Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2020;6(4):305-319

Methodology Article

A Note on Calibration of Clinical Prediction Models with Copas Statistics

James A. Koziol*

Proteomics Research Institute for Systems Medicine, La Jolla, California 92037, USA.

ARTICLE INFO ABSTRACT

Received 21.11.2020
Revised 29.11.2020
Accepted 08.12.2020
Published 10.12.2020

Key words:
Calibration;
Clinical prediction models;
Copas statistics

Background: Calibration of clinical prediction models often entails assessing goodness of fit with independent, non-identically distributed Bernoulli random variables. We here investigate two statistics studied by Copas in this setting.

Materials and Methods: We present distribution theory and a simulation study to compare the operating characteristics of the Copas statistics.

Results: In our simulation study with relatively small sample sizes, we found a simple Cornish-Fisher approximation tail quantiles of the distributions of the Copas statistics to perform adequately. Upon illustrating their use in a calibration study relating to prediction of atherosclerotic cardiovascular disease risk, power properties appear to reflect differential weighting accorded to observations, as evinced with other goodness-of-fit statistics.

Conclusion: The Copas statistics are easily implemented, have proven value in other contexts, and appear to be underutilized in calibration studies. They ought to be part of the armamentarium of calibration tools for all researchers

Introduction

Clinical prediction models have been increasingly utilized in disease management for individualized risk assessment and treatment choice[2]. In this regard, prior to model adoption for routine application in clinical practice, the accuracy of the model predictions needs to be established, leading in turn to issues of validation relating to discrimination and calibration.

Commonly used methods for assessment of performance of prediction models include the concordance or c statistic for discriminative ability[3], and the Hosmer-Lemeshow chi-squared test for goodness of fit or calibration[4,5,6] . Various authors[5,7,8] have pointed out elements of arbitrariness in these statistics, and development of novel or

* . Corresponding Author Email:JKoziol@prism-sd.org

refined alternatives to these statistics is an active research area.

The Hosmer-Lemeshow test assesses the level of agreement between observed outcomes and model predictions (expected outcomes). Although originally designed for assessing goodness of fit of binary response models with logistic regression, it is also widely used for calibration of clinical prediction models[2]. Hosmer and colleagues[5] studied power properties of the Hosmer-Lemeshow test, and found that two procedures suggested by Copas[1] had reasonable properties for assessing goodness of fit with binary response models. In this note, we consider these two procedures, and a variant, in the context of assessing goodness of fit with binary outcomes. We examine their limiting distributions in the next section, then briefly investigate some power properties, which point to potential limitations. We give an example related to prognosis of atherosclerotic cardiovascular disease in Section 4, and conclude with some remarks.

Methods: Theoretical Development

Let X_i, i=1,...,n, denote independent Bernoulli random variables with respective success probabilities π_i. We will consider the following statistics:

$$S_n = \sum_{i=1}^{n} (X_i - \pi_i)^2$$

$$T_n = \sum_{i=1}^{n} \frac{(X_i - \pi_i)^2}{\pi_i(1-\pi_i)}$$

$$D_n = \sum_{i=1}^{n} |X_i - \pi_i|.$$

S_n and T_n were investigated by Copas[1]. From linear algebra, S_n is the square of the Euclidean distance (also known as the L_2 norm) between the 1xn vectors $X=(X_1, X_2, \ldots, X_n)$ and $\pi=(\pi_1, \pi_2, \ldots, \pi_n)$, and D_n is the Manhattan distance (also known as the L_1 norm) between the two vectors.

S_n, T_n, and D_n are each sums of independent, non-identically distributed random variables. Unfortunately, closed form expressions for their distributions are in general intractable. Instead, we will initially rely upon the moments of these statistics to derive approximate distributions. It is easily shown that the moment generating functions (mgfs) for these statistics are

$$M_S(t) = \prod_{i=1}^{n} \left[\pi_i \exp(t(1-\pi_i)^2) + (1-\pi_i)\exp(t\pi_i^2) \right]$$

$$M_T(t) = \prod_{i=1}^{n} \left[\pi_i \exp(t\frac{1-\pi_i}{\pi_i}) + (1-\pi_i)\exp(t\frac{\pi_i}{1-\pi_i}) \right]$$

and

$$M_D(t) = \prod_{i=1}^{n} \left[\pi_i \exp(t(1-\pi_i)) + (1-\pi_i)\exp(t\pi_i) \right]$$

respectively. Means and variances of the statistics are straightforward; and, higher order moments of the statistics can be easily obtained from the mgfs or the corresponding cumulant generating functions. In particular,

$$E(S_n) = \sum_{i=1}^{n}\left(\pi_i(1-\pi_i)^2 + (1-\pi_i)\pi_i^2\right) = \sum_{i=1}^{n}\pi_i(1-\pi_i)$$

$$\mathrm{Var}(S_n) = \sum_{i=1}^{n}\left(\pi_i(1-\pi_i)^4 + (1-\pi_i)\pi_i^4 - (\pi_i(1-\pi_i)^2 + (1-\pi_i)\pi_i^2)^2\right)$$

$$= \sum_{i=1}^{n}\pi_i(1-\pi_i)(1-2\pi_i)^2.$$

We also have

$$E(T_n) = n,$$

$$Var(T_n) = \sum_{i=1}^{n}\left(\frac{\pi_i^2}{1-\pi_i} + \frac{(1-\pi_i)^2}{\pi_i}\right) - n$$

$$= \sum_{i=1}^{n}\frac{(1-2\pi_i)^2}{\pi_i(1-\pi_i)},$$

and,

$$E(D_n) = \sum_{i=1}^{n}2\pi_i(1-\pi_i),$$

$$Var(D_n) = \sum_{i=1}^{n}\left(\pi_i(1-\pi_i)^2 + (1-\pi_i)\pi_i^2 - 4\pi_i^2(1-\pi_i)^2\right)$$

$$= \sum_{i=1}^{n}\pi_i(1-\pi_i)(1-2\pi_i)^2.$$

By Lyapunov's central limit theorem, if the π_i are bounded away from 0 and 1, and not all equal to ½, the standardized statistics $Z_S=(S_n-ES_n)/SD(S_n)$, $Z_T=(T_n-ET_n)/SD(T_n)$, and $Z_D=(D_n-ED_n)/SD(D_n)$ will each converge in distribution to the standard normal distribution as n increases (SD denoting standard deviation).

