Journal of Cataract & Refractive Surgery: January 2021 - Volume 47 - Issue 1 - p 65-77
doi: 10.1097/j.jcrs.0000000000000370

ARTICLE

Review and recommendations for univariate statistical analysis of spherical equivalent prediction error for IOL power calculations

Jack T. Holladay, MD, MSEE, Rand R. Wilcox, PhD, Douglas D. Koch, MD, Li Wang, MD, PhD

Purpose: To provide a reference for study design comparing intraocular lens (IOL) power calculation formulas, to show that the standard deviation (SD) of the prediction error (PE) is the single most accurate measure of outcomes, and to provide the most recent statistical methods to determine *P* values for type 1 errors.

Setting: Baylor College of Medicine, Houston, Texas, and University of Southern California, Los Angeles, California, USA.

Design: Retrospective consecutive case series.

Methods: Two datasets comprised of 5200 and 13 301 single eyes were used. The SDs of the PEs for 11 IOL power calculation formulas were calculated for each dataset. The probability density functions of signed and absolute PE were determined.

Results: None of the probability distributions for any formula in either dataset was normal (Gaussian). All the original signed PE distributions were not normal, but symmetric and leptokurtotic (heavy tailed) and had higher peaks than a normal distribution. The absolute distributions were asymmetric and skewed to the right. The heteroscedastic method was much better at controlling the probability of a type I error than older methods.

Conclusions: (1) The criteria for patient and data inclusion were outlined; (2) the appropriate sample size was recommended; (3) the requirement that the formulas be optimized to bring the mean error to zero was reinforced; (4) why the SD is the single best parameter to characterize the performance of an IOL power calculation formula was demonstrated; and (5) and using the heteroscedastic statistical method was the preferred method of analysis was shown.

A PubMed search of the past 5 years revealed 239 articles published on intraocular lens (IOL) power calculation formulas. The sample sizes ranged from 1 to 18 001 cases, and outcomes of mean prediction error (PE), mean absolute PE, median PE, associated SDs, and mean absolute deviations (MADs) were reported. There was no consistency in the reporting or the statistical methods used to compare formulas, techniques, or devices, although most converted to absolute values for the statistical analysis.

In their article in 2015 in *American Journal of Ophthalmology*, Hoffer et al. recommended optimizing the lens constant so that the arithmetic PE is zero, converting PEs to absolute values, and comparing median absolute errors because the distribution is not normal.[1] Aristodemou et al. in a Letter pointed out the deficiencies in the recommendations and proposed statistical comparisons between 2 formulas using the Wilcoxon signed-rank test and with 3 or more the Friedman test.[2–5] They also pointed out that, if the *P* value is not statistically significant, post hoc analysis can be performed to find out which group or groups are responsible for the null hypothesis being rejected, which might be used to correct for the multiple comparisons made as recommended by Benavoli et al.[6] We disagree with the comments by Aristodemou et al. and the response offered by Hoffer et al and will recommend current statistical techniques that overcome errors in the *P* value of the not normal, symmetrical, and heavy tailed PE distributions and that allow the use of the original signed value of the spherical equivalent (SEQ) PE (not the absolute value).[2,7]

In this study, we provided recommendations for designing a prospective study and characterized the distribution of PE and specific preoperative variables such as preoperative refraction, axial length, corneal power (keratometry), anterior chamber depth, and crystalline lens

Submitted: June 21, 2020 | Final revision submitted: July 15, 2020 | Accepted: July 22, 2020

From the Department of Ophthalmology, Baylor College of Medicine (Holladay, Koch, Wang), Houston, Texas, and Department of Psychology, University of Southern California, Los Angeles (Wilcox), Los Angeles, California, USA.

Corresponding author: Jack T. Holladay, MD, Department of Ophthalmology, Baylor College of Medicine, Cullen Eye Institute, 6565 Fannin St, Houston, TX 77030. Email: holladay@docholladay.com.

0886-3350/$ - see frontmatter
https://doi.org/10.1097/j.jcrs.0000000000000370

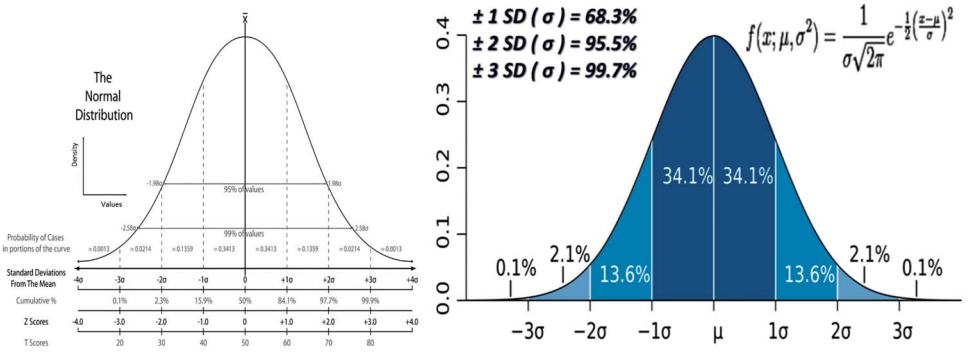


Figure 1. A normal probability density distribution (Gaussian) has 68.3% of the data within ±1 SD, 95.5% within ±2 SD, and 99.7% within ±3 SD.

thickness to show their relationship and effect on PE. Knowledge of the PE distributions is necessary to determine the appropriate statistics for comparison of the outcomes for formulas, procedures, and devices. We then proposed a new method for statistical analysis of univariate SEQ PE.

DESIGNING A PROSPECTIVE STUDY

Inclusion criteria should include no preoperative or postoperative pathology, 1 eye only from each patient, and corrected visual acuity of greater than or equal 20/30 to increase the likelihood that the postoperative refraction sphere and cylinder is accurate to ±0.25 diopters (D) with vertex distance specified. Multiple contributing surgeons broaden the applicability of the results. Studies performed by 1 surgeon have unique factors that can affect the results, including patient population, incision design, IOL insertion and placement, surgical instrumentation, and postoperative medications.[8,9] In particular, the method of capsulorhexis (eg, manual, femtosecond laser, or Zepto) along with its size and location have all been shown to not only affect the strength of the bag but also the contraction that occurs postoperatively.[10,11] This contraction can cause axial and lateral displacement of the IOL, which affects the lens effectivity and the final postoperative effective lens position (ELP) and refraction. Although changes in the SEQ power (not astigmatic) of the cornea are stable by 3 to 4 months or earlier,changes in the actual ELP, which directly affect the refraction, are usually not stable until 6 to 12 months postoperatively.[12–14] The U.S. Food and Drug Administration typically requires 12-month studies to assure the stability of the results. For results to be reliable and refractions stable, the 6-month visit is a good compromise. Shorter postoperative periods for reporting results have more variability and lower reliability.

Definitions

The proper statistical analysis of univariate (SEQ) and outcomes after cataract, refractive, and corneal surgery is challenging, even for biostatisticians. The following discourse provides the basis for evaluating PE using the appropriate statistics and explains the interpretation for the reader. The data that are used come from 2 large cataract surgery datasets.[15] The metric used for determining the accuracy of refractive outcomes is called PE.[16] It is the difference in the actual refraction and the predicted refraction using a specific formula:

Prediction Error (D) = Actual SEQ Refraction (D) – Predicted SEQ Refraction (D) (1)

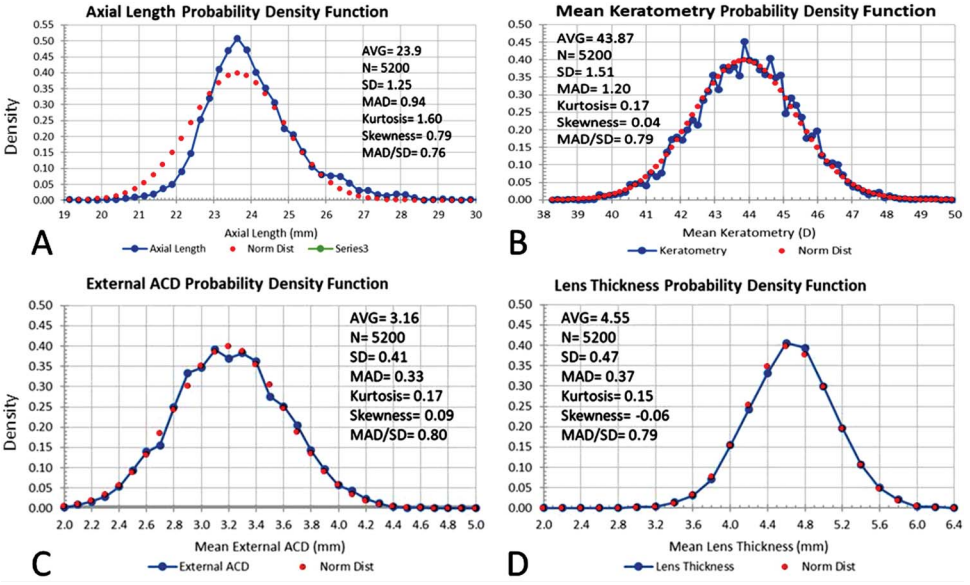


Figure 2. Using dataset 1, the distributions of axial length (*blue line*), SEQ keratometry, anatomic ACD, and CLT along with a normal distribution (*red line*) (ACD = anterior chamber depth; CLT = crystalline lens thickness; MAD = mean absolute deviation; SEQ = spherical equivalent).

