1-7-2005

Standardizing Markers to Evaluate and Compare their Performances

Margaret S. Pepe
University of Washington, mspepe@u.washington.edu

Gary M. Longton
Fred Hutchinson Cancer Research Center, glongton@fhcrc.org

INTRODUCTION

Markers are sought to detect conditions or predict future onset of conditions. Examples include childhood screening tests, tests for genetic abnormalities, and markers for cardiovascular disease such as serum lipids and inflammatory indicators. Biomarkers for cancer detection include prostate specific antigen and CA-125. Some of these same markers are used as markers of treatment response and of disease progression. The emergence of new technologies such as gene and protein expression arrays promise the development of more sophisticated markers in the near future.[1,2]

The issue here is how to evaluate the performance of a marker. The importance of rigorously evaluating a marker's performance before it is adopted in routine medical practice is of particular concern to regulatory agencies and has recently been highlighted in the popular press.[3] The ultimate validation of a marker requires large population studies and consideration of disease-specific costs and benefits associated with incorrect and correct classification by the marker.[4] Preliminary to such studies are smaller studies that simply assess the marker's ability to discriminate subjects with the condition from those without. The statistical evaluation of a marker's discriminatory capacity is the specific topic we discuss in this paper.

How should one measure the discriminatory capacity of a marker? An appropriate measure should not depend on the measurement units of the marker. If it does, it cannot be used to compare markers measured in different units. For example, the odds ratio (or relative risk) per unit increase in the marker, although commonly used, is not a self-contained summary statistic of discrimination and cannot be compared across different markers.[6]

We propose an approach that first involves standardizing the marker values relative to a normative population (those without the condition). This standardization puts different markers

on a common scale, thereby facilitating comparisons amongst markers. In addition we show that the distribution of the standardized marker among subjects with the condition is closely related to the receiver operating characteristic (ROC) curve, a statistical tool that has long been used for evaluating diagnostic tests. [7—9] The ROC curve is appropriate for evaluating the discriminatory capacity of any marker.[10] Its interpretation as relating to the distribution of standardized marker values is appealing. In particular it may be of interest to those researchers already comfortable with statistical concepts of standardization and frequency distributions, but who are not familiar with ROC analysis.

METHODS

Datasets

To illustrate concepts we apply statistical techniques to two simple datasets. The data are online at http://www.fhcrc.org/labs/pepe/book/. In the first, two serum biomarkers for pancreatic cancer, CA-125 and CA19-9, were measured for 90 patients with pancreatic cancer and 51 without[11] (Figure 1). Questions of interest are: (i) how to quantify the capacities of the two markers to distinguish between the patients with and without cancer; and (ii) to compare the two markers.

The second dataset pertains to a marker of hearing impairment at the 1416 Hz freqeuncy for 57 hearing impaired ears and 147 unimpaired ears. The marker is the signal-to-noise ratio (SNR) from the distortion product otoacoustic emissions (DPOAE) test. The test was performed using 9 different sound stimulus intensity levels, 3 of which are included in this dataset. Thus for each ear we have an SNR value for each of the intensity levels (Figure 2). Details of the original study and data selection can be found in Stover et al[12] and in Pepe,[13] respectively.

Both of these studies employed case-control designs. Cross-sectional cohort studies can be analysed in the same way.

Approach

To explain the general approach we adopt the convention that higher values of the marker are more indicative of the presence of the condition. (We can always redefine the marker if necessary to ensure this, using negation for example. See the audiology data in Figure 2.) The basic idea is to use the distribution of marker values in the *unaffected* population, *without* the condition, as a reference distribution for standardizing marker values in the *affected* population, i.e., those *with* the condition. The standardization for an affected subject with marker value *Y* is simply to calculate the frequency of unaffected subjects with marker values greater than *Y*. Thus if marker values for 20% of unaffected subjects exceed *Y*, the standardized marker value is 0.20. We call the standardized value its *placement value*.[14—16]

The concept of calculating a placement value is closely related to that of calculating a percentile value relative to a healthy reference population as is the common practice for reporting anthropometric measurements in children.[17] Here, rather than reporting the percentile, the proportion of the reference population *less* than *Y*, we report the proportion *greater* than *Y*. Although the concepts are equivalent, we will see that calculation of placement values rather than percentiles facilitates connections with ROC methodology.