We remark that if all the π_i are identically ½, then the distributions are simple point masses: $S_n=n/4$, $T_n=n$, and $D_n=n/2$, with probabilities one. Note in addition that for any $X_i$with $\pi_i = ½$, there is no contribution from that term to any of Z_S, Z_T, or Z_D.

We also note the following relationship between S_n and D_n: with Bernoulli (0-1) random variables X_i, $(X_i-\pi_i)^2 - \pi_i(1-\pi_i) = |X_i-\pi_i| - 2\pi_i(1-\pi_i)$, that is, $S_n-E(S_n) = D_n-E(D_n)$. It follows that the standardized variables Z_S and Z_D are numerically identical, though higher order moments will differ.

For large n, the normal approximation to the exact distributions of these statistics should be sufficient, but might be improved on. Copas[1] details a χ^2 approximation [that is, $a\chi_b^2$] for S_n and T_n, where a and b are obtained by matching the first two moments.

Further improvements might be possible with higher order moment corrections, e.g., Edgeworth or saddlepoint approximations.
In this regard, we illustrate two straightforward approximations in a limited simulation study, as follows. For each n from 10 to 80 in steps of 10, we first constructed n x 1 probability vectors p by setting p_i = i/(n+1), i=1,2,...n. We then generated 10000 independent n x 1 X vectors of 0's and 1's by randomly taking X_i ~ Bernoulli(p_i), i=1,2,...n. We next calculated 10000 S_n and T_n statistics to determine their empirical distributions, based on the random X and fixed p vectors.

Our first approximation to the distributions of S_n and T_n is the normal approximation, based on the exact means and variances of the statistics as given above. Our second approximation is based on the Cornish-Fisher expansion[9]. Briefly, given a "near normal" cumulative distribution function F, the Cornish- Fisher approximation for the value y_p at quantile p of the F distribution is y_p ~ m + s*w, where m and s are the mean and standard deviation of the F distribution, and w is given by

$$w = x + \gamma_1 h_1(x) + \gamma_2 h_2(x) + \gamma_1^2 h_{11}(x) + ...$$

Here, $x=\Phi^{-1}(p)$, the pth quantile of the standard normal distribution,

$$\gamma_{r-2} = \frac{\kappa_r}{\kappa_2^{r/2}}, r = 3,4,...$$

$$h_1(x) = \frac{x^2 - 1}{6},$$

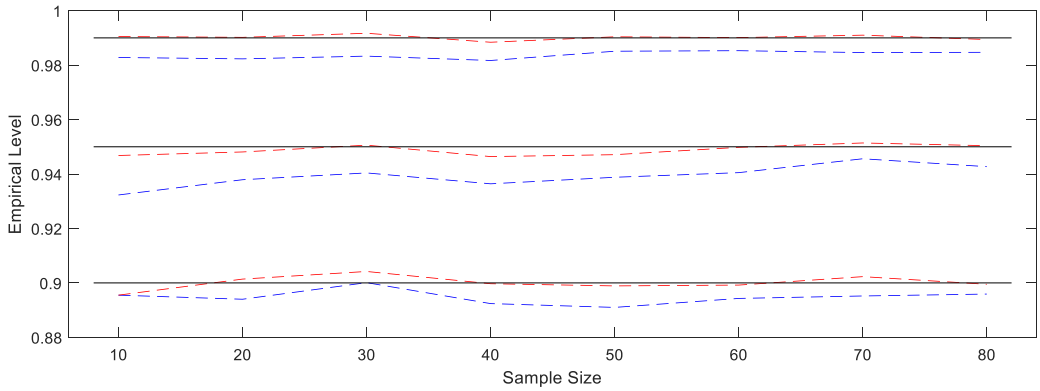
$$h_2(x) = \frac{x^3 - 3x}{24},$$

$$h_{11}(x) = -\frac{2x^3 - 5x}{36},$$

and the k_r are the cumulants of the distribution F.

For each n, we calculated the estimated upper 90th, 95th, and 99th percentiles of the distributions of S_n and T_n from both the normal approximation and the Cornish-Fisher expansion, and determined the observed levels of these percentiles by comparison with the empirical distributions of S_n and T_n we had previously generated. We plot these observed levels in Figure 1. The normal approximation tends to underestimate the percentiles for both S_n and T_n, an unsurprising finding given the marked skewness of the distributions of S_n and T_n for these small values of n. On the other hand, even the simple three term Cornish-Fisher expansion we have used seems to estimate the percentiles of S_n rather accurately, but shows some variability with T_n at the 95th and especially the 90th percentile.