Table 1. The 1% and 5% probability points of G_g.

Size of Sample n'	DOF n*	Upper Limit of G_g	Probability Points of G_g				Lower Limit of G_g	Mean of G_g	SD of G_g
			Upper 1%	Upper 5%	Lower 5%	Lower 1%			
6	5	1.000	0.980	0.954	0.696	0.626	0.4472	0.8385	0.0786
11	10	1.000	0.941	0.911	0.710	0.656	0.3162	0.8180	0.0613
16	15	1.000	0.916	0.891	0.720	0.677	0.2681	0.8113	0.0516
21	20	1.000	0.902	0.879	0.728	0.691	0.2236	0.8079	0.0454
26	26	1.000	0.892	0.870	0.734	0.701	0.2000	0.8059	0.0410
31	30	1.000	0.884	0.864	0.739	0.709	0.1826	0.8046	0.0376
36	36	1.000	0.878	0.859	0.743	0.715	0.1690	0.8036	0.0350
41	40	1.000	0.873	0.855	0.746	0.720	0.1581	0.8029	0.0328
46	45	1.000	0.869	0.851	0.749	0.725	0.1491	0.8023	0.0310
61	60	1.000	0.865	0.849	0.751	0.728	0.1414	0.8019	0.0295
76	76	1.000	0.863	0.839	0.759	0.741	0.1155	0.8005	0.0242
101	100	1.000	0.846	0.834	0.764	0.748	0.1000	0.7999	0.0210
501	500	1.000	0.820	0.814	0.783	0.776	0.0447	0.7983	0.0095
1001	1000	1.000	0.813	0.809	0.787	**0.782**	0.0316	0.7981	0.0067

*Degrees of freedom

This definition is opposite to what we have defined in some earlier articles, but agrees with our most recent publication.[15] We have chosen this definition because when the PE is negative, it is myopic, just similar to the refraction, and when PE is positive, it is hyperopic. This avoids the problem of the PE having the opposite the sign of the SEQ refraction.

This definition of PE is true whether the refractions are vectors representing an astigmatic refraction or scalar values such as the SEQ refraction, but we will limit our current discussion to scalar values. The SEQ refraction is defined as the sphere plus one half of the cylinder in the spherocylindrical form or one half of each cylinder in the cross-cylinder form as follows:

$$\text{SEQ} = \text{sphere} + \tfrac{1}{2}\,\text{cylinder} = \tfrac{1}{2}\,(\text{cylinder 1} + \text{cylinder 2}) \quad (2)$$

Statistical Terms

Mean, SD, Mean Deviation, Median, and Mean Absolute Error Calculating the mean PE is no different than calculating any other mean. The sample mean is the arithmetic sum of the prediction errors (PEi) divided by the number (n) of values in the dataset:

$$\text{Mean PE} = \bar{x} = \text{Sum PEi}/n \quad (3)$$

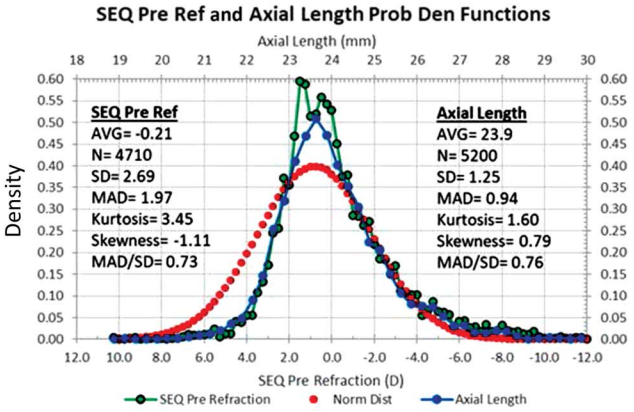
Note that, in Microsoft Excel, the function for the mean is AVERAGE.

The SD of the PE is the square root of the mean of the sum of the squares (root mean square [RMS]) about the mean of the $(\text{PE}_i - \bar{x})$ values.[17] For a normal distribution, 68.3% of data are within ±1.0 SD, and in Microsoft Excel, the function is STDEV.S:

$$\text{SD of the PE} = \sqrt{\frac{\text{Sum}(\text{PE}_i - \bar{x})^2}{n-1}} \quad (4)$$

Another statistical measure of variation is the mean absolute deviation (MAD). The MAD of the PE is calculated by taking the mean of the absolute values of the PEi about the mean. Note that, in Microsoft Excel 2010, the function for the MAD is AVEDEV:

$$\text{MAD} = \text{Sum} \mid \text{PEi} - \bar{x}\mid /n \quad (5)$$



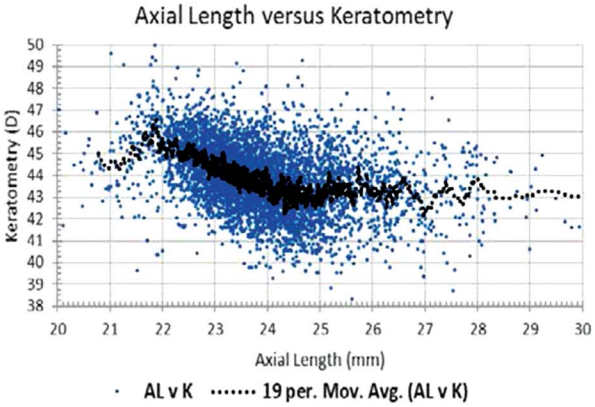


Figure 3. A: The correlation of SEQ preoperative refractions and axial lengths for 4710 patients of the 5200 cases, before affected by cataract. The axial length (*blue line*) is the primary factor determining the extreme peak of emmetropia (60%, *green line*). *B*: The peak for SEQ preoperative refraction is 8% higher than the axial length peak due to the inverse correlation (*black dots*) of keratometry and axial length from 22 to 24.5 mm (AL = axial length; K = keratometry; SEQ = spherical equivalent).

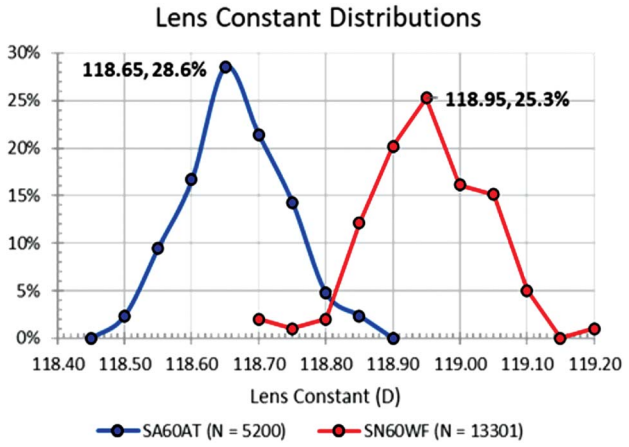


Figure 4. The distributions by surgeon of the individual lens constants for both datasets; 81% of dataset 1 and 86% of dataset 2 had individual surgeon lens constants that were within ±0.10 D of the mean.

For a normal distribution, the MAD is 0.7979 × SD and comprises 57.51% of the cases.

Another statistic often reported is the median of the absolute values (Median in Excel) where the number of values above and below is equal (MedAE). The MedAE is 0.6745 × SD, where exactly 50% of the absolute values are within this value and 50% outside of it.[17]

As we have defined earlier, the mean PE $\bar{x}$ is the arithmetic mean of the data. We will see that the datasets are rarely normal because our goal is to have a PE of zero, so formulas, procedures, and devices have (1) peaks that are much higher and narrower than a normal distribution and (2) tails that are usually heavier. The asymmetry or skewness of the PE distribution is usually minimal because the chances of myopic or hyperopic PEs are formulated to be equal. Even though the PE distributions are not normal, it will be helpful to review some of the characteristics of a normal distribution for comparison.

Normal Distribution When a probability density distribution is normal (Gaussian) as shown in Figure 1, 68.3% of the data are within ±1 SD, 95.5% within ±2 SD, and 99.7% within ±3 SD. Note that, in Microsoft Excel, the function for the SD of a population is STDEV.P.

