figure 2. This value provides us with a value against which we can compare the RMSE value we obtained from the sample, namely 0.449. Since 0.449 lies above the critical value of 0.348, we can conclude that this RMSE value is less likely than 5% if H_0 is true.

Table 4 shows quantile values similarly obtained, that is, from 100,000 simulated RMSE values calculated under the assumption H_0, for samples sized from 3 to 50, which allows us to perform the above test for a range of sample sizes and significance levels.

	Quantile				
n	50%	75%	90%	95%	99%
3	0.259	0.386	0.460	0.485	0.505
4	0.288	0.369	0.455	0.505	0.583
5	0.284	0.360	0.437	0.486	0.582
6	0.276	0.347	0.419	0.466	0.561
7	0.269	0.334	0.402	0.447	0.539
8	0.260	0.322	0.386	0.429	0.519
9	0.253	0.310	0.373	0.414	0.501
10	0.246	0.301	0.361	0.401	0.488
11	0.238	0.291	0.349	0.387	0.470
12	0.233	0.284	0.339	0.377	0.458
13	0.227	0.276	0.329	0.366	0.444
14	0.221	0.269	0.320	0.355	0.432
15	0.217	0.263	0.314	0.348	0.421
16	0.212	0.257	0.306	0.339	0.410
17	0.208	0.252	0.298	0.331	0.400
18	0.204	0.246	0.292	0.324	0.392
19	0.200	0.241	0.286	0.317	0.384
20	0.197	0.238	0.281	0.311	0.376
21	0.194	0.233	0.277	0.307	0.370
22	0.190	0.229	0.271	0.300	0.361
23	0.187	0.225	0.266	0.294	0.355
24	0.185	0.222	0.262	0.289	0.350
25	0.182	0.218	0.258	0.286	0.344
26	0.180	0.216	0.255	0.281	0.341

	Quantile				
n	50%	75%	90%	95%	99%
27	0.177	0.212	0.250	0.277	0.334
28	0.174	0.209	0.246	0.272	0.327
29	0.172	0.206	0.244	0.269	0.324
30	0.170	0.204	0.241	0.266	0.319
31	0.168	0.201	0.237	0.262	0.314
32	0.166	0.199	0.234	0.259	0.312
33	0.164	0.196	0.231	0.254	0.307
34	0.162	0.194	0.228	0.251	0.301
35	0.161	0.192	0.226	0.249	0.301
36	0.159	0.190	0.224	0.246	0.295
37	0.157	0.188	0.220	0.243	0.292
38	0.156	0.186	0.219	0.241	0.289
39	0.154	0.184	0.216	0.238	0.286
40	0.153	0.182	0.214	0.236	0.283
41	0.151	0.180	0.211	0.233	0.279
42	0.150	0.179	0.210	0.231	0.276
43	0.148	0.177	0.208	0.229	0.274
44	0.147	0.175	0.206	0.227	0.270
45	0.146	0.174	0.204	0.224	0.268
46	0.145	0.172	0.202	0.222	0.265
47	0.143	0.170	0.200	0.220	0.265
48	0.142	0.169	0.199	0.218	0.261
49	0.141	0.168	0.196	0.216	0.258
50	0.140	0.166	0.195	0.214	0.256

Table 4: RMSE statistic quantiles under the assumption H_0, based on 100,000 Monte Carlo simulations for sample sizes 3 to 50..

Whether we consider the p-value or the 5% critical value in the case of the example sample, we can treat our results as sufficient evidence to reject H_0, at least at the 5% significance level but even at the 1% significance level. In short, the RMSE value associated with a QQ-plot can be used as a statistic that can be used as the basis for a formal normality test.

Ideally, however, we would like to be able to model the sampling distribution of RMSE statistics under H_0 so that we do not need to rely on simulations each time we would like to perform an RMSE inference test. It turns out that the RMSE statistic sampling distribution is well modeled using a log-normal distribution, as is demonstrated next.

FITTING MODELS TO RMSE STATISTIC SAMPLING DISTRIBUTIONS

The log-normal probability density function is:

$$f(x \mid \mu, \sigma) = \frac{e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}}{x\sigma\sqrt{2\pi}}$$

Given a sample of simulated RMSE values X of size n, the *likelihood density function* L is the function that maps the sample to the likelihood that the sample was drawn from a log-normal population with parameters μ and σ, which turn out to be the mean and standard deviation respectively of the logarithms of the values in X, as we see shortly.

If the values in the sample are drawn independently, L is simply the product of the probability densities of the values in the sample:

$$\begin{aligned} L(\mathrm{X}) &= \prod_{x \in \mathrm{X}} f(x \mid \mu, \sigma) \\ &= \prod_{x \in \mathrm{X}} \frac{e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}}{x\sigma\sqrt{2\pi}} \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{x \in \mathrm{X}} \frac{1}{x} \prod_{x \in \mathrm{X}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} \end{aligned}$$

Fitting experimentally obtained values obtained from Monte Carlo simulations to a log-normal distribution, then, amounts to finding values of parameters μ and σ that maximize the value of L.

The *log-likelihood density function* λ is the function that maps sample X to the log of the likelihood that it was drawn from a log-normal population with parameters μ and σ; that is, $\lambda(\mathrm{X}) = \ln L(\mathrm{X})$. It is often more convenient to work with λ than L, both in terms of representing the values in computers, since products can grow very large or very small very quickly, as well as in terms of finding optimal model parameter values, as we will do shortly; of course, finding parameter values that optimize λ is equivalent to finding values that optimize L, since λ increases and decreases with L (Meeker & Escobar, 1998, p. 157).

In this case, the log-likelihood function is:

$$\begin{aligned}\lambda(\mathrm{X}) &= \ln L(\mathrm{X}) \\ &= \ln\left(\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{n} \prod_{x\in\mathrm{X}}\frac{1}{x} \prod_{x\in\mathrm{X}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}\right) \\ &= -n\ln\sigma - n\ln\sqrt{2\pi} - \sum_{x\in\mathrm{X}}\ln x - \frac{1}{2\sigma^2}\sum_{x\in\mathrm{X}}\left(\ln x-\mu\right)^2\end{aligned}$$

To optimize this function, we can take the partial derivatives with respect to μ and σ, set those derivatives to zero, and then solve for μ and σ.

Differentiating λ with respect to μ, we get:

$$\frac{\partial\lambda}{\partial\mu} = \frac{1}{\sigma^2}\left(\sum_{x\in\mathrm{X}}\ln x - n\mu\right)$$

Setting $\frac{\partial\lambda}{\partial\mu} = 0$ and solving for μ, we get:

$$\mu = \frac{1}{n}\sum_{x\in\mathrm{X}}\ln x$$

From this we can see that μ is the mean of the logarithms of the RMSE values generated by the Monte Carlo simulations.

Similarly, differentiating λ with respect to σ, we get:

$$\frac{\partial\lambda}{\partial\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{x\in\mathrm{X}}\left(\ln x - \mu\right)^2$$

Setting $\frac{\partial\lambda}{\partial\sigma} = 0$, we get:

$$\sigma^2 = \frac{1}{n}\sum_{x\in\mathrm{X}}\left(\ln x - \mu\right)^2$$
$$\sigma = \sqrt{\frac{1}{n}\sum_{x\in\mathrm{X}}\left(\ln x - \mu\right)^2}$$

From this we can see that σ is the population standard deviation of the logarithms of the RMSE values generated by the Monte Carlo simulations.