A



B

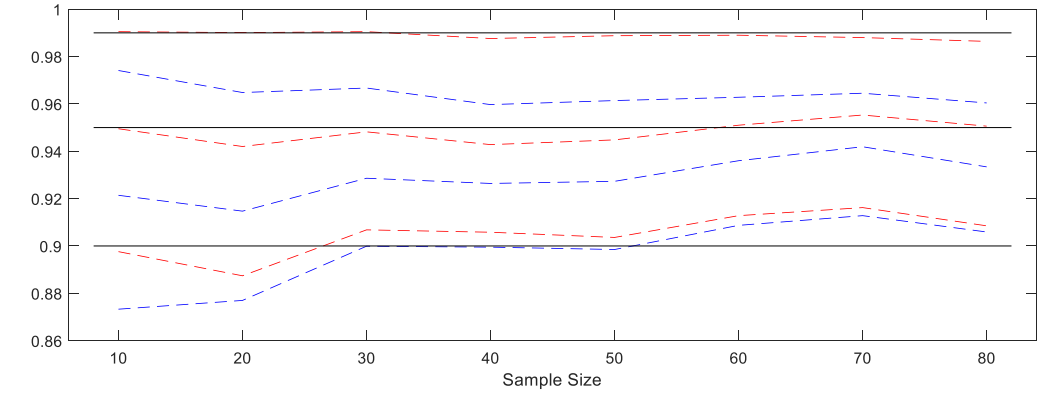


Figure 1. Achieved levels of the statistics S_n (A) and T_n (B) in a simulation study involving 10000 replications of S_n and T_n at sample sizes n=10 to 80 in steps of 10. The levels were calculated using estimated critical values at alpha levels .90, .95, and .99, demarcated by solid horizontal black lines. The blue dashed lines depict the observed levels from critical values determined from normal approximations to the distributions of S_n and T_n, and the red dashed lines depict the observed levels from critical values determined from three term Cornish-Fisher approximations based on the cumulants of S_n and T_n. The normal approximations tend to underestimate the 90^{th}, 95^{th}, and 99^{th} percentiles of S_n and T_n, whereas the Cornish-Fisher approximations generally estimate these percentiles fairly accurately.

Operating characteristics

In the previous section, we have discussed approximate distributions of S_n, T_n, and D_n when the X_i are independent Bernoulli

random variables with success probabilities π_i. In a general goodness of fit scenario, the π_i would be known and prespecified. Let us consider an alternative hypothesis, that each of the π_i is shifted by a fixed, small positive amount δ. How are the distributions of the statistics affected by this shift?

We find that

$$E(S_n \mid \delta) = \sum_{i=1}^{n} \pi_i(1-\pi_i) + \delta \sum_{i=1}^{n} (1-2\pi_i),$$

$$Var(S_n \mid \delta) = \sum_{i=1}^{n} \pi_i(1-\pi_i)(1-2\pi_i)^2 + \sum_{i=1}^{n} (1-2\pi_i)^2(\delta(1-2\pi_i) - \delta^2),$$

$$E(T_n \mid \delta) = n + \delta \sum_{i=1}^{n} \frac{1-2\pi_i}{\pi_i(1-\pi_i)},$$

$$Var(T_n \mid \delta) = \sum_{i=1}^{n} \frac{(1-2\pi_i)^2}{\pi_i(1-\pi_i)} + \sum_{i=1}^{n} \frac{(1-2\pi_i)^2(\delta(1-2\pi_i) - \delta^2)}{\pi_i^2(1-\pi_i)^2},$$

and

$$E(D_n \mid \delta) = 2\sum_{i=1}^{n} \pi_i(1-\pi_i) + \delta \sum_{i=1}^{n} (1-2\pi_i),$$

$$Var(\mathrm{D}_n \mid \delta) = \sum_{i=1}^{n} \pi_i(1-\pi_i)(1-2\pi_i)^2 + \sum_{i=1}^{n} (1-2\pi_i)^2(\delta(1-2\pi_i) - \delta^2)$$

under this shift. Note that, even if δ is positive, the increments in means and hence $E(Z_S)$, $E(Z_T)$, and $E(Z_D)$ can be negative (e.g., if all $\pi_i > ½$) or 0 (again, trivially, if all $\pi_i = ½$, or more generally if the π_i and π' are paired so that $\pi_i = 1-\pi'$). The upshot is, it may be that none of the Z statistics derived from S_n, T_n, or D_n will be sensitive to shifts in the magnitudes of the π_i, depending on the originally hypothesized values of the πi. These statistics are not consistent against all global alternatives.

We illustrate these points by simulation. First, we generated 1000 π_i from a uniform U(0, .45) or a U(.50, .95) distribution, then generated the X_i as Bernoulli(π_i) random variables, to calculate the empirical null distributions of Z_S and Z_T. We then shifted the π_i to the right by $\delta = .05$, and recalculated the sampling distributions of Z_S and Z_T under this shift. [The X_i are now Bernoulli($\pi_i+\delta$), but the null means and standard deviations are incorporated into Z_S and Z_T.] The resulting empirical histograms of Z_S and Z_T are given in Figures 2 and 3 respectively.