Kurtosis, Skew (Asymmetry), and Geary Ratio

As stated earlier, the values given for the percentage of cases within ±1 SD or for 1 MAD are only true if the distribution is normal. We will see from our analysis of actual datasets that the PE distribution is not a normal distribution. There are 2 properties of any distribution that are usually tested: symmetry (skewness) and kurtosis (tailedness). If the data are not symmetrical about the mean and have a longer tail to the right, they are said to be positively skewed (Figure 2, A). By contrast, if the tail to the left is longer, the data are said to be negatively skewed. The *standardized* measure for skewness (*G*1) for a population in Excel (Skew) and in most modern software packages is the adjusted Fisher-Pearson standardized moment coefficient, given by the following formula:

$$G_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3 \qquad (6)$$

If skewness is positive, the data are positively skewed, and if negative, the data are negatively skewed. A rule of thumb is that:

1. If skewness is less than −1 or greater than +1, the distribution is highly skewed.
2. If skewness is between −1 and −½ or between +½ and +1, the distribution is moderately skewed.
3. If skewness is between −½ and +½, the distribution is approximately symmetric.

Another characteristic of the normal distribution for which we can test is *kurtosis*. In probability theory and statistics, kurtosis (from Greek: κυρτός, kyrtos or kurtos, meaning curved or arching) is a measure of the tailedness of the probability distribution of a real-valued random variable. The standard measure of a distribution's kurtosis, originating with Karl Pearson, is a scaled version of the fourth moment of the distribution.[18] This number is related to the tails of the distribution, not its peak; hence, the sometimes-seen characterization of kurtosis as peakedness is incorrect.[19] For this measure, higher kurtosis corresponds to greater extremity of deviations (or outliers) and not the configuration of data near the mean.

One measure used for this characteristic is the *excess* kurtosis function:

$$\gamma = \frac{m_4}{m_2^2} - 3, \qquad (7)$$

where, the fourth moment $m_4 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{n}$ and m_2^2 is the second moment (variance) squared. By subtracting 3, γ is zero for a normal distribution and increases above zero with increasing leptokurtosis. In Excel, the kurtosis function is KURT.

Although we will see that none of the formula PE datasets are normal, we should mention that Shapiro-Wilk test for univariate normality and the Anderson-Darling test for multivariate normality have the best power for a given significance using Monte Carlo simulations. An older test, proposed by Geary more than 100 years ago, compares the ratio (G_g) of the MAD with the SD.[20] The normal value for MAD/SD is $\sqrt{2/\pi} = 0.80$, and departure from this value for a given sample size determines the *P* value for not being normal. This ratio is easy to compute, and the values are readily available, so it is a quick way of confirming that the distribution is not normal. Table 1 lists the 1% and 5% probability points of the G_g distribution for samples from 6 to 1001. We will see that all 11 formulas have PEs for both datasets that are far below the **0.782**, indicating they are not normal at a *P* value much less than .01.

Distributions of Biometric Measurements–Axial Length, Keratometry, Anatomic Anterior Chamber Depth, and Lens Thickness Using dataset 1, the distributions of axial length, SEQ keratometry, anatomic anterior chamber depth, and crystalline lens thickness along with a

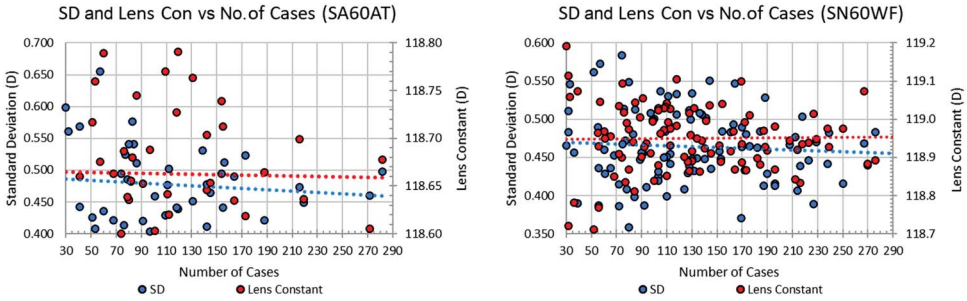


Figure 5. The PEs varied from 0.400 to 0.540 D for each surgeon who contributed more than 30 cases: SA60AT IOL (*A*) and SN60WF IOL (*B*). There was no significant correlation between the number of cases and the SD of the PE or the individual surgeon's lens constant, although the spread for both decreased as the number of cases contributed increased (PE = prediction error).

normal distribution (in *red dots*) are shown in Figure 2. For reference, a true normal density distribution will have a peak (at the mean) of 0.40. Notice that the axial length has a narrower, higher peak (0.52) than normal and is skewed to the right (toward longer axial lengths). The other 3 distributions are normal.

Figure 3A shows the correlation of SEQ preoperative refractions and axial lengths (for 4710 patients of the 5200 cases) before affected by cataract. The axial length (*blue line*) is the primary factor determining the extreme peak of emmetropia (60%, *green line*). The peak for SEQ preoperative refraction is 8% higher than the axial length peak due to the inverse correlation (*black line*) of keratometry and axial length from 22 to 24.5 mm (Figure 3, B). In this region, as the axial length increases, mean keratometry decreases, balancing their effects and, thereby, increasing

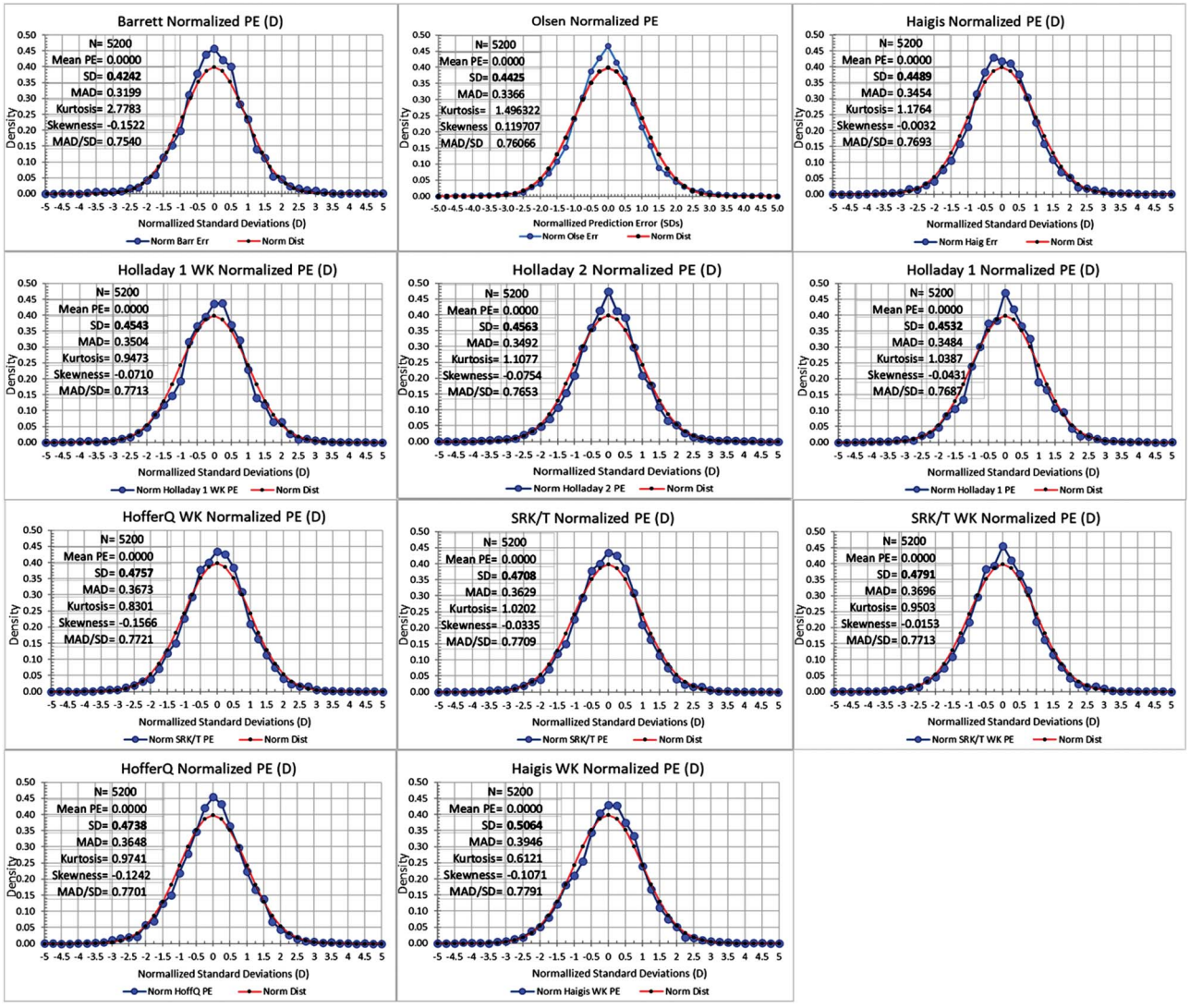


Figure 6. The actual PE distributions for 11 formulas from the 5200 cases in dataset 1 (*blue line*) and a normal distribution (*red line*). The mean, SD, MAD, kurtosis, skewness, and Geary ratio (MAD/SD) are shown on the inset for each graph (PE = prediction error).