The pseudocode for fitting a log-normal model to the statistics generated by the Monte Carlo simulations might thus be expressed as:

```
function LOG_NORMAL_FIT
  input: values, set of Monte Carlo generated statistic values

  logs ← MAP(x → LN(x) for x in values)
  mu ← MEAN(logs)
  sigma ← STANDARD_DEVIATION(logs)

  return mu, sigma
```

Continuing with the example sample, figure 3 shows the histogram of Monte Carlo RMSE statistic values generated earlier fitted to a log-normal model; the maximum

likelihood parameters turn out to be $\mu = -1.53$ and $\sigma = 0.29$. As we can see, the model fits the histogram very well.

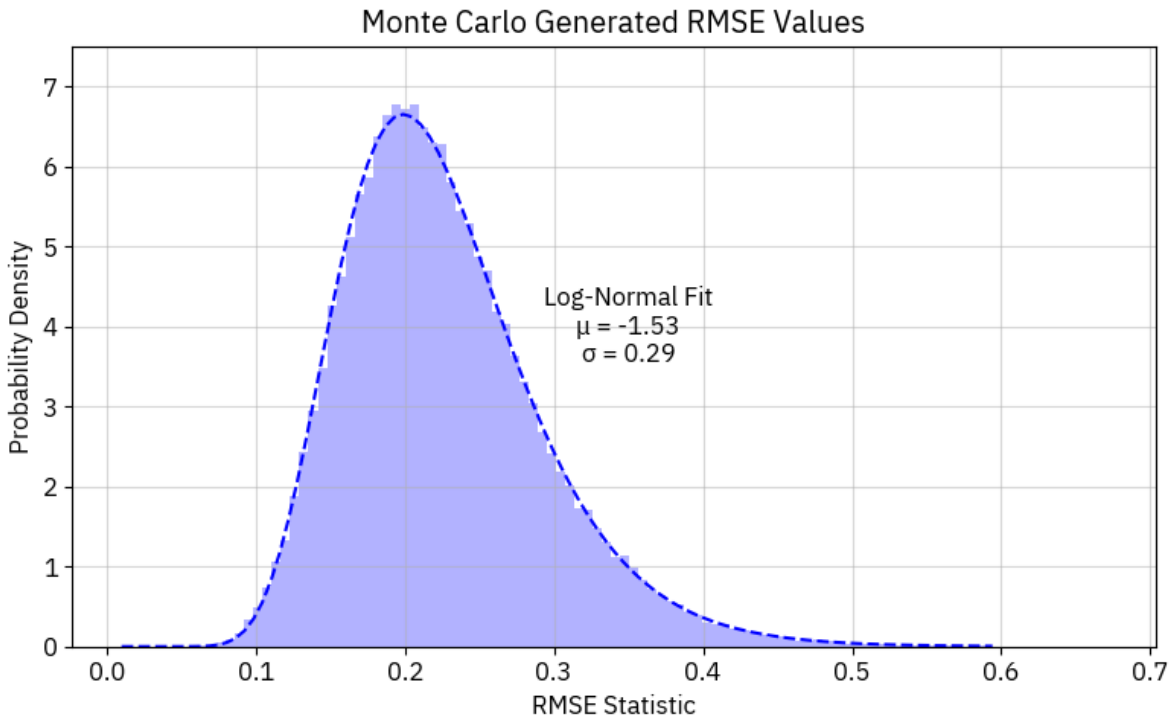


Figure 3: The distribution of Monte Carlo generated RMSE statistic values from figure 2 with a fitted log-normal model shown as a dotted curve.

Table 5 shows the maximum likelihood parameter estimates over a range of sample sizes from 3 to 50 obtained similarly; figure 4 shows the plot of the parameter estimates for sample sizes up to 200.

	Parameter Values	
n	μ	σ
3	-1.570	0.753
4	-1.348	0.518
5	-1.311	0.421
6	-1.318	0.374
7	-1.335	0.347
8	-1.361	0.332
9	-1.387	0.320
10	-1.413	0.313
11	-1.441	0.305
12	-1.461	0.299
13	-1.486	0.294
14	-1.510	0.291
15	-1.530	0.289
16	-1.550	0.285
17	-1.571	0.283
18	-1.590	0.280
19	-1.607	0.278
20	-1.624	0.276
21	-1.639	0.276
22	-1.657	0.274
23	-1.673	0.271
24	-1.688	0.270
25	-1.702	0.269
26	-1.714	0.269

	Parameter Values	
n	μ	σ
27	-1.729	0.266
28	-1.744	0.264
29	-1.756	0.266
30	-1.767	0.265
31	-1.780	0.263
32	-1.791	0.262
33	-1.805	0.262
34	-1.816	0.260
35	-1.824	0.260
36	-1.835	0.259
37	-1.846	0.258
38	-1.855	0.258
39	-1.865	0.257
40	-1.875	0.257
41	-1.885	0.255
42	-1.893	0.255
43	-1.903	0.254
44	-1.911	0.254
45	-1.921	0.253
46	-1.929	0.253
47	-1.937	0.252
48	-1.945	0.252
49	-1.954	0.251
50	-1.963	0.251

Table 5: Parameter values for log-normal models fitted to 100,000 Monte Carlo generated RMSE statistic values for each row.