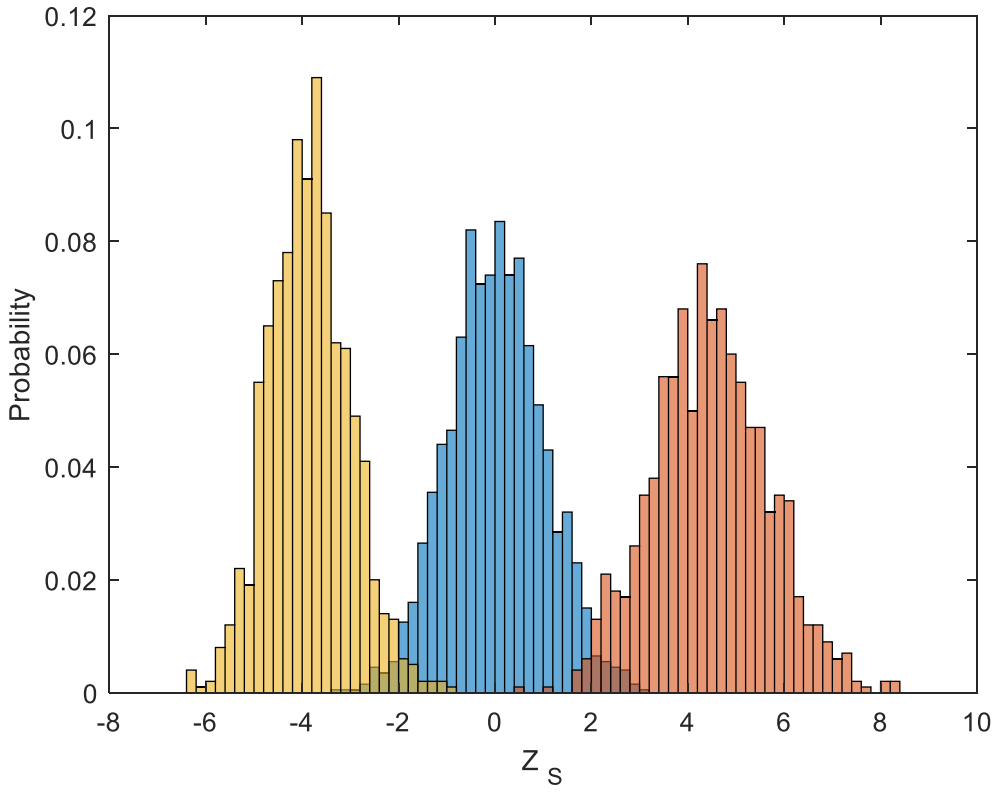


Figure 2. Empirical histograms of Z_S, normalized to probabilities, under three scenarios. The middle histogram is derived from 1000 replicates of Z_S in which, for each Z_s, 1000 π_i were generated from a uniform U(0,.45) distribution, the X_i were Bernoulli(π_i) random variables, ES_n and $SD(S_n)$ were computed from these π_i, plus 1000 replicates of Z_S in which, for each Z_S, 1000 π_i were generated from a uniform U(.5, .95) distribution, the X_i were again Bernoulli(π_i) random variables, and, ES_n and $SD(S_n)$ were computed from these π_i,. The middle histogram should approximate a standard normal [N(0,1)] distribution. Computing formulas are detailed in Section 2. The right histogram is derived from 1000 replicates of Z_S in which, for each Z_s, 1000 π_i were generated from a uniform U(.05, .50) distribution, the X_i were Bernoulli(π_i) random variables, but ES_n and $SD(S_n)$ were taken from the U(0, .45) simulations. The left histogram is derived from 1000 replicates of Z_S in which, for each Z_s, 1000 π_i were generated from a uniform U(.55, 1) distribution, the X_i were Bernoulli(π_i) random variables, but ES_n and $SD(S_n)$ were taken from the U(.5, .95) simulations.

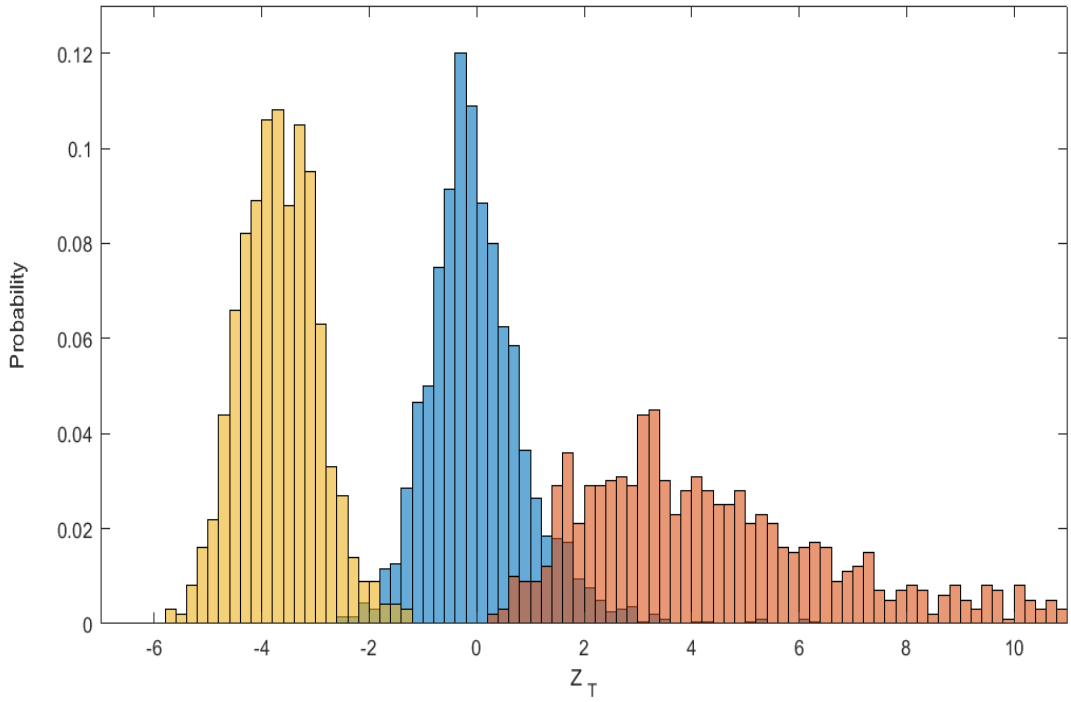


Figure 3. Empirical histograms of Z_T, normalized to probabilities, under three scenarios. The middle histogram is derived from 1000 replicates of Z_T in which, for each Z_T, 1000 π_i were generated from a uniform U(0,.45) distribution, the X_i were Bernoulli(π_i) random variables, ET_n=1000, SD(T_n) was computed from these π_i, plus 1000 replicates of Z_T in which, for each Z_T, 1000 π_i were generated from a uniform U(.5, .95) distribution, the X_i were again Bernoulli(π_i) random variables, and, ET_n=1000, and SD(T_n) was computed from these π_i,. The middle histogram should approximate a N(0,1) distribution. Computing formulas are detailed in Section 2. The right histogram is derived from 1000 replicates of Z_T in which, for each Z_T, 1000 π_i were generated from a uniform U(.05, .50) distribution, the X_i were Bernoulli(π_i) random variables, but ET_n and SD(T_n) were taken from the U(0, .45) simulations. The left histogram is derived from 1000 replicates of Z_T in which, for each Z_T, 1000 π_i were generated from a uniform U(.55, 1) distribution, the X_i were Bernoulli(π_i) random variables, but ET_n and SD(T_n) were taken from the U(.5, .95) simulations.