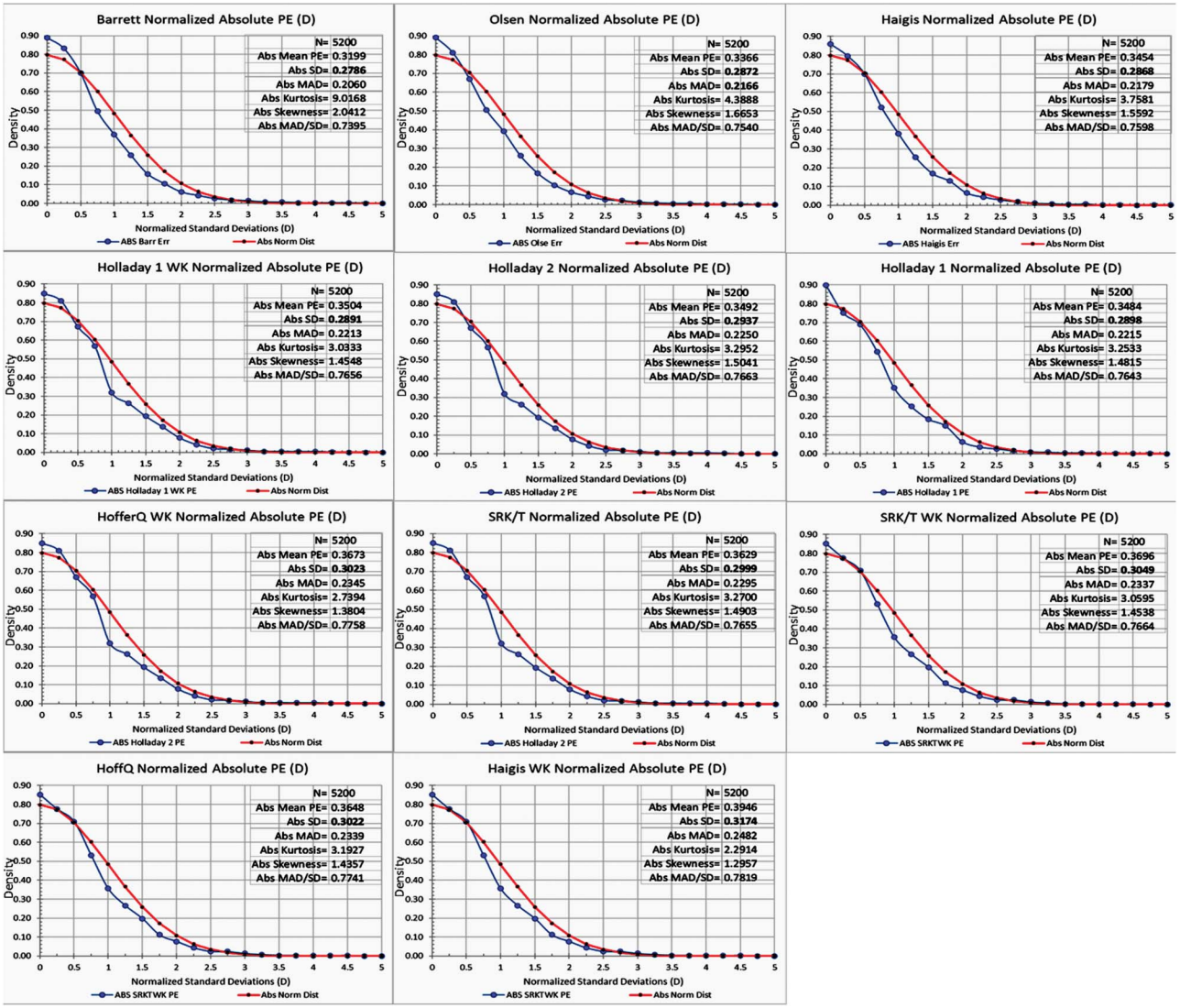


Figure 7. The absolute PE distributions for 11 formulas from the 5200 cases in dataset 1 (*blue line*) and a normal distribution (*red line*). The mean, SD, MAD, kurtosis skewness, and Geary ratio (MAD/SD) are shown on the inset for each graph (PE = prediction error).

the prevalence of emmetropia. Sorsby and Leary and later Rubin called this process emmetropization.[21,22]

The current emphasis will be restricted to the proper statistical methods for comparing means and SDs for PEs to determine the probabilities of whether values are different. For example, the statistics could be used to determine if an IOL power calculation formula performed better than another, a surgical technique was superior to another, or a laser performed better than another.

Characterizing PE Distributions Over the past 43 years, we have been fortunate to have been involved in reporting cataract outcomes and have hundreds of datasets with the raw data. The most recent publication with Melles et al. is representative of the biometry and SEQ PE distributions for 11 formulas.[15] Dataset 1 comprised of 5200 single eyes with a spherical IOL (SA60AT), and dataset 2 comprised of 13 301 single eyes with an aspheric IOL (SN60WF) for SEQ PE.

Lens Constant Optimization

The first step in the analysis is to adjust the arithmetic mean of the PE to zero for each formula by adjusting the lens constant (to at least to 6 decimal places 0.000000). This requires knowledge of which IOLs were implanted in each eye and access to the formulas to calculate the predicted refraction for each patient and to be able to optimize the lens constant. In Excel you must have the Add In Analysis ToolPak installed; under the Data Tab, there will be a Solver in the Analyze Section. You set the cell with the Target of the sum of the PEs to zero by changing the cell with the Lens Constant.

Ideally, although one should optimize the lens constant for each surgeon in the dataset, the benefit is minimal, and the effort is much greater. We had 95 surgeons and 127 surgeons in datasets 1 and 2, respectively, with implantation occurring between July 1, 2014, and December 31, 2015. In dataset 1 the difference in the lens constant by optimizing all 5200 cases as one surgeon vs each of the 95

Table 2A. Prediction error distribution parameters (*N* = 5200).

Formula	Mean	SD (D)	± 1.0 SD (%)	MAD/SD	Kurtosis	Skew
Barrett	0.000	0.424	72.87	0.755	2.778	−0.152
Olsen	0.000	0.443	71.94	0.761	1.496	0.120
Haigis	0.000	0.449	71.88	0.768	1.176	−0.003
Holladay 1	0.000	0.453	72.08	0.768	1.039	−0.043
Holladay 1 WK	0.000	0.454	72.46	0.771	0.947	−0.071
Holladay 2	0.000	0.456	72.46	0.765	1.108	−0.075
SRK/T	0.000	0.471	71.71	0.771	1.020	−0.033
Hoffer Q	0.000	0.474	70.94	0.770	0.974	−0.124
Hoffer Q WK	0.000	0.476	70.87	0.771	0.830	−0.157
SRK/T WK	0.000	0.479	71.75	0.772	0.974	−0.124
Haigis WK	0.000	0.506	70.04	0.781	0.612	−0.107

resulted in SDs of the PE of 0.4726 individually and 0.4791 globally, a difference of 0.0065 D, which is not clinically or statistically significant. Because the difference is negligible, it is acceptable to determine a single global lens constant that optimizes the PE to zero. The distributions by surgeon of the individual lens constants are shown in Figure 4 for both datasets; 81% of dataset 1 and 86% of dataset 2 had individual surgeon lens constants that were within ±0.10 D of the mean. The PEs varied from 0.400 to 0.540 D for each surgeon who contributed more than 30 cases (Figure 5). There was no significant correlation between the number of cases and the SD of the PE or the individual surgeon's lens constant, although the spread for both decreased as the number of cases contributed increased.

Figure 6 shows the actual PE distributions for 11 formulas from the 5200 cases in dataset 1 in *blue* and a normal distribution in *red*. The mean, SD, MAD, kurtosis, skewness, and Geary ratio (MAD/SD) are shown on the inset for each graph.

Figure 7 shows the absolute PE distributions for 11 formulas from the 5200 cases in dataset 1 in *blue* and a normal distribution in *red*. The mean, SD, MAD, kurtosis skewness, and Geary ratio (MAD/SD) are shown on the inset for each graph.

From Table 1, by taking the ratio of the MAD/SD (G_g), one might determine whether a dataset is not normal at the 1% probability. In both datasets, we have well more than 1001 cases, so for a normal distribution, there is a 0.01 probability of a value less than 0.782 (bottom of column 7). We see from Tables 2 and 3 and Figure 4 that all our formulas are far less than this value and, therefore, not normal. Furthermore, we see that converting to absolute values exaggerates the kurtosis, introduces significant skewness (asymmetry), and lowers the Geary ratio (Gg) even further.