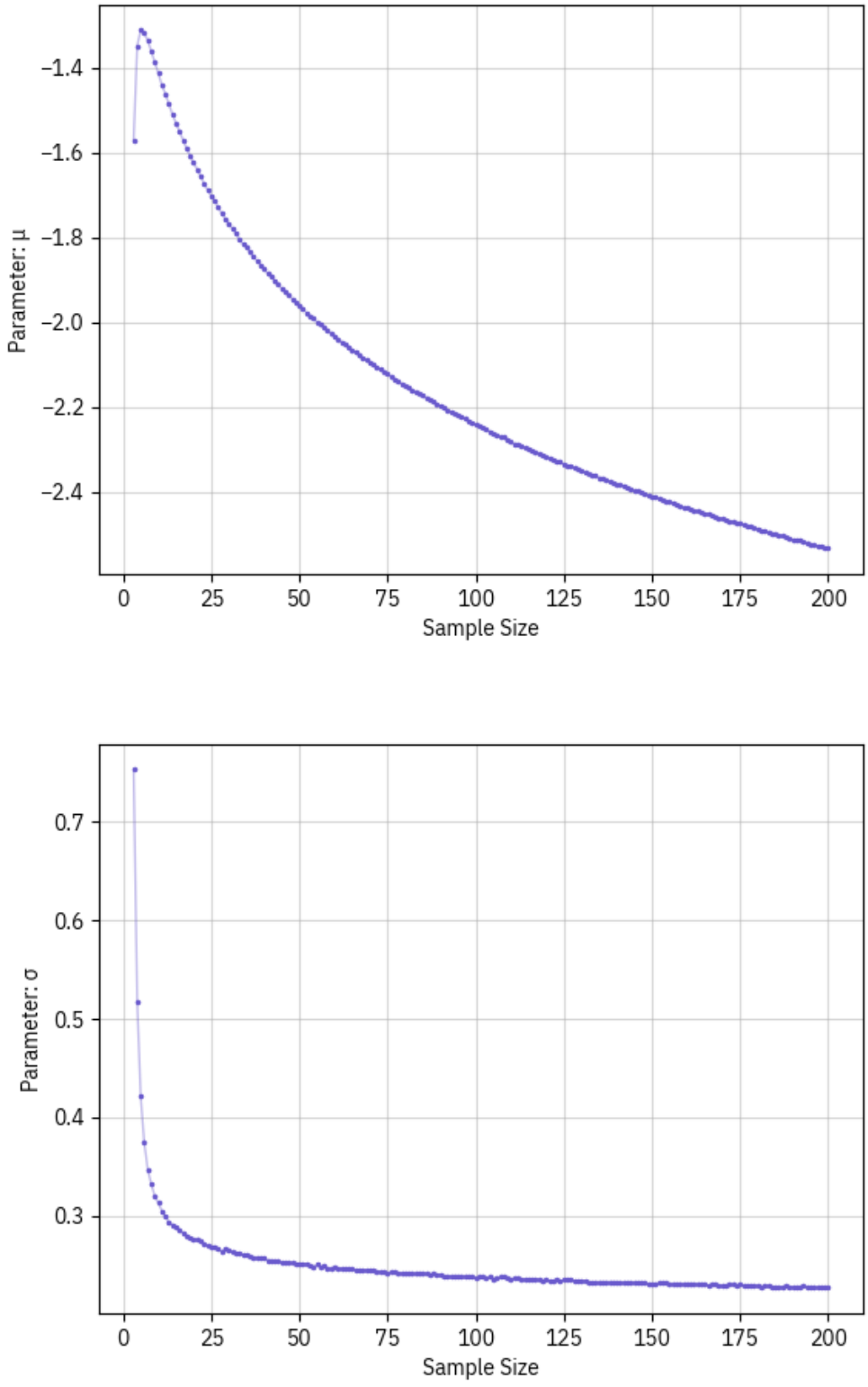


Figure 4: Parameter values for log-normal fits of the generated RMSE values.

	Quantile				
n	50%	75%	90%	95%	99%
3	0.208	0.346	0.546	0.718	1.200
4	0.260	0.368	0.504	0.609	0.867
5	0.270	0.358	0.462	0.539	0.718
6	0.268	0.344	0.432	0.495	0.639
7	0.263	0.333	0.410	0.466	0.590
8	0.256	0.321	0.392	0.442	0.554
9	0.250	0.310	0.376	0.422	0.525
10	0.243	0.301	0.363	0.407	0.504
11	0.237	0.291	0.350	0.391	0.482
12	0.232	0.284	0.341	0.380	0.466
13	0.226	0.276	0.330	0.367	0.449
14	0.221	0.269	0.321	0.357	0.435
15	0.217	0.263	0.314	0.348	0.424
16	0.212	0.257	0.306	0.339	0.412
17	0.208	0.251	0.299	0.331	0.401
18	0.204	0.246	0.292	0.323	0.391
19	0.200	0.242	0.286	0.317	0.383
20	0.197	0.238	0.281	0.311	0.375
21	0.194	0.234	0.276	0.306	0.369
22	0.191	0.229	0.271	0.299	0.361
23	0.188	0.225	0.266	0.293	0.353
24	0.185	0.222	0.261	0.288	0.347
25	0.182	0.219	0.257	0.284	0.341
26	0.180	0.216	0.254	0.280	0.336

	Quantile				
n	50%	75%	90%	95%	99%
27	0.178	0.212	0.250	0.275	0.329
28	0.175	0.209	0.245	0.270	0.324
29	0.173	0.207	0.243	0.267	0.320
30	0.171	0.204	0.240	0.264	0.316
31	0.169	0.201	0.236	0.260	0.311
32	0.167	0.199	0.233	0.257	0.307
33	0.164	0.196	0.230	0.253	0.302
34	0.163	0.194	0.227	0.250	0.298
35	0.161	0.192	0.225	0.247	0.295
36	0.160	0.190	0.222	0.244	0.291
37	0.158	0.188	0.220	0.241	0.287
38	0.156	0.186	0.218	0.239	0.285
39	0.155	0.184	0.215	0.236	0.281
40	0.153	0.182	0.213	0.234	0.279
41	0.152	0.180	0.210	0.231	0.275
42	0.151	0.179	0.209	0.229	0.273
43	0.149	0.177	0.207	0.227	0.270
44	0.148	0.175	0.205	0.224	0.267
45	0.146	0.174	0.203	0.222	0.264
46	0.145	0.172	0.201	0.220	0.262
47	0.144	0.171	0.199	0.218	0.259
48	0.143	0.169	0.197	0.216	0.257
49	0.142	0.168	0.195	0.214	0.254
50	0.141	0.167	0.194	0.213	0.253

Table 6: RMSE statistic quantiles under the assumption H_0, based on the fitted log-normal RMSE sampling distribution for sample sizes 3 to 50.

With a model for the RMSE statistic sampling distribution, we no longer need to perform Monte Carlo simulations to determine critical RMSE values or p-values associated with a given sample; rather, we can simply use the respective log-normal model for the given sample size. Table 6 shows the same quantiles as shown in table 4, but obtained from the fitted log-normal models instead of directly from Monte Carlo simulations.

Now that we have a formal test for normality, the next consideration is how well it performs as a statistical test, which we address next.

ASSESSING THE PERFORMANCE OF THE RMSE TEST

Notice that there are two possible outcomes for the test described in the previous section, namely to accept H_0 or to reject H_0. There are thus two ways in which we may err based on the results of the test, namely, to reject H_0 when H_0 is actually true, called *type I error*, and to accept H_0 when H_0 is actually false, called *type II error*. Similarly, there are two ways in which the test may yield correct results, namely, if it results in us accepting H_0 when H_0 is actually true, and rejecting H_0 when H_0 is actually false.

When performing a test, we typically *choose* the type I error rate, that is $\Pr\left(\text{reject } H_0 \mid H_0\right)$, that we are willing to tolerate, as the *significance level* of the test. Previously, for example, we chose a 5% significance level and rejected H_0 since the

RMSE statistic obtained from the example sample was more extreme than the associated critical RMSE value; this is equivalent to choosing a 5% type I error rate we are willing to tolerate for the test, since if the sample were drawn from a normally distributed population, we would expect 5% probability of getting as extreme a value, in which case we would incorrectly reject H_0 based on the test results.

The probability with which we correctly accept H_0, called the *specificity* of the test, is the complement of the type I error rate, that is:

$$\begin{aligned} \text{specificity} &= \Pr\left(\text{accept } H_0 \mid H_0\right) \\ &= 1 - \text{type I error rate} \\ &= 1 - \Pr\left(\text{reject } H_0 \mid H_0\right) \end{aligned}$$

Thus, choosing a 5% significance level, for example, which, as already mentioned is equivalent to choosing a 5% type I error rate we are willing to tolerate, is also equivalent to choosing a 95% desired specificity.