Both Z_S and Z_T are sensitive to the shifts, but in the case of U(.50, .95) shifting to U(.55, 1), Z_S and Z_T turn negative. The implication is that two-sided alternatives to the null distributions of Z_S and Z_T ought to be examined. The variability in the empirical distribution of Z_T under the U(0, .45) to U(.05, .5) shift is also pronounced.

We also looked at beta alternatives. We generated 1000 π_i from a uniform U(0, 1) distribution, then generated the X_i as Bernoulli(π_i) random variables, for the empirical null distributions of Z_S and Z_T. We examined two alternatives: (a) the π_i are from

a Beta(2,2) distribution; (b) the π_i are from a Beta(.5, .5) distribution. The Beta(2,2) distribution is symmetric and unimodal, with mode .5, whereas the Beta(.5, .5) distribution is symmetric U-shaped over (0, 1). The resulting empirical histograms of Z_S and Z_T are given in Figures 4 and 5 respectively.

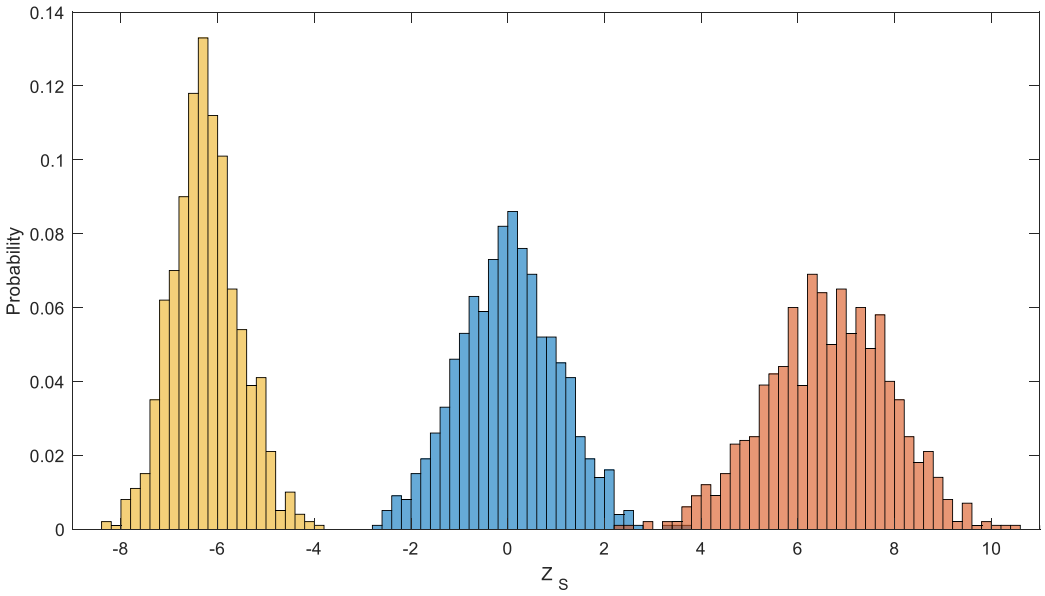


Figure 4. Empirical histograms of Z_S, normalized to probabilities, under three scenarios. The middle histogram is derived from 1000 replicates of Z_S in which, for each Z_s, 1000 π_i were generated from a uniform U(0, 1) distribution, the X_i were Bernoulli(π_i) random variables, and ES_n and $SD(S_n)$ were computed from these π_i. The middle histogram should approximate a N(0,1) distribution. Computing formulas are detailed in Section 2. The right histogram is derived from 1000 replicates of Z_S in which, for each Z_s, 1000 π_i were generated from a beta B(2,2) distribution, the X_i were Bernoulli(π_i) random variables, but ES_n and $SD(S_n)$ were taken from the U(0, 1) simulations. The left histogram is derived from 1000 replicates of Z_S in which, for each Z_s, 1000 π_i were generated from a beta B(.5, .5) distribution, the X_i were Bernoulli(π_i) random variables, but ES_n and $SD(S_n)$ were taken from the U(0, 1) simulations.