Placement values are proportions taking values between 0 and 1. Since higher marker values are more indicative of the condition, having the condition is associated with having smaller placement values. The smallness of the placement value indicates how extreme a subject's marker value is relative to the reference population. Moreover, a marker for which most

affected subjects have very small placement values is a good marker because it identifies most affected subjects as being extreme relative to the reference population.

A key attribute of placement values is that they do not have measurement units associated with them. Different markers are converted to a common scale by the placement value standardization. This facilitates comparisons amongst them. Thus if a diseased subject has a placement value of .50 for marker 1 and placement value .01 for marker 2, then marker 2 is the better disease indicator for him. He is identified as extreme in regards to marker 2 while he appears to be well within the reference (non-diseased) population in regards to marker 1. To determine which of two markers is better at discriminating *the population* of affected subjects from the unaffected population, one must consider *the population distributions* of placement values in affected subjects for each of the markers. The marker with a higher frequency of small placement values is preferred.

ROC Curve

The ROC curve is a statistical device for illustrating the classification accuracy achievable with a diagnostic test, or marker.[9,10,16] For each possible threshold value, c, one can define a positive classification rule based on the marker, $Y \geq c$ indicating that the condition is present. The associated true positive rate (TPR(c)) and false positive rate (FPR(c)) are

TPR (c) = proportion of affected subjects with $Y \geq c$

and

FPR (c) = proportion of unaffected subjects with $Y \geq c$,

respectively. The ROC curve plots TPR (c), the test sensitivity, versus FPR (c), 1-specificity, for all values of c. It shows the range of (FPR, TPR) achievable. Since good classification accuracy pertains to low FPRs and high TPRs, a good marker has an ROC curve with points in the upper

left corner of the $(0,1)\times(0,1)$ square. The area under the ROC curve (AUC) is the most popular ROC summary statistic. An AUC of 1.0 corresponds to a perfect marker.

RESULTS

Pancreatic Cancer Biomarkers

Using the 51 subjects without pancreatic cancer as the reference group we standardized each of the markers for the 90 subjects with pancreatic cancer by calculating placement values. The frequency distributions are displayed in Figure 3a. The CA-19-9 placement values are smaller than the CA-125 values indicating that pancreatic cancer patients are more extreme relative to the non-cancer reference in regards to CA19-9 than in regards to CA-125

The average (sd) of the placement values is .14 (.26) for CA 19-9 and .29 (.25) for CA-125. A simple paired t-test could be applied to compare the averages. However it is not quite appropriate because a finite sample of only 51 non-cancer patients was used to standardize the markers. A different sample of non-cancer patients would have produced a somewhat different standardization. The sampling variability in the reference group used to calculate the placement values for the 90 diseased subjects must be accounted for in calculating a *p*-value that compares mean CA-19-9 and CA-125 placement values. The bootstrapping technique[18] described in the electronic appendix does this and yields $p < .01$.

The scatterplot (Figure 3(b)) shows that although CA19-9 is the better marker overall, there are a substantial number of cancer patients for whom CA-125 is better in the sense that they are normal in regards to CA19-9 but abnormal in regards to CA-125. For example, 5 patients with CA-19-9 placement values exceeding 20% had CA-125 values less than 10%.

Audiology Testing

Distributions of standardized –SNR values (negative SNR) are shown in Figure 4 for hearing impaired subjects. It appears that the test is more discriminatory when the sound stimulus is at a lower intensity since the placement values are smaller at the 55dB intensity level versus at the 60 and 65 dB levels. The average (sd) values are .029 (.057), .053 (.106), and .071 (.127), respectively. The *p*-value for comparing the averages at 55 and 65dB is < .01 using the bootstrap technique. Interestingly the 55 dB stimulus appears to work better than the 65dB stimulus for most individuals as can be seen from the scatterplot in Figure 4 (b). That is, the test results for most hearing impaired subjects appeared more abnormal with the lower intensity stimulus, as evidenced by smaller placement values.