The probability of a type II error, that is, $\Pr\left(\text{accept } H_0 \mid \text{not } H_0\right)$, on the other hand, is more difficult to determine a priori, because it depends not only on the chosen significance level, but on the unknown population distribution. The probability of correctly rejecting H_0, called the *power* of the test, is the complement of the type II error rate, that is:

$$\begin{aligned}\text{power} &= \Pr\left(\text{reject } H_0 \mid \text{not } H_0\right) \\ &= 1 - \text{type II error rate} \\ &= 1 - \Pr\left(\text{accept } H_0 \mid \text{not } H_0\right)\end{aligned}$$

As such, the power of the test, like the type II error rate, is contextual and difficult to determine a priori. What researchers tend to to when evaluating the performance of statistical tests, as we will see later on, is assess the power of a test over a range of contexts with specified sample sizes, non normal population distributions, and significance levels.

To estimate the power of the RMSE test for a specified sample size, underlying non normal population distribution, and significance level, we can once again resort to Monte Carlo simulations by following the steps:

1. Draw a random sample of the given sample size from from the given non normal population.
2. Calculate the RMSE statistic for the sample.
3. Use the RMSE test to accept or reject H_0 by comparing the RMSE statistic to the respective critical value for the given sample size and significance level.
4. Repeat from step 1 a large number of times.
5. Calculate the proportion of samples for which H_0 was correctly rejected as an estimate of the power of the test in the given context.

In pseudocode, this algorithm might be expressed as follows.

```
function MONTE_CARLO_POWER
  input: sample-size,  size of sample to be drawn from population
         POPULATION,    function to generate random value
         cutoff,       critical value
         experiments,  number of experiments to perform

  successes ← 0

  repeat experiments times
    sample ← RANDOM_SAMPLE(sample-size, POPULATION)
    statistic ← RMSE_STATISTIC(sample)
    if statistic > cutoff then successes += 1

  return successes / experiments
```

For example, to assess the power of the RMSE test for a sample size of 15, a non normal population with a t distribution with 1 degree of freedom, and a significance level of 5%, we would set *sample-size* to 15, POPULATION to a pseudorandom function that returns a random element from the given t distribution, and *cutoff* to 0.348, obtained from table 6. It turns out that the power of the test in this context is about 0.83; in other words, the test is able to correctly reject H_0 in this context with a probability of 0.83.

In the next section we compare the power of the RMSE test to that of existing normality tests over a range of contexts explored in selected studies.

PERFORMANCE COMPARISONS TO EXISTING TESTS

As mentioned in the previous section, the power of a normality test depends in part on the distribution of the underlying non normal population, and it is a common practice, when comparing normality tests in terms of power, for authors to choose a range of non normal populations over which to perform the tests and assess their power. In this section, we consider four pertinent power comparison studies and investigate how well the RMSE test performs in the contexts presented in those studies.

As with the RMSE test, the tests considered in this section are based on respective sampling distributions of defined statistics that are influenced by the distribution of the population a sample is drawn from. While the definitions of the statistics tests in the studies investigated are not explicitly considered in this paper, they are presented from sources in appendix A.

The basic steps followed for each of the four studies are:

- From each study, a selection of normality tests, sample sizes and non normal distributions are selected as contexts

- For each context, the power of the RMSE test is estimated using the MONTE_CARLO_POWER algorithm described earlier, using 100,000 simulations. Practically, random value generators from the *random* module of the *Numpy* Python

package (Harris et al, 2020), which provides pseudorandom generators for a wide range of probability distributions, are used for the POPULATION argument.

- The RMSE power estimates are presented along with respective power estimates reported in the study.

In addition, a sketch of the non normal population distribution is included with a normal distribution with the same mean and standard deviation overlaid to illustrate how dissimilar the non normal populations are from analogous normal populations.

Study 1: Shapiro, Wilk & Chen (1968)

Shapiro, Wilk & Chen (1968) categorized non normal populations into five groups based on symmetry, tail length and similarity to the normal distribution, and then used normal points published by RAND (1955) to generate Monte Carlo estimates to compare the power of several tests at the 10% significance level on a selection of non normal populations representative of each group, on sample sizes 10, 15, 20, 35 and 50. The tests that they compared in terms of power include their own *Shapiro-Wilk* (SW) test, the *Kolmogorov-Smirnov* (KS) test, and the *weighted Cramer-von Mises* (CM) test, based on sampling distributions of statistics shown in table 7, and the non normal population distributions tested on include the *log-normal*, *Weibull, chi-squared*, *beta* and *t* distributions.

A selection of the results presented by Shapiro, Wilk & Chen for each of the categories identified in the study, along with the power of the RMSE test in analogous contexts, are shown in tables 7—11.

We notice that the power of the RMSE test roughly matches that of the SW test for asymmetric long-tailed and near normal populations, is somewhat less powerful than the SW test for asymmetric short-tailed populations, but still more powerful than KS or CM, and is less powerful than KS and CM for symmetric long-tailed populations, though it outperforms these tests for populations in this category that are more similar to the normal population.

Group 1: Asymmetric Long-Tailed Non Normal Populations

n	RMSE Test	SW	KS	CM
Log-Normal				
10	0.68	0.72	0.28	0.34
15	0.87	0.88	0.47	0.43
20	0.95	0.96	0.57	0.81
35	1.00	1.00	1.00	0.98
50	1.00	1.00	1.00	1.00
Weibull (0.5)				
10	0.93	0.94	0.58	0.63
15	0.99	1.00	0.65	0.77
20	1.00	1.00	1.00	0.99
35	1.00	1.00	1.00	1.00
50	1.00	1.00	1.00	1.00
Chi-Squared (1)				
10	0.80	0.82	0.41	0.42
15	0.95	0.94	0.51	0.48
20	0.99	0.99	0.58	0.72
35	1.00	1.00	1.00	0.93
50	1.00	1.00	1.00	1.00
Chi-Squared (4)				
10	0.33	0.33	0.20	0.14
15	0.49	0.43	0.22	0.16
20	0.63	0.65	0.26	0.23
35	0.88	0.84	0.33	0.34
50	0.97	0.97	0.44	0.47

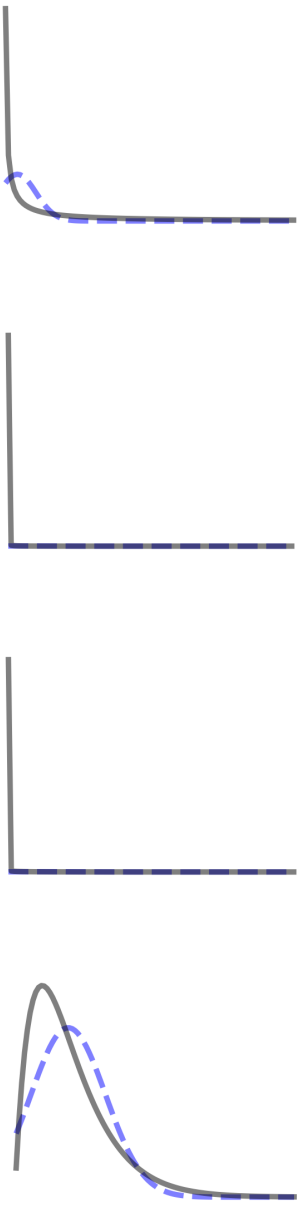
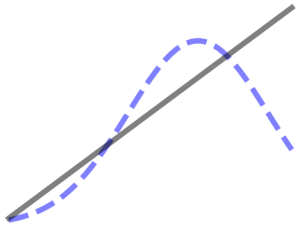


Table 7: The 10% significance power of the RMSE test to discriminate against asymmetric long-tailed non normal populations, compared to analogous power of other tests as reported by Shapiro, Wilk & Chen (1968).