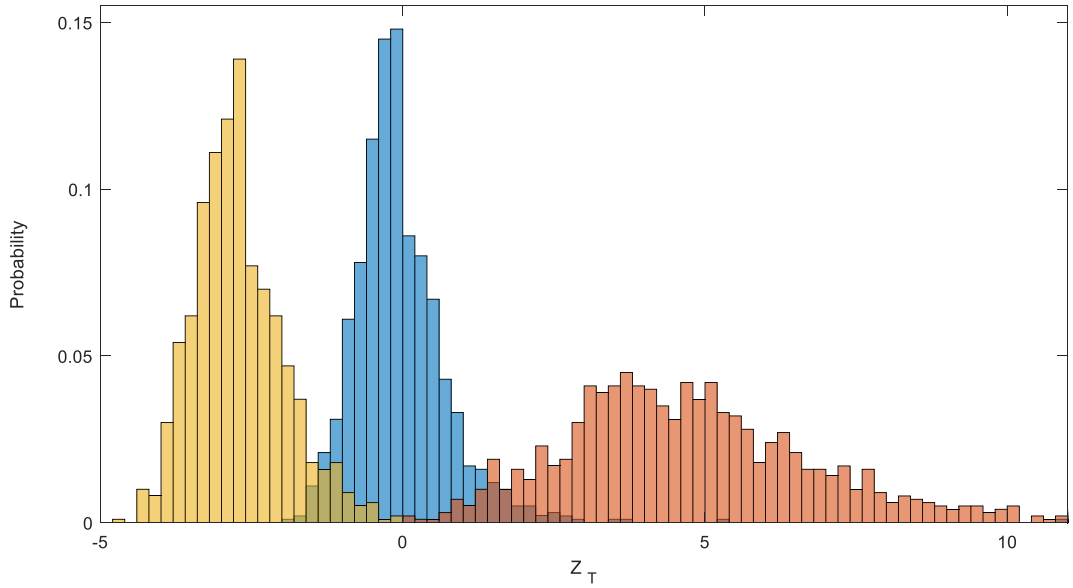


Figure 5. Empirical histograms of Z_T, normalized to probabilities, under three scenarios. The middle histogram is derived from 1000 replicates of Z_T in which, for each Z_T, 1000 π_i were generated from a uniform U(0, 1) distribution, the X_i were Bernoulli(π_i) random variables, ET_n=1000, and SD(T_n) was computed from these π_i. The middle histogram should approximate a N(0,1) distribution. Computing formulas are detailed in Section 2. The right histogram is derived from 1000 replicates of Z_T in which, for each Z_T, 1000 π_i were generated from a beta B(2,2) distribution, the X_i were Bernoulli(π_i) random variables, but ET_n and SD(T_n) were taken from the U(0, 1) simulations. The left histogram is derived from 1000 replicates of Z_T in which, for each Z_T, 1000 π_i were generated from a beta B(.5, .5) distribution, the X_i were Bernoulli(π_i) random variables, but ET_n and SD(T_n) were taken from the U(0, 1) simulations.

As with the previous example, Z_S and Z_T turn negative for one of the alternatives, here, when the π_i are from a Beta(.5, .5) distribution. Again, the Z_T observations are widely dispersed for one alternative, Beta(2,2). Relative to power, Sn would likely be preferred over T_n for these alternatives.

An example

The American College of Cardiology jointly with the American Heart Association have recently published a set of equations for estimating 10-year atherosclerotic cardiovascular disease (ASCVD) risk[10]. We will assess calibration of these risk equations in an independent cohort of individuals enrolled in MESA (multi-ethnic study of atherosclerosis). MESA was funded by the US National Heart, Lung and Blood Institute to study preclinical atherosclerosis, with the intent of identifying risk factors involved in the progression of atherosclerosis to clinical ASCVD. A total of 6800 men and women, aged 45 to 84, and free of ASCVD at baseline examination, were recruited into the study between July 2000 and September 2002.

We chose a cohort of 6520 individuals from the study, with full information on 10 year outcomes and clinical characteristics allowing calculation of the ACC/AHA risk equations. Of these individuals, 930 experienced an ASCVD diagnosis or event (including death) within 10 years, and 5590 did not. A calibration plot depicting the frequencies of the observed and predicted events is given in Figure 6. There seems to be reasonable agreement between the observed outcomes and the predicted outcomes from the ACC/AHA risk equations, though higher frequencies do not seem fully in accord, as suggested by the lowess smoother. The risk equations perform adequately at discriminating between individuals with or without ASCVD events, with an area under the curve of .753. On the other hand, the risk equations slightly underestimate the total numbers of events, with an expected to observed ratio of .934 (868.8/930).

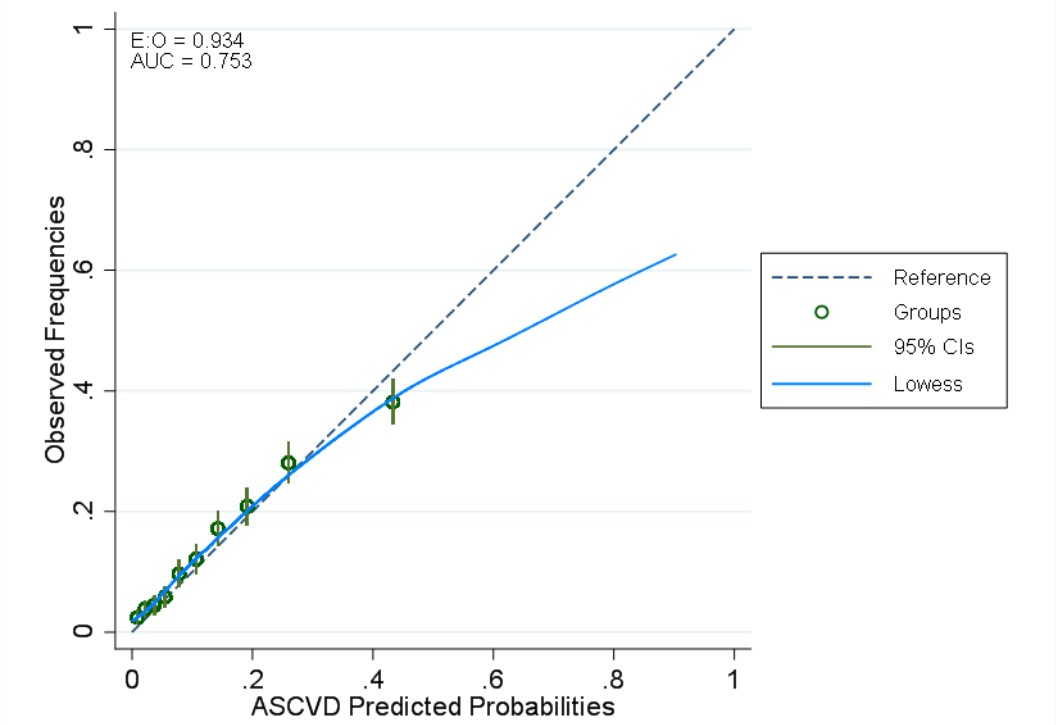


Figure 6. Calibration plot of prediction performance of the ACC/AHA risk equations applied to an independent cohort of 6520 individuals enrolled in the Multi-Ethnic Study of Atherosclerosis (MESA). The outcome of interest is occurrence of atherosclerotic cardiovascular disease (ASCVD) within 10 years. Predicted risks were used to divide the cohort into 10 equally sized groups. 95% confidence intervals for the observed proportions of events are shown for each of the 10 groups. A lowess smoother[15] is also depicted. Summary statistics O:E (observed:expected) = .934, and AUC (area under the curve) = .753, are also given. The figure was rendered in Stata 14 with the module PMCALPLOT[16].