In these datasets, none of the PE distributions with any formula are normal. The peak is always higher than a normal distribution, and the tails are always heavier (higher). The distributions (not absolute) are very symmetrical with skewness values within ±0.50 D. It is not surprising that the peaks are higher than normal. The goal of an IOL calculation formula is to have a PE of zero: a perfect formula would have a 100% peak at zero. The better the formula, the closer it is to that goal.

However, in the absolute distributions in Figure 7, the higher peak and tails are due to moving some of the intermediate PEs (1.0 to 3.0 SDs) to the peak and the remainder to the tails. In Figures 6 and 7, the area below the *red line* (normal distribution) must equal the area above the *red line* because the area under the probability curve must be 1.0 for both the *blue* and *red curves*.

In Table 4, we can see the exact amounts that have been moved for each interval of the SDs for dataset 2 (N = 13 301). The cases within ±1 SD for all formulas are higher than the

Table 2B. Absolute prediction error distribution parameters (*N* = 5200).

Formula	Mean	SD (D)	MAD (D)	Median Abs (D)	± 0.25 D (%)	± 0.50 D (%)	± 0.75 D (%)	± 1.00 D (%)	MAD/SD	Kurtosis	Skew
Barrett	0.320	0.279	0.206	0.252	49.8	80.0	92.7	97.2	0.739	9.017	2.041
Olsen	0.337	0.287	0.217	0.268	47.1	78.0	91.5	96.7	0.754	4.389	1.665
Haigis	0.345	0.287	0.218	0.278	45.3	76.3	90.9	96.8	0.760	3.758	1.559
Holladay 1	0.348	0.290	0.222	0.281	45.1	75.9	90.1	96.9	0.764	3.253	1.482
Holladay 1 WK	0.350	0.289	0.221	0.283	44.6	75.8	90.1	96.8	0.766	3.033	1.455
Holladay 2	0.349	0.294	0.225	0.277	46.1	75.3	90.4	96.6	0.766	3.295	1.504
SRK/T	0.363	0.300	0.230	0.290	43.7	74.1	89.5	96.0	0.765	3.270	1.490
Hoffer Q	0.365	0.302	0.234	0.292	43.7	73.0	89.4	96.3	0.774	3.193	1.436
Hoffer Q WK	0.367	0.302	0.234	0.295	43.4	72.9	89.0	96.0	0.776	2.739	1.380
SRK/T WK	0.370	0.305	0.234	0.298	42.4	73.5	88.8	95.9	0.766	3.059	1.454
Haigis WK	0.395	0.317	0.248	0.321	40.3	69.6	86.8	95.0	0.782	2.291	1.296

Table 3A. Prediction error distribution parameters (*N* = 13 301).

Formula	Mean	SD (D)	± 1.0 SD (%)	MAD/SD	Kurtosis	Skew
Barrett	0.000	0.404	71.6	0.770	1.192	−0.012
Olsen	0.000	0.424	71.7	0.767	1.274	0.146
Haigis	0.000	0.437	70.9	0.773	1.063	0.067
Holladay 1 WK	0.000	0.439	70.6	0.774	0.956	−0.055
Holladay 2	0.000	0.450	70.7	0.778	0.858	−0.050
Holladay 1	0.000	0.453	70.7	0.775	0.926	0.036
Hoffer Q WK	0.000	0.461	70.4	0.781	0.747	−0.129
SRK/T	0.000	0.463	70.4	0.778	0.834	−0.080
SRK/T WK	0.000	0.467	70.5	0.777	0.817	−0.056
Hoffer Q	0.000	0.473	70.3	0.780	0.687	−0.071
Haigis WK	0.000	0.490	70.0	0.782	0.688	−0.017

normal, ranging from a high of 3.46% (formula 2) to a low of 1.72% for formula 11 (first column). Those cases came from the 1 to 3 SDs that are all negative (lower than the normal). The cases with a PE above 3 SD are above the normal (positive), which explains why the distributions are considered heavy tailed. Notice in the last column that the sum in each row of the percentages of variations of the distributions from normal is zero for each formula.

STATISTICAL COMPARISON OF PE FOR IOL POWER CALCULATION FORMULAS

There are 2 main types of statistical comparison we will consider: (1) dependent paired samples from the same group and (2) independent samples that are from 2 different groups. The typical comparison of the power calculation formulas is the first type, which involves the use of 1 dataset for which the actual refraction is compared with the predicted refraction for each formula. Because only 1 dataset is used for all formulas, the comparisons are paired dependent samples for which the PEs for 2 formulas are used for each patient.

The second type of comparison of independent samples from 2 different groups might be comparisons of 2 different IOLs, surgical techniques, or devices such as femtosecond laser vs manual capsulorhexis. When independent, the samples should be randomized and might be slightly different in the total number of cases.

When analyzing data of the first type, where, for example, the PE of IOL calculation formulas in the same group is the random variable used to assess the performance, we will show that using the SD of PE is the single, most accurate assessment of performance and accurately predicts other measures such as the percentage of cases within an interval (eg, ±0.50), the MAD, and the median. We will also provide the appropriate, contemporary methods of determining the whether the performance differences in the SDs are statistically significant.

As shown in Equation 4, the SD is the square root of the mean of the sum of the squares (RMS) divided by the mean. The RMS value is used throughout science in every discipline to accurately compare differences in complex shapes, surfaces, or waveforms. In electrical engineering, the energy of an alternating sinusoidal current can be compared with the energy of a constant direct current using the RMS value. The energy of a sinusoid with amplitude of 1 has an RMS value of 70.7% of its peak and is equal in energy to the DC value of this amplitude. A perfect wavefront is a circular flat disk, but the ocular wavefront of the human eye is irregular and looks similar to a deformed potato chip. By computing the RMS value of the deviations from the perfect disc, we find a mean value of 0.38 ± 0.14 μm in the normal human.[23] Even though every human has a unique wavefront, the RMS value can be used to compare the visual quality between individuals with different wavefronts. In statistics, the RMS value about the mean is the SD. It allows

Table 3B. Absolute prediction error distribution parameters (*N* = 13 301).

Formula	Mean	SD (D)	MAD (D)	Median Abs (D)	± 0.25 D (%)	± 0.50 D (%)	± 0.75 D (%)	± 1.00 D (%)	MAD/SD	Kurtosis	Skew
Barrett	0.311	0.258	0.197	0.252	49.8	80.8	93.7	97.8	0.762	3.883	1.550
Olsen	0.325	0.272	0.208	0.258	48.8	78.7	92.5	97.4	0.762	3.975	1.578
Haigis	0.338	0.277	0.212	0.275	46.1	77.0	91.9	97.3	0.767	3.697	1.504
Holladay 1 WK	0.340	0.277	0.214	0.275	45.9	76.6	91.7	97.2	0.771	3.360	1.459
Holladay 2	0.350	0.283	0.218	0.287	44.5	75.4	90.9	97.0	0.770	3.117	1.425
Holladay 1	0.351	0.287	0.221	0.285	44.7	75.0	90.7	96.8	0.771	3.204	1.445
Hoffer Q WK	0.360	0.288	0.223	0.295	43.1	74.0	90.2	96.5	0.774	2.856	1.375
SRK/T	0.360	0.291	0.225	0.292	43.3	74.0	90.0	96.5	0.774	3.082	1.408
SRK/T WK	0.363	0.294	0.227	0.295	43.1	73.6	89.7	96.5	0.775	3.001	1.395
Hoffer Q	0.369	0.296	0.230	0.303	42.5	73.0	89.3	96.1	0.776	2.601	1.347
Haigis WK	0.383	0.305	0.237	0.318	40.6	71.0	88.3	95.6	0.777	2.744	1.342

Table 4. Difference from normal distribution (*N* = 13 301).

SD -->	0–1 (%)	1–1.5 (%)	1.5–2 (%)	2–3 (%)	3–4 (%)	>4 (%)	Sum (%)
Normal dist	68.27	18.37	8.81	4.28	0.26	0.01	100.00
Barrett	3.34	−2.17	−1.40	−0.32	0.40	0.14	0.00
Olsen	3.46	−2.15	−1.64	−0.33	0.50	0.15	0.00
Haigis	2.65	−1.79	−1.10	−0.33	0.43	0.14	0.00
Holladay 2	2.45	−1.41	−1.26	−0.24	0.36	0.10	0.00
Holladay 1	2.42	−1.60	−0.95	−0.39	0.39	0.13	0.00
Holladay 1 WK	2.31	−1.26	−1.31	−0.27	0.38	0.14	0.00
SRK/T WK	2.26	−1.54	−0.89	−0.22	0.29	0.11	0.00
Hoffer Q WK	2.18	−1.17	−1.19	−0.19	0.26	0.11	0.00
SRK/T	2.18	-1.37	-1.08	-0.17	0.33	0.11	0.00
Hoffer Q	2.06	-1.25	-1.15	-0.06	0.30	0.10	0.00
Haigis WK	1.72	-0.90	-1.11	-0.05	0.24	0.11	0.00

comparison of different shaped probability distributions with a single number.