Group 2: Asymmetric Short-Tailed Non Normal Populations

n	RMSE Test	SW	KS	CM
Beta (2, 1)				
10	0.19	0.23	0.10	0.17
15	0.28	0.30	0.10	0.13
20	0.38	0.47	0.14	0.20
35	0.68	0.78	0.16	0.18
50	0.88	0.96	0.22	0.28
Beta (3.2)				
10	0.09	0.16	0.12	0.12
15	0.10	0.21	0.06	0.08
20	0.11	0.19	0.08	0.12
35	0.18	0.22	0.12	0.08
50	0.26	0.42	0.10	0.11



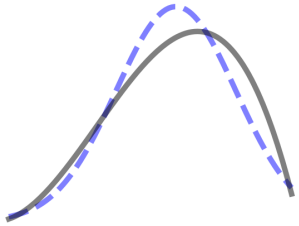
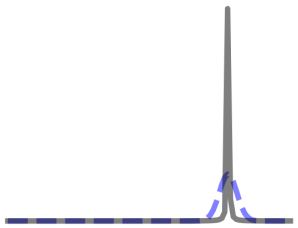


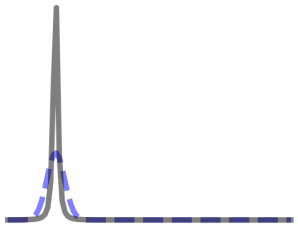
Table 8: The 10% significance power of the RMSE test to discriminate against asymmetric short-tailed non normal populations, compared to analogous power of other tests as reported by Shapiro, Wilk & Chen (1968).

.

Group 3: Symmetric Long-Tailed Non Normal Populations

n	RMSE Test	SW	KS	CM
T (1)				
10	0.68	0.42	0.30	0.95
15	0.83	0.81	0.47	0.98
20	0.92	0.92	0.65	0.99
35	0.99	0.99	0.86	1.00
50	1.00	1.00	0.95	1.00
T (2)				
10	0.40	0.41	0.15	0.63
15	0.53	0.51	0.18	0.69
20	0.64	0.58	0.23	0.81
35	0.83	0.80	0.24	0.92
50	0.92	0.84	0.43	0.99
T (4)				
10	0.22	0.23	0.11	0.14
15	0.29	0.26	0.09	0.07
20	0.35	0.27	0.11	0.11
35	0.49	0.38	0.19	0.17
50	0.60	0.42	0.22	0.20





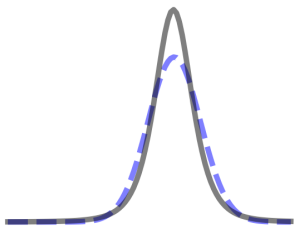


Table 9: The 10% significance power of the RMSE test to discriminate against symmetric long-tailed non normal populations, compared to analogous power of other tests as reported by Shapiro, Wilk & Chen (1968).

Group 4: Symmetric Short-Tailed Non Normal Populations

n	RMSE Test	SW	KS	CM
Beta (2, 2)				
10	0.07	0.13	0.12	0.11
15	0.07	0.12	0.07	0.09
20	0.08	0.12	0.17	0.17
35	0.12	0.22	0.14	0.15
50	0.18	0.37	0.14	0.12
Beta (1.1)				
10	0.12	0.15	0.14	0.14
15	0.16	0.28	0.11	0.08
20	0.23	0.39	0.19	0.19
35	0.51	0.76	0.18	0.18
50	0.77	0.96	0.26	0.27

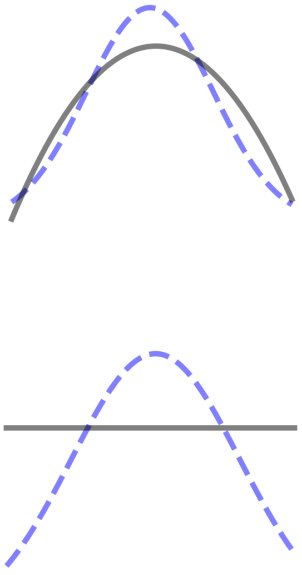


Table 10: The 10% significance power of the RMSE test to discriminate against symmetric short-tailed non normal populations, compared to analogous power of other tests as reported by Shapiro, Wilk & Chen (1968).

.

Group 5: Near Normal Non Normal Populations

n	RMSE Test	SW	KS	CM
T (10)				
10	0.13	0.13	0.09	0.08
15	0.15	0.19	0.06	0.11
20	0.18	0.18	0.11	0.13
35	0.22	0.17	0.06	0.09
50	0.26	0.21	0.09	0.09

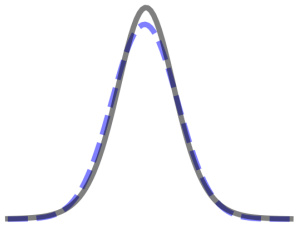


Table 11: The 10% significance power of the RMSE test to discriminate against symmetric a near normal population, compared to analogous power of other tests as reported by Shapiro, Wilk & Chen (1968).

Study 2: Filliben (1975)

Filliben (1975) proposed using the correlation of the points in the QQ-plot as the basis for a test for normality and compared his test (F) in terms of power to other tests, including the *D'Agostino* (D) and *Shapiro-Franka* (SF) tests, over a range of non normal populations including *chi-squared*, *log-normal*, *uniform* and *Weibull* distributions, for sample sizes of 20, and 50; in addition, he presented power estimates for his own test on these populations for sample size 100. As in the previous study, random numbers published by RAND (1955) were used by Filliben for the simulations.

A selection of the results presented by Filliben in the study, along with the power of the RMSE test in analogous contexts for a 5% significance level, are shown in table 12. Although the critical RMSE value for sample size 100 is not shown in table 6, is was obtained using the same methods described earlier and turns out to be 0.157, which was used to get the RMSE test power estimates displayed.