Relationship with ROC analysis

Figures 3(c) and 4(c) show the cumulative distributions (cdf) of standardized markers for cancer patients and for hearing impaired subjects respectively. The cdf corresponding to p on the x-axis is the proportion of values that are $\leq p$. Interestingly these cumulative distribution curves are identical to ROC curves for the markers. The general argument is as follows: Let c be the threshold value that corresponds to the false-positive rate *p*, $\text{FPR}(c) = p$. Consider the point cdf(*p*) on the cumulative distribution curve. Observe that a subject's placement value is $\leq p$ if and only if his marker value $Y \geq c$. Therefore the proportion of affected subjects with placement values $\leq p$, namely cdf(*p*), is equal to the proportion with marker values $\geq c$, i.e., $\text{TPR}(c)$. So, each point (*p*, cdf(*p*)) on the cumulative distribution curve is a point (FPR(c), TPR(c)) on the ROC curve and vice versa. A mathematical argument is given in Pepe and Cai.[15]

There are two interpretations then for the curves shown in Figures 3(c) and 4(c). Interpreted as cumulative distribution functions, we see the proportion of affected subjects with standardized marker values as or more extreme than *p*. Interpreted as ROC curves we see the

trade-offs between sensitivity and specificity that are possible when we apply thresholding classification rules to the marker in the population. Both interpretations are meaningful and useful. The accuracy of CA19-9 for classifying subjects with or without pancreatic cancer is clearly superior to CA-125. For example, the thresholding rule with specificity of 80% (FPR=.20) yields a sensitivity of 78% for CA19-9 but only 49% for CA-125. Said another way, 78% of cancer patients have standardized CA-19-9 below 0.2 while only 49% have standardized CA-125 below 0.2. Similarly we see from Figure 4 (c) that classification accuracy is better when the lower sound intensity is employed.

ROC Summary Statistics

The areas under the ROC curves in Figure 3(c) are .86 for CA 19-9 and .71 for CA-125. Those in Figure 4(c) yield AUCs of .97 at 55 dB, .95 at 60 dB, and .93 at 65 dB. Observe that these are exactly the same as 1 minus the mean placement values calculated earlier. The result holds in general that averaging standardizing markers for affected subjects yields 1–AUC.

$$\text{average (placement value)} = 1 - \text{AUC}.$$

It is intuitive for the perfect marker since all placement values for affected subjects are equal to 0 and AUC = 1 for the perfect marker. Mathematical arguments for the general result are available.[14,15]

The implication of this result is that statistical comparisons between markers using areas under ROC curves are the same as statistical comparisons between markers using placement value averages for diseased subjects. Therefore the *p*-values cited earlier that pertain to average placement values are also valid for comparing the AUCs in Figures 3(c) and 4(c).

DISCUSSION

The main contribution of this paper is to suggest a standardization procedure to facilitate the evaluation of markers. Use of a reference distribution is a familiar concept. In laboratory medicine for example values outside of a normal healthy reference range often flag patients as having a medical condition. Standardization with respect to an age and gender matched reference population is used for anthropometric measurements. Standardization not only provides better clinical interpretations but makes possible valid comparisons of different populations. Our standardization can be used to compare a marker's discriminatory capacity across different populations. One could compare placement values in diseased men and diseased women, for example, to determine if the marker performs better in men or women. An additional compelling attribute of the standardization we propose is that it makes possible valid comparisons of *different markers* across the *same* population, as demonstrated with our two datasets.

We also noted the close connection between analysing standardized markers of affected subjects and ROC analysis. With our approach one can analyze standardized markers in familiar ways, as we did for pancreatic cancer and hearing impairment markers, without explicitly considering operating characteristics of thresholding decision rules. Nevertheless we have shown that such considerations are implicitly at play and the approach is fundamentally the same as ROC analysis.

Our approach offers avenues for addressing questions that should be, but are typically not asked about marker performance. In particular, regression analysis applied to placement values can be used to determine if covariates affect the capacity of a marker to distinguish cases from controls.[15,19] Covariates may relate to characteristics of subjects tested or to the test itself. [16] To

illustrate with the audiology data, the following linear regression model was fit to the placement values for hearing impaired subjects:

$$Z = Z\,(\text{placement value}) = \alpha_0 + \alpha_1\ \text{Intensity} + \varepsilon$$

with Z, the normal deviate corresponding to the placement value, and covariate, Intensity, being the sound stimulus intensity. The estimate $\alpha_1 = 0.023$ (95% confidence interval =(-.019, .076); se=.024) indicates a trend for higher intensity levels being associated with larger placement values among hearing impaired subjects, i.e., reduced marker performance. Figure 4c shows the corresponding cumulative distributions of placement values. The more frequent occurrence of small placement values at the lower intensity levels is obvious from these curves. The better performance at lower intensity is also evident with the ROC curve interpretation. More complex models that include multiple independent variables simultaneously can easily be fit too. As noted earlier, bootstrapping is applied to arrive at appropriate standard errors and *p*-values. Alternatively, recent work[15,19] provides theory for making statistical inference about regression models using placement values.