Note that the preponderance of predicted risks from the ACC/AHA risk equations is quite small as shown in Figure 7. The calibration plot shows this from the groupings, but from Figure 7 it should be clear that selection of cutpoints for the groups, as for example with the Hosmer-Lemeshow chi-squared statistic, is somewhat arbitrary.

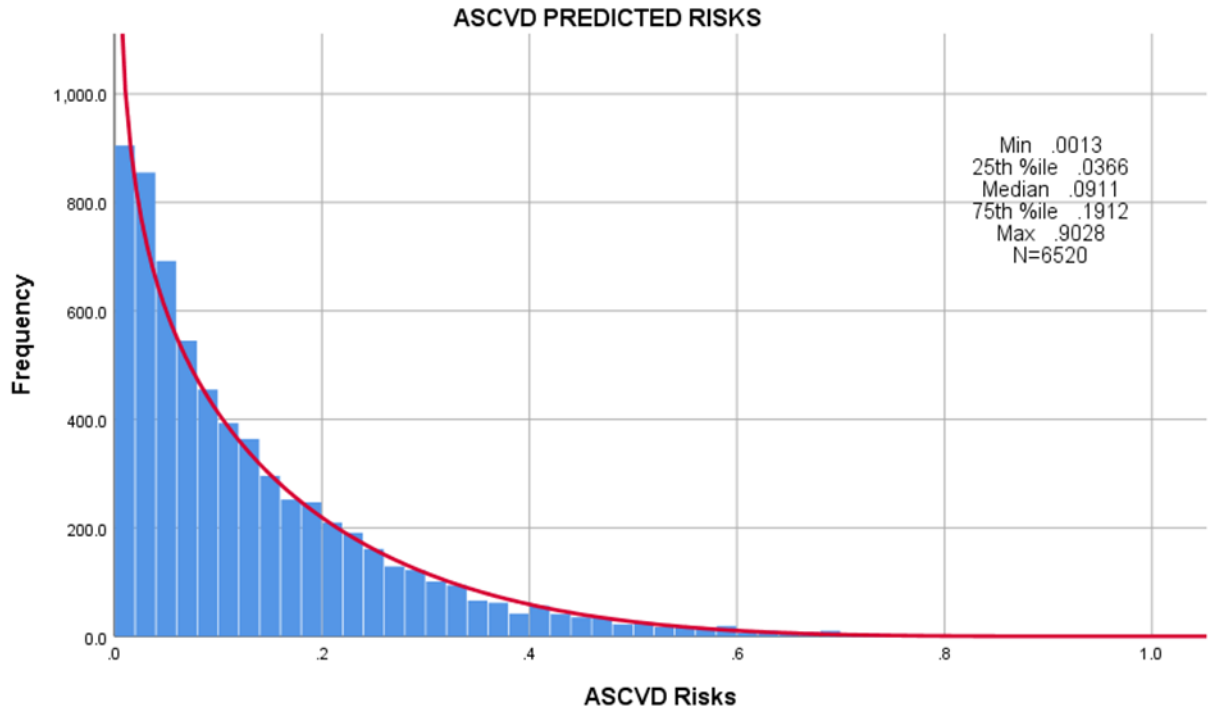


Figure 7. Histogram of the ACC/AHA risk equation predictions for the cohort of 6520 individuals in the MESA study. Summary statistics are also given. A beta function fit via maximum likelihood with estimated parameters a=.77, b=5.02 is depicted in red.

We proceed to assess calibration of the risk equations with the statistics introduced in Section 2. Calibration in this setting devolves to assessment of goodness-of-fit, that is, how well the predicted risks π_i from the ACC/AHA risk equations accord with the observed outcomes X_i ($X_i = 1$ for an ASCVD event within 10 years, 0 otherwise), for the cohort of 6520 individuals indexed by i. Summary statistics are given in Table 1.

Table 1. Summary statistics for assessing goodness of fit of the ACC/AHA risk equations.

Statistic	**Observed**	**Expected**	**Variance**	**Z Statistic**
S_n	718.24	642.19	253.03	4.78
T_n	9105.05	6520	164462.4	6.37
D_n	1360.43	1284.38	253.03	4.78

Notes: Observed and expected values, and variances, were calculated using the formulas in Section 2; the Z statistics are (Observed – Expected)/√(Variance).

As noted earlier, the Z statistics for S_n and D_n are numerically identical. The statistic T_n provides the strongest evidence against goodness of fit. For comparative purposes, we also computed the Hosmer-Lemeshow statistic, after dividing the cohort into 10 equally sized groups using predicted risks as in Figure 6. We found Hosmer-Lemeshow $X^2 = 42.3$, $p<10^{-5}$, consistent with Sn and Tn. It appears that the ACC/AHA risk equations are not very well calibrated with this cohort, in general providing slight underestimates of true risks especially when the risk predictions are small.