We notice that the power of the RMSE test is generally at least as powerful as D, SF and F. Like the RMSE statistic, F has the advantages of being conceptually closely related to the QQ-plot and simple in terms of computability and interpretation, and could be used to augment the QQ-plot in the way the RMSE statistic is intended to. However, the RMSE statistic appears to at least match F, and in some contexts outperform F, in terms of power while maintaining the advantages of F.

n	RMSE Test	D	SF	F
Chi-Squared (1)				
20	0.98	0.81	0.94	0.94
50	1.00	0.99	1.00	1.00
100	1.00	-	-	1.00
Log-Normal				
20	0.92	0.71	0.94	0.94
50	1.00	0.98	1.00	1.00
100	1.00	-	-	1.00
Uniform				
20	0.11	0.13	0.04	0.04
50	0.58	0.60	0.41	0.39
100	0.99	-	-	0.95
Weibull (3)				
20	0.04	0.03	0.02	0.02
50	0.05	0.06	0.04	0.03
100	0.07	-	-	0.01

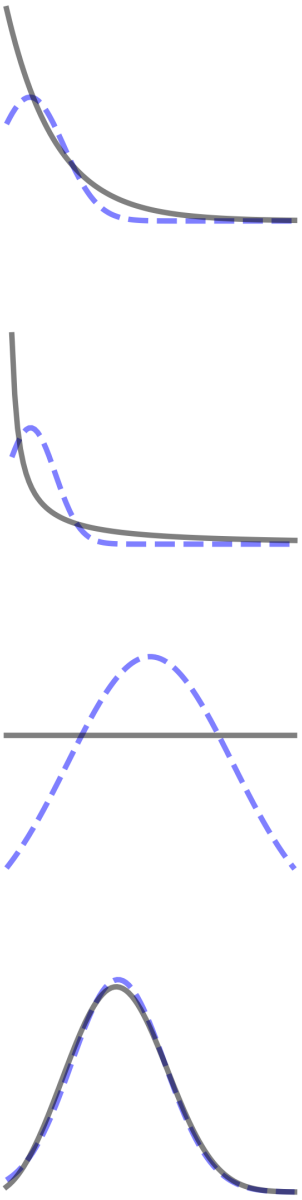


Table 12: The 5% significance power of the RMSE test to discriminate against a series of non normal populations, compared to analogous power of other tests as reported by Filliben (1975).

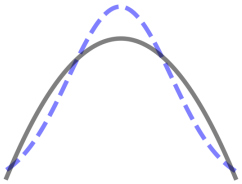
Study 3: Yazici & Yolacan (2007)

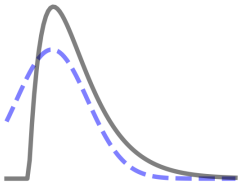
Yazici & Yolacan (2007) do not propose their own test, but similarly used simulations to compare several normality tests in terms of power over a range of non normal populations and sample sizes. They used a 5% significance level and sample sizes 20, 30, 40 and 50, tests considered included the *Jarque-Bera* (JB), *Chi-squared* (CS), *Anderson-Darling* (AD) and *modified Ajne* (MA) tests, and non normal population distributions included the *beta*, *gamma*, *log-normal* and *t* distributions.

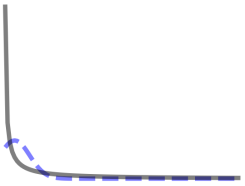
The power estimates of selected tests investigated by Yazici & Yolacan, along with that of the RMSE test, are shown in table 13.

We notice that the JB, CS, AD and MA tests have a much greater power than the RMSE test for populations that are similar to the normal distribution, such as beta (2, 2) and t (4), but, besides JB, are less powerful in dissimilar populations, such as the log-normal population. These tests, however, have not only been designed with normality in mind, however; for example, Jarque & Bera (1980) present their test as being a “three-directional test” that is sensitive to two other properties in addition to normality.

n	RMSE Test	JB	CS	AD	MA
Beta (2, 2)					
20	0.03	0.97	0.99	1.00	1.00
30	0.04	1.00	0.99	1.00	1.00
40	0.06	1.00	0.99	1.00	0.99
50	0.08	1.00	0.98	1.00	0.99
Gamma (2, 1)					
20	0.51	0.95	0.90	1.00	1.00
30	0.72	1.00	0.87	0.98	1.00
40	0.86	1.00	0.83	0.96	0.99
50	0.94	1.00	0.80	0.91	0.99
Log-Normal					
20	0.92	0.99	0.40	0.78	0.37
30	0.99	1.00	0.25	0.49	0.15
40	1.00	1.00	0.15	0.25	0.06
50	1.00	1.00	0.09	0.11	0.02
T (4)					
20	0.27	0.65	0.94	0.99	0.93
30	0.37	0.76	0.92	0.98	0.89
40	0.45	0.80	0.89	0.97	0.86
50	0.52	0.83	0.87	0.96	0.82







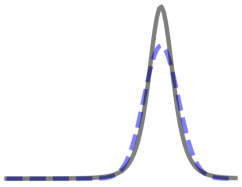


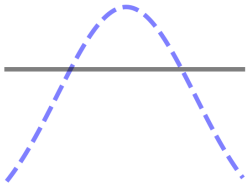
Table 13: The 5% significance power of the RMSE test to discriminate against a series of non normal populations, compared to analogous power of other tests as reported by Yazici & Yolacan (2007).

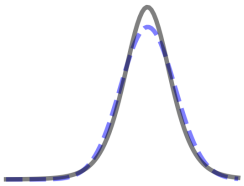
Study 4: Razali & Wah (2011)

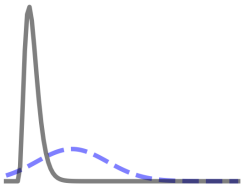
Razali & Wah (2011) also do not propose their own test, but similarly use simulation based techniques to compare the power of existing normality tests across a range of contexts. They use a 5% significance level on sample sizes 10, 20, 30 and 50; among the tests considered are the Shapiro-Wilk (SW) test, the Kalmogorov-Smirnov (KS) test, the Lillifors (L) test and the Anderson Darling (AD) test; non normal populations explored include the *uniform*, *t*, *gamma*, and *chi-squared* distributions. Table 15 shows a selection of their reported results with the analogous power of the RMSE test added in.

Again, we notice that the RMSE test performs relatively well: on par with SW and AD and outperforms KS and L in these contexts.

n	RMSE Test	SW	KS	L	AD
Uniform					
10	0.04	0.09	0.09	0.07	0.08
20	0.11	0.20	0.11	0.10	0.17
30	0.23	0.38	0.12	0.14	0.30
50	0.59	0.74	0.16	0.26	0.58
T (7)					
10	0.09	0.09	0.04	0.08	0.09
20	0.15	0.13	0.04	0.09	0.12
30	0.20	0.17	0.05	0.11	0.14
50	0.27	0.22	0.06	0.12	0.18
Gamma (4, 5)					
10	0.13	0.14	0.07	0.11	0.13
20	0.29	0.29	0.09	0.18	0.25
30	0.43	0.44	0.11	0.25	0.38
50	0.67	0.69	0.15	0.40	0.59
Chi-Squared (4)					
10	0.22	0.24	0.08	0.17	0.22
20	0.51	0.53	0.12	0.32	0.46
30	0.72	0.75	0.16	0.47	0.66
50	0.94	0.95	0.24	0.68	0.89







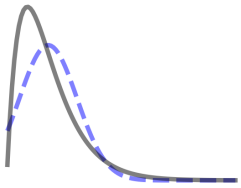


Table 14: The 5% significance power of the RMSE test to discriminate against a series of non normal populations, compared to analogous power of other tests as reported by Razali & Wah (2011).