For any probability density distribution, the total area under the curve is 1. The argument that the SD weighs the cases by the square of values is not true; on the contrary, it takes the square root of the mean of the sum of the squares so that the values are appropriately weighted and the area under the curve is 1. For a normal distribution, the ratio of the MAD/SD is 0.80, so the MAD is only 20% less than the SD.

Perhaps even more important is that the variance of the random variable (PE) can be expressed as a function of the variances and covariances of its constituent pieces (axial length, keratometry, predicted ELP, and pupil size).[24] Norrby identified 9 parameters that contributed more than 1% to the total PE, with the ELP accounting for 35%, the postoperative refraction 27%, the axial length 17%, corneal power 11%, and the pupil size 8% of the total PE. With other measures such as MAD and median, such an analysis is not possible because the total is not equal to the sum of the parts.

Because the PE measures are (1) dependent and (2) not normal, the standard F test for comparing SDs can not be used. It requires independence and normality. New perspectives in statistics have shown that, when not normal distributions are fairly symmetrical and heavy tailed, comparing the variances of 2 dependent variables can be improved by using the heteroscedastic (HC) method, a simple extension of the Morgan-Pitman test to the Spearman modification of the Morgan-Pitman test.[25] It is based partly on a HC method for making inferences about Pearson correlation.[26]

Performing these calculations is not practical for most clinicians, but fortunately, open access software is available from The R Project for Statistical Computing. Figure 8 shows, for the datasets 1 and 2 comprising 5200 and 13 301 eyes, the *P* values for each pair of variances for the 11 formulas for (1) the HC method, (2) the older Morgan-Pitman test based on Pearson correlation (MP), (3) the modification of the Morgan-Pitman suggested by McCulloch (1987) where Pearson correlation is replaced by Spearman correlation (SC), and (4) the Friedman test with the Nemenyi post hoc analysis used on the absolute values.[27–29] In Figure 8, we see that the Friedman test with the Nemenyi post hoc analysis using the absolute values for the PE (*yellow points*) results in much higher *P* values, especially when the paired formula SDs differences are lower. We also see that, in dataset 2 with 13 301 cases, the Friedman *P* value goes back up at the higher values (have an inflection). The cause of the erratic *P* values with the Friedman test is due primarily to the asymmetrical shape of the absolute values, the improper weighting of the values using the absolute values, and the post hoc analysis in the presence of heavy tails (Figure 7). Hoffer et al. had recommended the bootstrap method to deal with certain issues with datasets.[1] However, the erratic nature of the *P* values is exactly why Athreya states, "Unless one is reasonably sure that the underlying distribution is not heavy tailed, one should hesitate to use the naive bootstrap (post hoc analysis)."[30] For the methods using the SD (HC, MP, and SC), there is no reversal of the *P* values. The superiority of the HC method over the MP and SC methods becomes more apparent the smaller the sample size.[25] The original Morgan-Pitman test is based in part on testing the hypothesis of a zero Pearson correlation. The conventional method assumes homoscedasticity. As we move toward

Table 5. Matrix of paired standard deviation differences for dataset 1 (*N* = 5200).

Formula	SD	SD Dif	SD Dif	SD Dif	SD Dif	SD Dif	SD Dif	SD Dif	SD Dif	SD Dif	SD Dif	SD Dif
Barrett	0.424	0.000										
Olsen	0.443	0.018	0.000									
Haigis	0.449	0.025	0.006	0.000								
Holladay 1	0.453	0.029	0.011	0.004	0.000							
Holladay 1 WK	0.454	0.030	0.012	0.005	**0.001**	0.000						
Holladay 2	0.456	0.032	0.014	0.007	0.003	0.002	0.000					
SRK/T	0.471	0.047	0.028	0.022	0.018	0.017	0.015	0.000				
Hoffer Q	0.474	0.050	0.031	0.025	0.021	0.019	0.017	0.003	0.000			
Hoffer Q WK	0.476	0.051	0.033	0.027	0.023	0.021	0.019	0.005	0.002	0.000		
SRK/T WK	0.479	0.055	0.037	0.030	0.026	0.025	0.023	0.008	0.005	0.003	0.000	
Haigis WK	0.506	**0.082**	0.064	0.057	0.053	0.052	0.050	0.036	0.033	0.031	0.027	0.000
		Barrett	Olsen	Haigis	Holladay 1	Holladay 1 WK	Holladay 2	SRK/T	Hoffer Q	Hoffer Q WK	SRK/T WK	Haigis WK

Table 6. Matrix of paired adjusted HC *P* values for dataset 1 (*N* = 5200).

Formula	SD	HC adj *P*	HC adj *P*	HC adj *P*	HC adj *P*	HC adj *P*	HC adj *P*	HC adj *P*	HC adj *P*	HC adj *P*	HC adj *P*	HC adj *P*
Barrett	**0.424**	1										
Olsen	**0.443**	6.90E-04	1									
Haigis	**0.449**	9.50E-09	8.00E-01	1								
Holladay 1	**0.453**	8.00E-10	4.20E-01	8.00E-01	1							
Holladay 1 WK	**0.454**	1.30E-09	2.30E-01	8.00E-01	8.00E-01	1						
Holladay 2	**0.456**	3.50E-03	8.00E-01	8.00E-01	8.00E-01	8.00E-01	1					
SRK/T	**0.471**	2.70E-13	4.70E-06	3.70E-05	5.00E-09	9.40E-08	8.00E-01	1				
Hoffer Q	**0.474**	0.00E+00	1.50E-08	0.00E+00	1.50E-12	6.10E-06	4.20E-01	8.00E-01	1			
Hoffer Q WK	**0.476**	0.00E+00	6.20E-10	0.00E+00	1.50E-08	1.40E-12	2.30E-01	8.00E-01	8.00E-01	1		
SRK/T WK	**0.479**	0.00E+00	1.30E-09	8.00E-09	1.80E-10	0.00E+00	7.70E-02	2.30E-06	8.00E-01	8.00E-01	1	
Haigis WK	**0.506**	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	1.00E-08	3.70E-10	8.00E-10	0.00E+00	1.90E-06	1
		Barrett	**Olsen**	**Haigis**	**Holladay 1**	**Holladay 1 WK**	**Holladay 2**	**SRK/T**	**Hoffer Q**	**Hoffer Q WK**	**SRK/T WK**	**Haigis WK**

HC = heteroscedastic; Shaded Boxes = *P* value > 0.05 and are not considered statistically significant.

heavy-tailed distributions, heteroscedasticity becomes an issue. In effect, an incorrect estimate of the standard error is being used. The HC method addresses this problem. The SD is not only the best random variable to use statistically but also accurately predicts other key metrics. The first relationship that the SD accurately predicts is the percentage of cases within various intervals, such as within ±0.25, ±0.50, ±0.75, and ±1.00 D as shown in Figure 9.

The R^2 values are extremely high, ranging from 0.9173 to 0.9975. The SD also predicts the MAD and median even more accurately as shown in Figure 10, with the R^2 values are above 0.976. Table 5 summarizes the differences in each pair (55) of the SDs for the 11 formulas. The differences in SD for the 11 formulas range from a low of 0.001 D to a high of −0.082 D. The corresponding adjusted *P* values using the HC method are tabulated in Table 6. The *P* values have been adjusted because we conducted multiple comparisons on the same dependent variable with the 11 formulas. The chance of committing a type I error increases with multiple comparisons, thus increasing the likelihood of determining a significant result by pure chance. To correct for this and protect from type I errors, the Holm correction has been performed. The older Bonferroni correction can be criticized for being overly conservative, thus potentially excluding results of real significance.[31,32] The adjusted *P* values should be the values reported.

In Table 5, for dataset 1 (N = 5200), the smallest differences in SD correspond roughly with the 19 highest adjusted *P* values in Table 6 (pink-shaded boxes) that are above 0.05 and are not considered statistically significant. The adjusted *P* values along the diagonal are 1.0 because the probability of a formula being the same as itself is 1.0. For dataset 2 (N = 13 301), because the sample size was much larger, there were only 2 pairs of adjusted *P* values above 0.05.