REFERENCES

1. The Chipping Forecast. (1999). *Nat Genet.* 21 (suppl.)
2. Liotta LA, Ferrari M, Petricoin E. Clinical proteomics: written in blood. *Nature.* 2003;425(6961):905.
3. Pollack A. "New Cancer Test Stirs Hope and Concern." *New York Times.* 3 Feb. 2004, late ed., final: F1.
4. Etzioni R, Urban N, Ramsey S, et al. The case for early detection. *Nat Rev Cancer*. 2003 Apr;3(4):243–52.
5. Pepe MS, Etzioni R, Feng Z, et al. Phases of Biomarker Development for Early Detection of Cancer *J Natl Cancer Inst.* 2001;93:1054–1061.
6. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epid.* 2004;159(9):882–890
7. Hanley JA. Receiver Operating characteristic (ROC) methodology: The state of the art. *Crit Rev Diag Imag.* 1989;29:307–335.
8. Begg CB. Advances in statistical methodology for diagnostic medicine in the 1980s. *Stat Med*. 1991;10(12):1887–1895.
9. Zhou SH, McClish DK, Obuchowski NA. *Statistical methods in diagnostic medicine*. Wiley, New York; 2002.
10. Baker SG. The Central Role of Receiver Operating Characteristic (ROC) Curves in Evaluating Tests for the Early Detection of Cancer. *J Natl Cancer Inst.* 2003;95:511–515.

11. Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*. 1989;76:585–592.

12. Stover L, Gorga MP, Neely ST, Montoya D. Toward optimising the clinical utility of distortion product otoacoustic emission measurements. *J Acoust Soc Am*. 1996;100:956–967.

13. Pepe MS. Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics*. 1998;54:124–135.

14. Hanley JA and Hajian-Tilaki KO. Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update. *Acad Rad*. 1997;4:49–58.

15. Pepe MS, Cai T. The analysis of placement values for evaluating discriminatory measures. *Biometrics*. 2004;60(2):528–35.

16. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press; 2003.

17. Frischancho AR. *Anthropometric standards for the assessment of growth and nutritional status.* Ann Arbor: University of Michigan Press; 1990.

18. Efron B, Tibshirani RJ *An introduction to the bootstrap*. Chapman and Hall; 1993.

19. Cai T. Semi-parametric ROC regression analysis with placement values *Biostatistics*. 2004;5:45–60.

Appendix: Testing for a difference in the mean Placement Value between two disease markers: Bootstrap estimation of the achieved significance level.

With paired observations on the two markers, A and B, the test statistic is the average placement value difference between markers,

$$\hat{\theta} = \frac{1}{n}\sum_i \left(W_i^B - W_i^A\right),$$

where n = number of cases, and W_i is the placement value of the marker for case i and the superscript indicates the marker. The null hypothesis to be tested is $H_0 : \theta = 0$. The sampling variability of placement values calculated for disease cases depends not only on marker variability among the cases but also among the controls used to estimate the reference distribution.

In order to approximate the null distribution of $\hat{\theta}$ and estimate the achieved significance level, we sampled from the empirical distribution of $\hat{\theta}$ and centered the distribution at zero, the desired null mean. Specifically, samples of paired marker observations, equal in size to the original case and control samples, were drawn separately, with replacement, from the observed case and control samples. The test statistic, $\hat{\theta}_k$, was calculated for each set of case and control "bootstrap" samples, $k = 1, \ldots 1000$, and translated to conform to the null distribution by subtracting the original $\hat{\theta}_{obs}$ from the bootstrap sample, $\hat{\theta}_k^* = \hat{\theta}_k - \hat{\theta}_{obs}$. The achieved significance level of the test was then calculated as the proportion of the bootstrap $\hat{\theta}_k^*$'s more extreme than the observed $\hat{\theta}_{obs}$, i.e. $\left|\hat{\theta}_k^*\right| \geq \left|\hat{\theta}_{obs}\right|$.