CONCLUSIONS AND DISCUSSION

From the work presented in this paper, the RMSE test appears to have a performance comparable with many existing normality tests over a broad range of contexts. Several specialized tests appear to greatly outperform the RMSE test it in several narrow contexts, as we see in the selected results taken from Yazici & Yolacan (2007) in the previous section, however, such tests appear to be of greatest value for discriminating against populations that are very similarly distributed to a normal population. In many practical applications, however, that level of power might not be critical. For example, conclusions drawn from tests that assume normality might be practically valid even for populations that are not strictly normal but very nearly normal, in other words, when the assumption is only slightly violated. In that case, the RMSE test would still be of practical use even if it has a relatively low power to discriminate against populations that are near normal.

Given its natural association with the QQ-plot and simple interpretation, especially given that the QQ-plot is often used on its own as an informal visual test for normality, the RMSE test appears to have value as an augmentation of the QQ-plot in this use case. As already discussed, Filliben's normality test, which uses the correlation of the points in a QQ-plot as the test statistic, shares these advantages over several other normality tests, but appears to have a slightly lower statistical power in the contexts explored then the proposed RMSE test.

This paper only represents a cursory investigation of the viability of the RMSE statistic as a basis for a formal normality test, and further investigations that consider a more comprehensive comparison with existing normality tests might be of interest. Further, transformations of the RMSE statistic, such as the log RMSE, which has a normal sampling distribution, might be more convenient in practice or more sensitive to deviations from normality. Further, with minor updates as to how the theoretical z -scores are calculated in the THEORETICAL_Z_SCORES algorithm, such as INVERSE_NORMAL_CDF being replaced with the inverse cumulative probability distribution function of an alternate population, the RMSE test may be generalizable to test for deviation from populations that are not necessarily normal.

.

APPENDIX A: EXISTING TEST STATISTICS

Study 1: Shapiro, Wilk & Chen (1968)

Shapiro-Wilk

$$W = \frac{\left(\sum_{i=1}^{\left\lfloor \frac{n}{2} \right\rfloor} a_{n-i+1} \left(x_{n-i+1} - x_i \right) \right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2}$$

where a_{n-i+1} represents coefficients presented by Shapiro & Wilk, and indexing starting from 1.

Kolmogorov-Smirnov

$$\mathrm{KS} = \max \left(\left\{ \frac{i}{n} - \mathrm{F}(x_i) \;\middle|\; i \in \{1,\, 2,\, \ldots,\, n\} \right\} \right)$$

where F is the hypothesized normal cumulative probability function, and indexing starting from 1.

Weighted Cramer-von Mises

$$\mathrm{WCM} = n \int_0^1 \left(\mathrm{F}_n(y) - \mathrm{F}(y) \right)^2 \frac{\mathrm{dF}(y)}{\mathrm{F}(y)\left(1 - \mathrm{F}(y)\right)}$$

where F_n is the empirical distribution function.

Table A.1: Statistics used by a selection of tests explored by Shapiro, Wilk & Chen (1968).

Study 2: Filliben (1975)

Filliben (1975) assumed reader familiarity with tests discussed in his paper. The following information has been obtained from Arnastauskaitė, Ruzgas & Bražėnas (2021), who present summaries for a large number normality test statistics, including a more explicit definition of Filliben's own correlation statistic.

D'Agostino (D)
$$\mathrm{D} = \frac{\sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right) x_i}{n^2 \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2}}$$ with indexing starting from 1.

Shapiro-Franka (SF)
$$\mathrm{W_{SF}} = \frac{\left(\sum_{i=1}^{n} m_i x_i \right)^2}{\left(\sum_{i=1}^{n} x_i - \bar{x} \right)^2 \sum_{i=1}^{n} m_i^2}$$ where m_i is the expected value for x_i under the normal hypothesis, and with indexing starting from 1.

Filliben (F)
$$r = \frac{\sum_{i=1}^{n} x_i \Phi^{-1}(m_i)}{\sqrt{\sum_{i=1}^{n} \left(\Phi^{-1}(m_i) \right)^2} \sqrt{(n-1)\sigma^2}}$$ where m_i is the estimated proportion beneath x_i, and with indexing starting from 1.

Table A.2: Statistics used by a selection of tests explored by Filliben (1975), taken from Arnastauskaitė, Ruzgas & Bražėnas (2021).

Study 3: Yazici & Yolacan (2007)

Jarque-Bera (JB)
$$\mathrm{JB} = n\left(\frac{s^2}{6} + \frac{(k-3)^2}{24}\right)$$ where s is the sample *skewness* and k is the sample *kurtosis*.
Chi-Squared (CS)
$$\chi^2 = \sum_{i=1}^{n} \frac{\left(O_i - E_i\right)^2}{E_i}$$ where O_i is the observed value and E_i is the value expected under the normality assumption, and with indexing starting from 1.
Anderson-Darling
$$A^2 = -\sum_{i=1}^{n} \frac{2i-1}{n}\left(\ln z_i + \ln(1 - z_{n+1-i}\right) - n$$ with indexing starting from 1.
Modified Ajne (MA)
$$A^* = \left(A - \frac{0.7}{n} + \frac{0.9}{n^2}\right)\left(1 + \frac{1.23}{n}\right)$$ where A is the Ajne statistic; see Yazici & Yolacan (2007).

Table A.3: Statistics used by a selection of tests explored by Yazici & Yolacan (2007).

Study 4: Razali & Wah (2011)

In addition to the Shapiro-Wilk, Kolmogorov-Smirnov and Anderson-Darling tests, whose statistics have already been presented, Razali & Wah (2011) performed power comparisons for the Lillifors test, whose statistic is presented below.

Lillifors
$$\mathrm{D} = \max\left(\left\{ F(x) - S_n(x) \;\middle|\; x \in \mathrm{X} \right\}\right)$$ where F is the cumulative normal probability function with $\mu = \bar{x}$ and $\sigma^2 = s^2$, and S_n is the sample cumulative distribution function.

Table A.4: Statistics used by a selection of tests explored by Razali & Wah (2011).

APPENDIX B: A PYTHON IMPLEMENTATION

For convenience, the RMSE test presented in this paper has been implemented in Python in a library called *rmse_test*, which can be installed from its GitHub repository using:

```
pip install git+https://github.com/ram6ler/RMSE-Normality-Test
```

Example: Performing the RMSE Test on a Sample

The following example shows how the test may be performed for the example sample presented in this paper.