DETERMINING MODIFIED MP *P* VALUE OF 2 SDS FOR 2 INDEPENDENT DATASETS

The second type of statistical comparison using 2 or more independent datasets would be used to compare the aspheric (13 301) and nonaspheric (5200) IOLs. A generalization of the HC version of the Morgan-Pitman test based on Pearson correlation (MP) is used.[26,27] The aspheric IOL with 13 301 cases that reflects current aspheric IOLs has a lower SD (0.404 D) than the nonaspheric IOL (0.424 D) with 5200

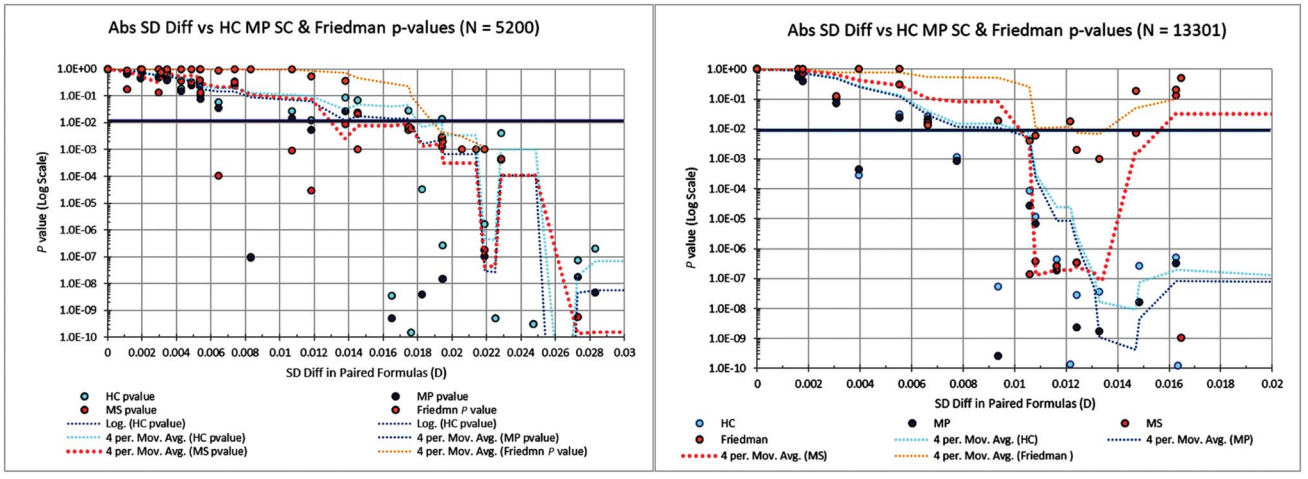


Figure 8. The Friedman test with the Nemenyi post hoc analysis using the absolute values for the PE (*yellow points*) results in much higher *P* values, especially when the paired formula SDs differences are lower. In dataset 2 with 13 301 cases, the Friedman *P* value goes back up at the higher values (has an inflection) (HC = heteroscedastic; MP = the older Morgan-Pitman test based on Pearson correlation; SC = Spearman correlation).

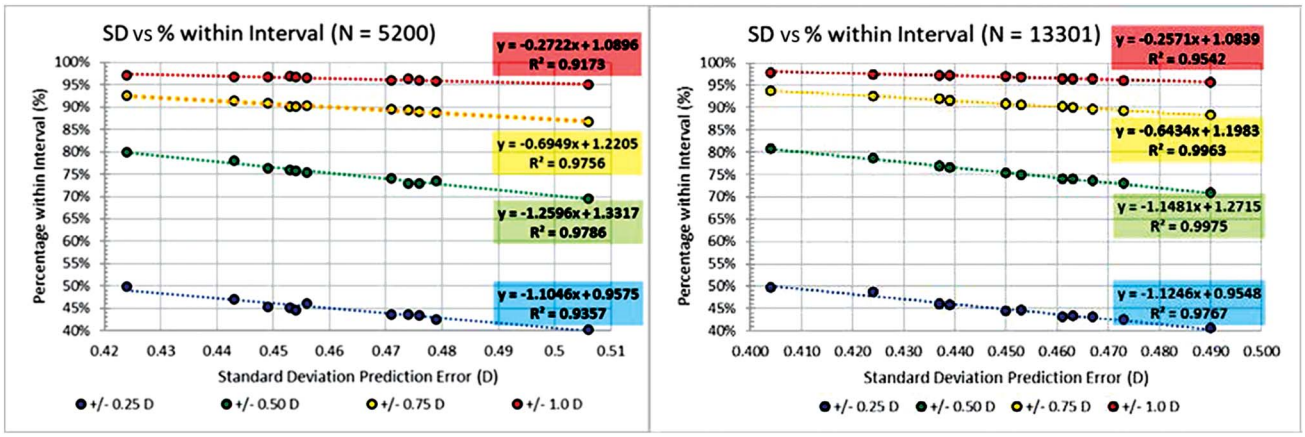


Figure 9. The SD accurately predicts the percentage of cases within various intervals, such as ±0.25, ±0.50, ±0.75, and ±1.00 D: dataset 1 (*A*) and dataset 2 (*B*). The R^2 values are extremely high, ranging from 0.9173 to 0.9975.

cases and a more accurate PE as predicted by Norrby.[24] The difference in SDs of 0.02 D results in 80.8% of PEs within ±0.50 D with the aspheric IOL vs 79.8% with the nonaspheric, a difference of 1.0% ($P = .005$). The smaller PE of the aspheric IOL is a result of the reduction in ocular spherical aberration. This reduction of spherical aberration reduces the influence of the pupil size on the effective IOL power for an individual patient and improves the quality of vision.[24,33] In fact, Norrby predicted in 2008 that the pupil size would be 8% of the SD, and we found it to be 5% (0.02/0.40).

If there were more than 2 independent random variables compared (3 or more IOLs), the *P* values must be adjusted because we conducted multiple comparisons on the same independent random variable (IOL type in this case) and the chance of committing a type I error increases. The Holm, Hommel, and Hochberg methods all represent a better balance excluding spurious positives without excluding true positives than the Bonferroni.[31,32,34,35]

Minimum Sample Size for Statistical Significance

A statistical significance with a *P* value of 0.01 is usually considered excellent and 0.05 is acceptable as a scientific minimal standard. The best way of determining a minimum sample size for statistical significance is empirically from previous studies. In Figure 8, dataset 1 with 5200 cases was able to show a difference of 0.02 D at a *P* value of 0.01 and dataset 2 with 13 301 cases was able to show a difference of 0.007 D at a *P* value of 0.01. Using the datasets 1 and 2, we can use sequentially ordered sampling to reduce the size of the sample and compute the HC *P* value. Figure 11 shows the number of cases and the resulting *P* values. The number of cases to achieve the same level of the difference in the SDs of 0.02 D for *P* value of approximately 0.01 for both datasets is between 300 and 700.

In this article, we have discussed the key elements for describing and analyzing the accuracy of IOL calculation formulas. We have (1) outlined the criteria for patient and data inclusion, (2) recommended the appropriate sample size, (3) reinforced the requirement that the formulas be optimized to bring the mean error to zero, (4) demonstrated why the SD is the single best parameter to characterize the performance of an IOL power calculation formula; it determines the percentage of cases within a given interval, the MAD, and the median of the absolute values; and (5) proposed that the HC statistical

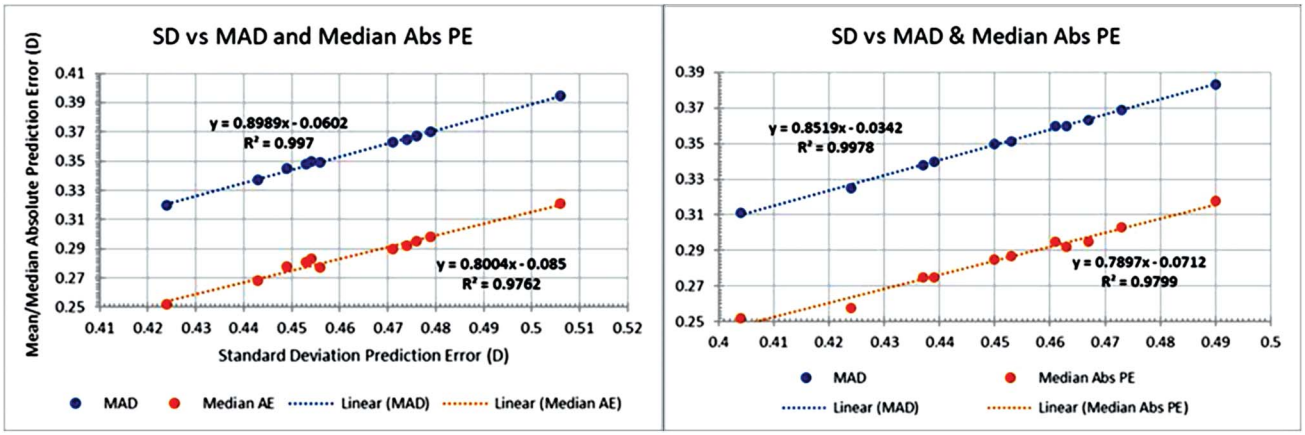


Figure 10. The SD also predicts the MAD and median even more accurately with the R^2 values are above 0.976 (AE = absolute error; PE = prediction error).