Discussion

Generally, the Hosmer-Lemeshow statistic compares observed and predicted events in 10 evenly spaced categories (deciles of risk). This is a convention and not a rigid rule, especially in situations in which the predicted risk is not evenly distributed across [0,1], as with the example in Section 4. Indeed, in a recent investigation that attempted to validate the ACC/AHA risk equations in a different cohort of nearly 11000 US adults[11] , patients were categorized into 4 groups according to their 10-year predicted ASCVD risk: less than 5%, 5% to less than 7.5%, 7.5% to less than 10%, and 10% or greater. [With our cohort, this grouping would yield bins with frequencies 2109 (32.3%), 744 (11.4%), 583 (9.0%), and 3084 (47.3%) respectively.]These authors also found that calibration for the overall population was poor: Hosmer-Lemeshow $X^2=84.2$, $p<.001$, though the level of statistical significance might reflect in part the large sample size[12].

As with the Muntner et al. study[11], calibration of clinical prediction models can involve sample sizes in the thousands, especially with data accruing from registries or long-term cohort studies. The standard normal approximations to the distributions of Z_S, Z_T, and Z_D should be appropriate in such settings. On the other hand, it is perhaps unexpected that two-sided alternatives to these limiting distributions ought to be considered in practice. One might alternatively invoke the chi-squared versions of these test statistics to avoid difficulties in interpretation.

A second implication of potentially large sample sizes is that, in such settings, one should be cautious about conflating statistical significance with practical import. In our example, underestimating the ASCVD risk in individual patients might have serious repercussions for clinical care, but this should be tempered with the realization that for most patients, the absolute risk is extremely small (Figure 7). Motivating patients to change behavior on the basis of small perceived risk is a fraught enterprise.

The two statistics S_n and T_n intrinsically differ in their weights assigned to the X_i: equal weights with S_n, but heavier weights for small or large π_i with T_n. Such differential weights are common with goodness-of-fit statistics, a close analog being Cramér-von Mises vs. Anderson-Darling quadratic tests based on the empirical distribution function. In the example in Section 4, Z_T seems to be more sensitive than ZS to a purported shift in magnitude of the π_i. On the other hand, one can envision scenarios in which the

differential weighting incorporated into T_n might be viewed as a detriment relative to power, as with the beta alternatives in the simulation study, or to over-dispersion.

We remark that the Copas statistic S_n is closely related to the Brier score[13] S_n/n, which has been widely studied and utilized both in the statistics literature[14] and in fields outside of traditional statistics[15]. There is no monopoly on these seminal ideas.

In summary, the Copas statistics are easily implemented, have proven value in other contexts, and appear to be underutilized in calibration studies. Along with Hosmer-Lemeshow, they ought to be part of the armamentarium of calibration tools for all researchers.

Future Directions

The Copas statistics have not enjoyed widespread interest in past years, and this brief study does not do them justice. In this regard, there are a host of follow-up studies that might be undertaken, as for example the following.

We utilized a normal approximation to the distributions of the Copas statistics in our example, noting that in this particular study, with a large sample size of well over 6000, the normal approximation should be adequate. In small samples (n=10 to 80), we found a simple Cornish-Fisher approximation to tail quantiles to be preferable to the normal approximation. Further investigation of the adequacy of approximations at intermediate sample sizes is clearly warranted. Are other approximations (e.g., Copas, saddlepoint) also worthwhile, especially with T_n? Is there anomalous behavior with π's near 0 or 1, compared to our uniform spacing of the π's? How are power properties related to alternatives of interest? A simulation study, perhaps patterned after HHLL[5], might examine power properties for various likely alternatives and provide guidance for definitive use in model calibration.

Acknowledgments

The author thanks Dr. Jill Waalen of The Scripps Research Institute for suggesting this problem, and Dr. Jonathan Unkart of the University of California, San Diego School of Medicine for providing the MESA data.

Thanks also to the reviewers, whose comments led to substantial improvements in the presentation.

Declaration of conflicting interests

The author declares no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

Research reported in this publication was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under Award Number 5P01HL119165.

References

1. Copas JB. Unweighted sum of squares test for proportions. Applied Statistics 1989;38:71-80.

2. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating (Statistics for Biology and Health). New York: Springer, 2009.

3. Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis , 2nd ed.(Springer

Series in Statistics). New York: Springer, 2015.

4. Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. Commun Stat Theor Meth 1980;9:1043-69.

5. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness of fit tests for the logistic regression model. Statistics in Medicine 1997;16:965-980.

6. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression 3rd Edition. New York: John Wiley, 2013.

7. Hand DJ. Measuring classifier performance: a coherent alternative to the area under the roc curve. Machine Learning 2009;77:103-23.

8. Austin PC, Reeves MJ. The relationship between the C-statistic of a risk-adjustment model and the accuracy of hospital report cards: a Monte Carlo study. Medical Care 2013;51:275-284.

9. Cornish EA, Fisher RA. Moments and cumulants in the specification of distributions. Revue de l'Institut International de Statistique / Review of the International Statistical Institute 1938;5:307–320.

10. Goff DC, Lloyd-Jones DM, Bennett G et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association task force on practice guidelines. Circulation 2014;129:S49

11. Muntner P, Colantonio LD, Cushman M et al. Validation of the atherosclerotic cardiovascular disease pooled cohort risk equations. JAMA 2014;311:1406-1415.

12. Paul P, Pennell ML, Lemeshow S. Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. Statistics in Medicine 2013;32:67-80.

13. Brier GW. Verification of forecasts expressed in terms of probability. Monthly Weather Review 1950;78:1-3.

14. Gerds TA, Cai T, Schumacher M. The performance of risk prediction models. Biometrical Journal 2008;50:457-479.

15. Bradley AA, Schwartz SS, Hashino T. Sampling uncertainty and confidence intervals for the Brier score and Brier skill score. Weather and Forecasting 2008;23:992-1006.

16. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Statistics in Medicine 2014;33:517-535.

17. Ensor J, Snell KIE, Martin EC. PMCALPLOT: Stata module to produce calibration plot of prediction model performance. Statistical Software Components S458486, Boston College Department of Economics, revised 1 November 2018.