- Create a test for a sample drawn from a population with an unknown distribution.

```
from rmse_test import RMSETest

sample = [
    1.09, 1.17, 1.26,
    1.66, 1.72, 1.88,
    1.89, 1.90, 2.05,
    2.64, 4.21, 4.62,
    5.31, 7.59, 7.77,
]

test = RMSETest(sample)
```

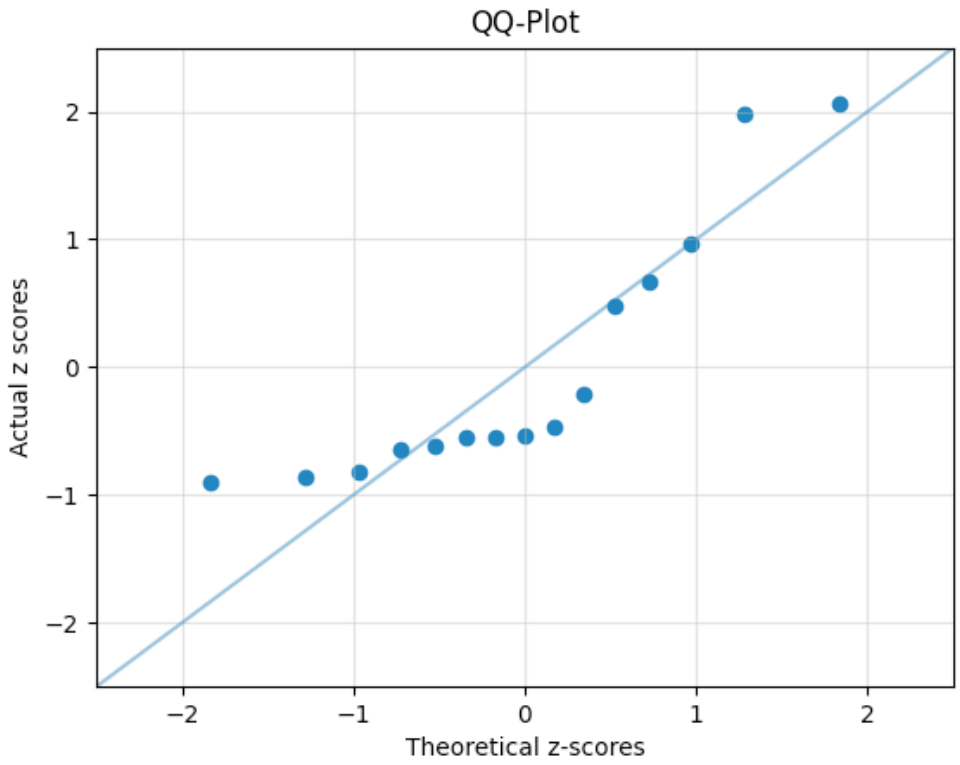
- Show a summary of the sample data.

```
print(test)
```

```
Key:
  x: sorted sample data
 tz: expected z-scores under normality hypothesis
  z: actual sample z-scores
  e: errors in predicting z with tz

Data:
.----------.----------.----------.----------.
|    x     |    tz    |    z     |    e     |
:----------:----------:----------:----------:
|   1.09   |  -1.834  |  -0.899  |  -0.935  |
|   1.17   |  -1.282  |  -0.863  |  -0.418  |
|   1.26   |  -0.967  |  -0.823  |  -0.144  |
|   1.66   |  -0.728  |  -0.646  |  -0.082  |
|   1.72   |  -0.524  |  -0.620  |   0.095  |
|   1.88   |  -0.341  |  -0.549  |   0.208  |
|   1.89   |  -0.168  |  -0.544  |   0.376  |
|   1.90   |   0.000  |  -0.540  |   0.540  |
|   2.05   |   0.168  |  -0.473  |   0.641  |
|   2.64   |   0.341  |  -0.212  |   0.552  |
|   4.21   |   0.524  |   0.484  |   0.040  |
|   4.62   |   0.728  |   0.666  |   0.062  |
|   5.31   |   0.967  |   0.972  |  -0.005  |
|   7.59   |   1.282  |   1.983  |  -0.701  |
|   7.77   |   1.834  |   2.063  |  -0.229  |
'----------'----------'----------'----------'

Results:
    n: 15
 RMSE: 0.449
    p: 0.006
```

- Plot the associated QQ-plot; for example:

```
import matplotlib.pyplot as plt

plt.xlim(-2.5, 2.5)
plt.ylim(-2.5, 2.5)
plt.plot([-2.5, 2.5], [-2.5, 2.5], alpha=0.5)
plt.grid(alpha=0.4)
plt.title("QQ-Plot")
plt.xlabel("Theoretical z-scores")
plt.ylabel("Actual z scores")
plt.scatter(test.theoretical_z_scores, test.z_scores)
plt.show()
```



- What is the RMSE statistic associated with this QQ-plot?

```
print(f"RMSE Statistic: {test.rmse}")
```

```
RMSE Statistic: 0.4494143469898124
```

- What is the 5% significance critical RMSE statistic value for a sample of this size?

```
print(f"Critical 5% RMSE: {test.critical_rmse(0.05)}")
```

```
Critical 5% RMSE: 0.34831873897411836
```

- What is the p-value associated with the RMSE statistic of this sample?

```
print(f"P-Value: {test.p_value}")
```

```
P-Value: 0.0057584819560067
```

Example: Estimating the Power of the RMSE Test in a Given Context

What is an estimate of the statistical power of the RMSE test for a sample of size 15 drawn from a log-normal population with parameters $\mu = 0$ and $\sigma = 1$, and using a significance level of 5%?

```
from rmse_test import Populations
from rmse_test.explore import rmse_power_estimate

Populations.seed(0)

print(
    rmse_power_estimate(
        sample_size=15,
        population=Populations.LOG_NORMAL(0.0, 1.0),
        experiments=100_000,
        significance_level=0.05,
    )
)
```

```
0.8049
```

For further information, please see *https://github.com/ram6ler/RMSE-Normality-Test*.

REFERENCES

Arnastauskaitė,J., Ruzgas, T. & Bražėnas, M. (2021). *An Exhaustive Power Comparison of Normality Tests*. Mathematics 9, 788. https:// doi.org/10.3390/math9070788

Filliben, J. (1975). *The Probability Plot Correlation Coefficient Test for Normality*. Technometrics, Feb., 1975, Vol. 17, No. 1, pp. 111-117

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. (2020). *Array programming with NumPy*. Nature 585, 357–362. DOI: 10.1038/s41586-020-2649-2.

Jarque, C. M. & Bera, A. K. (1980). *Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals*. Economics Letters 6, 255—259.

Marden, J. I. (2004). *Positions and QQ Plots*. Statistical Science, 19(4), 606–614. http://www.jstor.org/stable/4144431

Meeker, W. Q. & Escobar, L. A. (1998). *Statistical Methods for Reliability Data.* John Wiley & Sons. ISBN 0-47 1 - 14328-6.

RAND. (1955). A million random digits with 100,000 normal deviates. Glencoe, Ill. : Free Press. https://www.rand.org/pubs/monograph_reports/MR1418.html. Accessed September 2024.

Razali, N. & Wah, Y. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics, Vol.2 No.I, pp. 21-33

Shapiro, S. S., Wilk, M. B. & Chen, H. T. (1968). *A Comparative Study of Various Tests for Normality*. Journal of the American Statistical Association, Vol. 63, No. 324, pp. 1343-1372

Thode, H. (2002). *Testing For Normality*. State University of New York. Marcel Decker.

Wilk, M. & Gnanadesikan, R. (1968). *Probability Plotting Methods for the Analysis of Data*. Biometrika, Vol. 55, No. 1, pp. 1-17

Yazici, B. & Yolacan, S. (2007) *A comparison of various tests of normality*. Journal of Statistical Computation and Simulation, 77:2, 175-183, DOI: 10.1080/10629360600678310