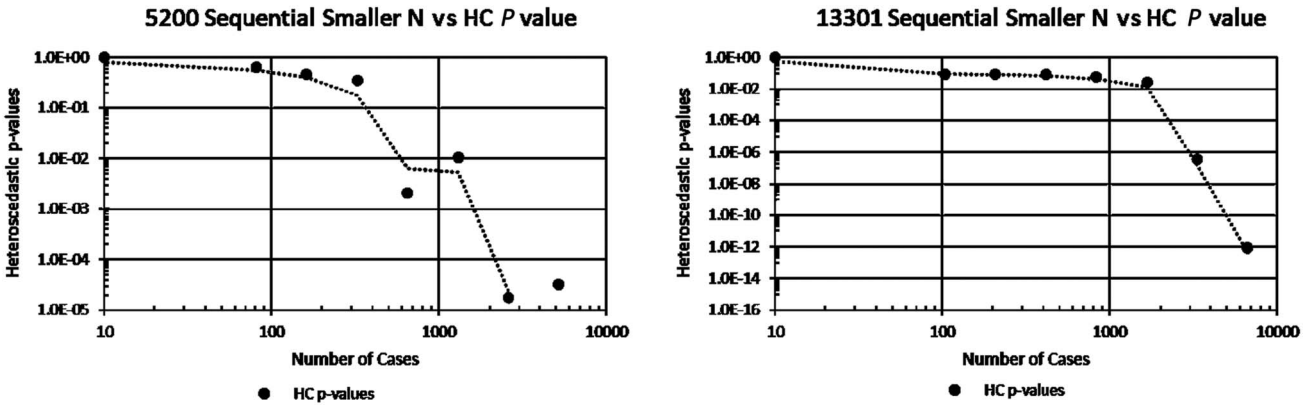


Figure 11. Graph showing that the number of cases to achieve a *P* value of .01 for an SD difference of 0.02 D in dataset 1 (*A*) is 300 and for an SD difference 0.007 D in dataset 2 (*B*) is 2000 (HC = heteroscedastic).

method is the preferred method of analysis, especially for smaller datasets; it is much better at controlling the probability of a type I error when the marginal distributions have heavy tails but are still symmetric. Details for downloading the open access software from The R Project for Statistical Computing can be found at https://www.r-project.org/ and details regarding how to implement HC analysis are in the README files at https://osf.io/nvd59/quickfiles.

WHAT WAS KNOWN

- Prediction error in diopters, the difference between the SEQ of the actual and predicted refraction, is used to measure the accuracy of intraocular lens power calculation formulas.
- The prediction error is usually converted to absolute values, creating a random variable that is not normal, asymmetric, and heavy tailed; the formulas are then compared using decades-old statistical methods that are erratic and unreliable for predicting the *P* values for a type 1 error.
- Other measures such as mean absolute deviation, median absolute error, mean absolute error, and percentages within various dioptric intervals are often reported to ameliorate this variability and limitation when using absolute values, but they also do not eliminate the problem.

WHAT THIS PAPER ADDS

- The original, signed prediction error is the random variable that is not normal, symmetric, and heavy tailed.
- The SD is the single most accurate measure of prediction error when comparing intraocular lens power calculation formulas; it predicts the results with other measures such as those mentioned earlier with extremely high R^2 values ranging from 0.9173 to 0.9975.
- The SD of the prediction error allows the use of modern, contemporary heteroscedastic statistical methods specifically intended for use with not normal, symmetric, heavy tailed random variables, providing accurate *P* values for type 1 errors and a valid method for comparing the formulas.

REFERENCES

1. Hoffer KJ, Aramberri J, Haigis W, Olsen T, Savini G, Shammas HJ, Bentow S. Protocols for studies of intraocular lens formula accuracy. Am J Ophthalmol 2015;160:403–405.e1
2. Aristodemou P, Cartwright NEK, Sparrow JM, Johnston RL. Statistical analysis for studies of intraocular lens formula accuracy. Am J Ophthalmol 2015;160:1085–1086
3. Wilcoxon F. Individual comparisons by ranking methods (PDF). Biometrics 1945;1:80–83
4. Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc 1937;32:675–701
5. Friedman M. A correction: the use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc 1939;34:109
6. Benavoli A, Corani G, Mangili F. Should we really use post-hoc tests based on mean-ranks? J Mach Learn Res 2016;17:1–10
7. Hoffer KJ, Aramberri J, Haigis W, Olsen T, Savini G, Shammas HJ, bentow S. Reply: to PMID 26117311. Am J Ophthalmol 2015;160:1086–1087
8. Holladay JT, Prager TC, Ruiz RS, Lewis JW, Rosenthal H. Improving the predictability of intraocular lens power calculations. Arch Ophthalmol 1986;104;539–541
9. Holladay JT, Prager TC, Chandler TY, Musgrove KH, Lewis JW, Ruiz RS. A three-part system for refining intraocular lens power calculations. J Cataract Refract Surg 1988;14:17–24
10. Findl O, Hirnschall N, Draschl P, Wiesinger J. Effect of manual capsulorhexis size and position on intraocular lens tilt, centration, and axial position. J Cataract Refract Surg 2017;43:902–908
11. Toto L, Mastropasqua R, Mattei PA, Agnifili L, Mastropasqua A, Falconio G, Nicola MD, Mastropasqua L. Postoperative IOL axial movements and refractive changes after femtosecond laser-assisted cataract surgery versus conventional phacoemulsification. J Refractive Surg 2015;31:524–530
12. Caglar C, Batur M, Eser E, Demir H, Yaşar T. The stabilization time of ocular measurements after cataract surgery. Semin Ophthalmol 2017;32:412–417
13. de Juan V, Herreras JM, Pérez I, Morejón Á, Río-Cristóbal A, Martín R, Fernández I, Rodríguez G. Refractive stabilization and corneal swelling after cataract surgery [published correction appears in Optom Vis Sci. 2013 Apr;90(4):e134. Cristóbal, Ana Río-San [corrected to Río- Cristóbal, Ana]]. Optom Vis Sci 2013;90:31–36
14. Gray A. Modern Differential Geometry of Curves and Surfaces. Boca Raton, FL: CRC Press; 1993:375–387;279–285
15. Melles RB, Holladay JT, Chang WJ. Accuracy of intraocular lens calculation formulas. Ophthalmology 2018;125:169–178
16. Holladay JT, Moran JR, Kezirian GM. Analysis of aggregate surgically induced refractive change, prediction error, and intraocular astigmatism. J Cataract Refract Surg 2001;27:61–79
17. Remington RD, Schork MA. Statistics with Applications to the Bilogical and Health Sciences. Englewood Cliffs, NJ: Prentice-Hall Inc; 1970;32–36
18. Zenga M, Fiori A. Karl Pearson and the origin of kurtosis. Int Stat Rev 2009;77:40–50
19. Westfall PH. Kurtosis as peakedness, 1905—2014. R.I.P. Am Stat 2014;68:191–195
20. Geary RC. The ratio of the mean deviation to the standard deviation as a test of normality. Biometrika 1935;27:310–332
21. Sorsby A, Leary GA. A longitudinal study of refraction and its components during growth. Spec Rep Ser Med Res Counc (G B) 1969;309:1–41
22. Rubin ML. Optics for Clinicians. 2nd ed. Gainesville, FL: Triad Scientific Publishers; 1974:129

23. McCormick GJ, Porter J, Cox IG, MacRae S. Higher-order aberrations in eyes with irregular corneas after laser refractive surgery. Ophthalmology 2005;112:1699–1709
24. Norrby S. Sources of error in intraocular lens power calculation. J Cataract Refract Surg 2008;34:368–376
25. Wilcox RR. Comparing the variances of two dependent variables. J Stat Distributions Appl 2015;2:7
26. Wilcox RR. Introduction to Robust Estimation and Hypothesis Testing. 5th ed. San Diego, CA: Academic Press; (in Press)
27. Morgan WA. A test for the significance of the difference between two variances in a sample from a normal bivariate population. Biometrika 1939; 31:13–19
28. Pitman EJG. A note on normal correlation. Biometrika 1939;31:9–12
29. McCulloch CE. Tests for equality of variance for paired data. Commun Stat Theor Methods 1987;16:1377–1391
30. Athreya KB. Bootstrap of the mean in the infinite variance case. Ann Stats 1987;15:724–731
31. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979;6:65–70
32. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilità. Florence, Italy: Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze. 1936;8:3–62
33. Holladay JT, Piers PA, Koranyi G, van der Mooren M, Norrby NE. A new intraocular lens design to reduce spherical aberration of pseudophakic eyes. J Refract Surg 2002;18:683–691
34. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 1988;75:383–386
35. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 1988;75:800–802

Disclosures: *None of the authors have a financial or proprietary interest in any material or method mentioned.*

First author:
Jack T. Holladay, MD, MSEE

Department of Ophthalmology, Baylor College of Medicine, Houston, Texas