FIGURE LEGENDS

Figure 1. Distributions of pancreatic cancer biomarkers in 51 subjects with pancreatic cancer and 90 subjects without cancer.

Figure 2. Distributions of –SNR values from the DPOAE test in 57 ears with hearing impairment and in 147 ears without impairment at the 1416 HZ frequency. The test was applied using input stimulus of intensities 55 dB, 60 dB and 65 dB. Shown are –SNR values (rather than SNR values) to agree with the convention of higher marker values being more indicative of hearing impairment.

Figure 3. Distributions of standardized biomarkers (placement values) in 90 subjects with pancreatic cancer. Shown are (a) frequency distributions (b) scatter plots and (c) cumulative distributions .

Figure 4. Distributions of placement value standardized –SNR in 57 hearing impaired ears at 3 stimulus intensity levels (55 dB, 60 dB, and 65 dB). Shown are (a) frequency distributions, (b) scatter plots and (c) cumulative distributions .

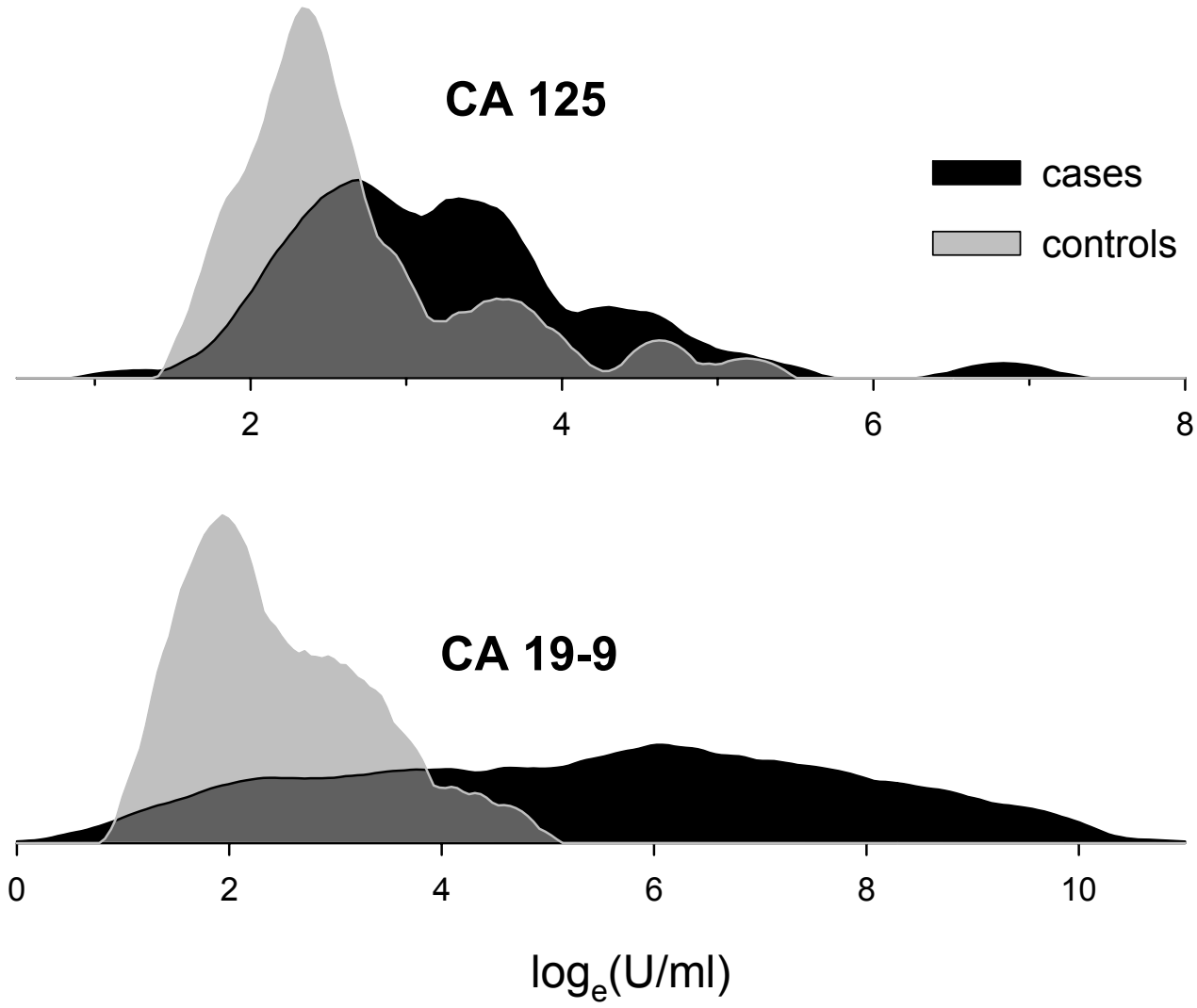

Figure 1

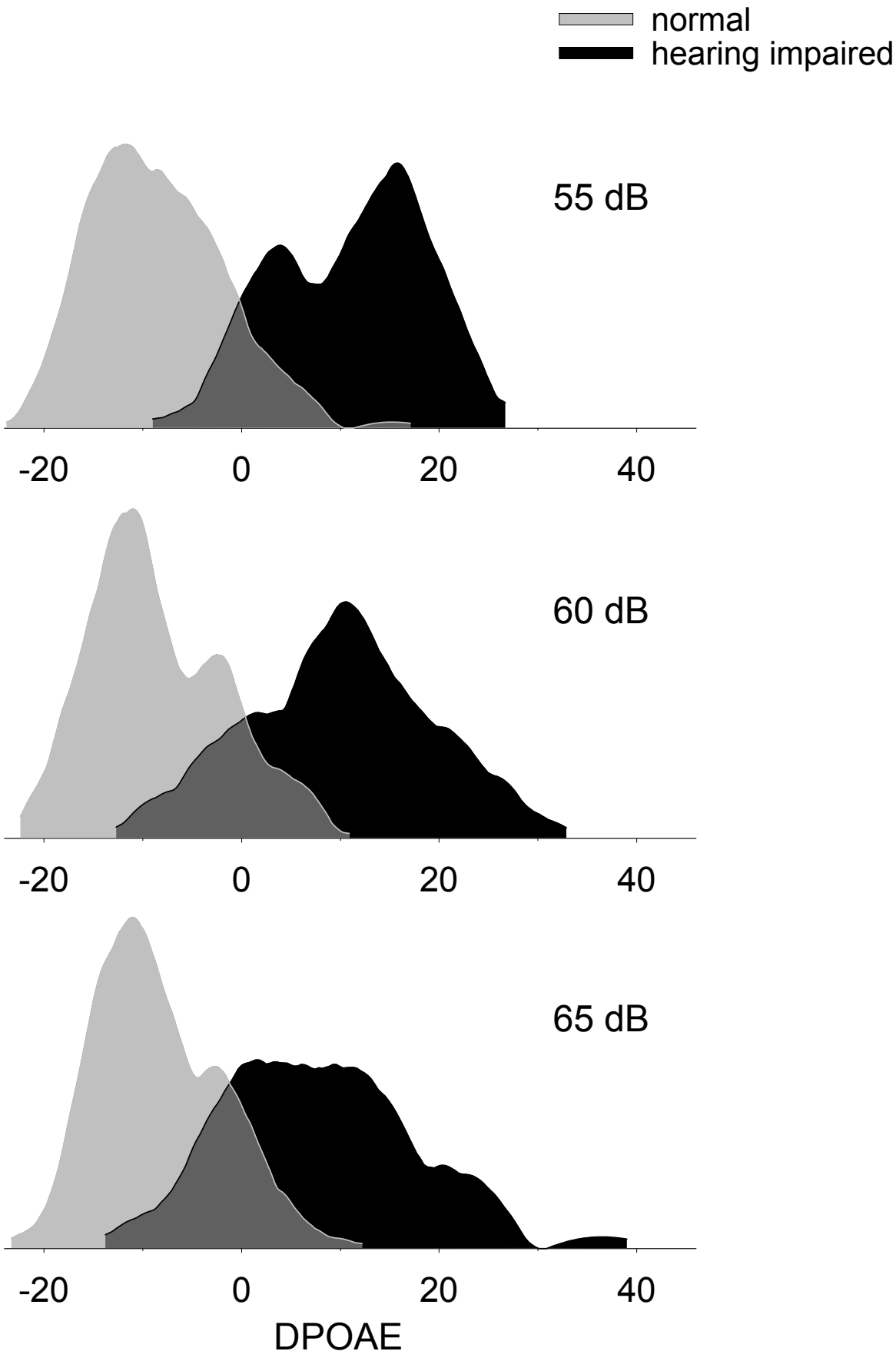

Figure 2

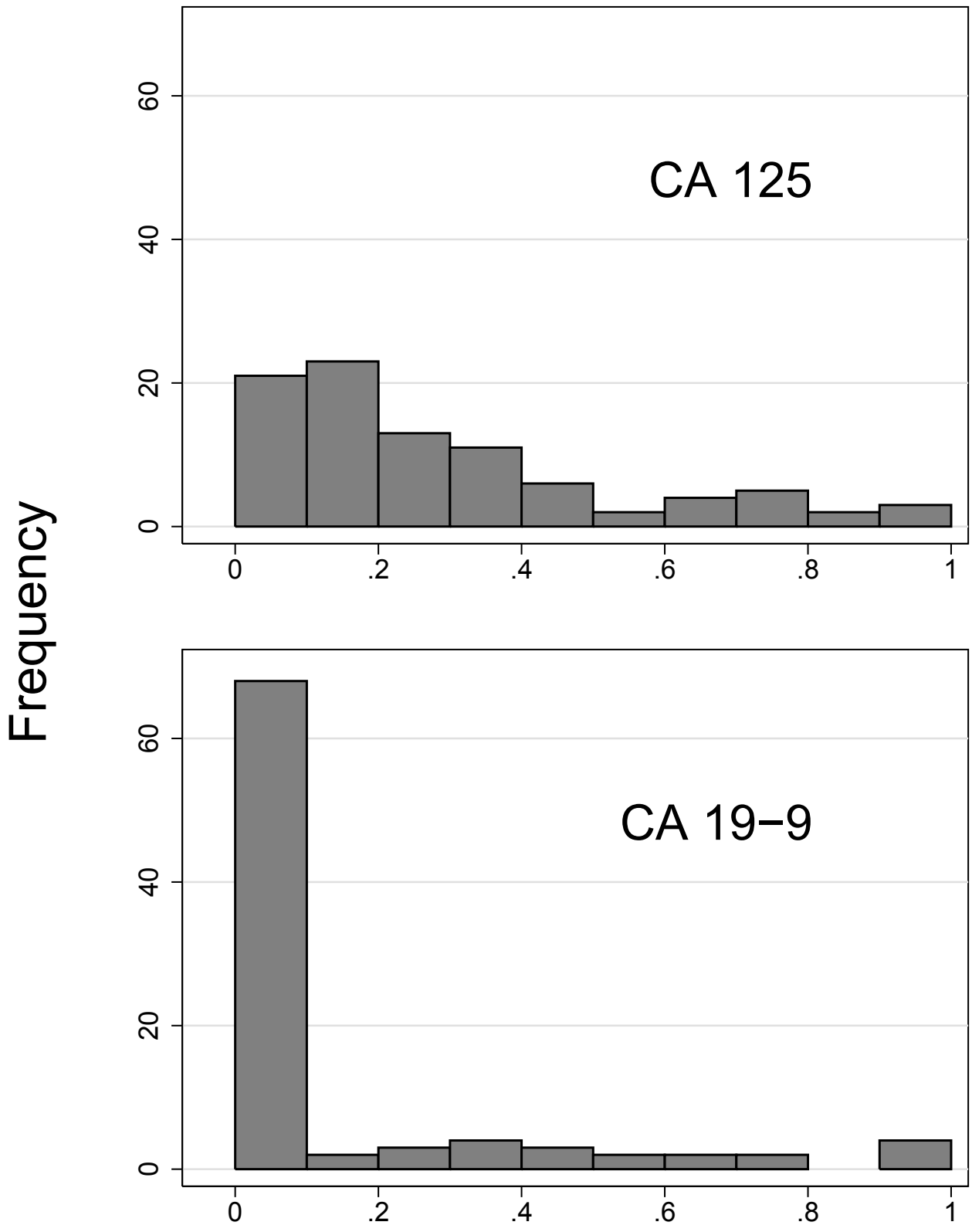

Figure 3a

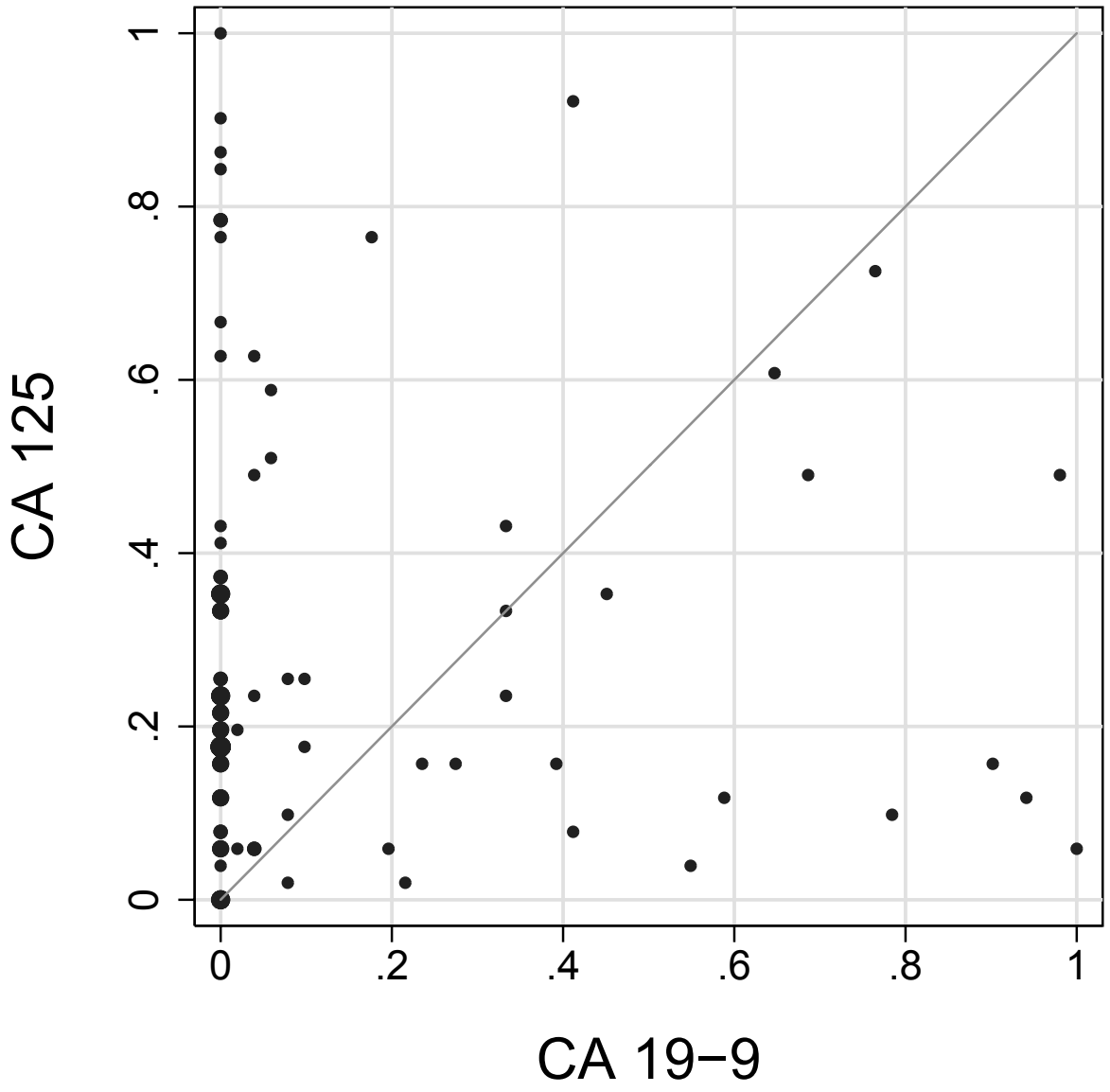

Figure 3b

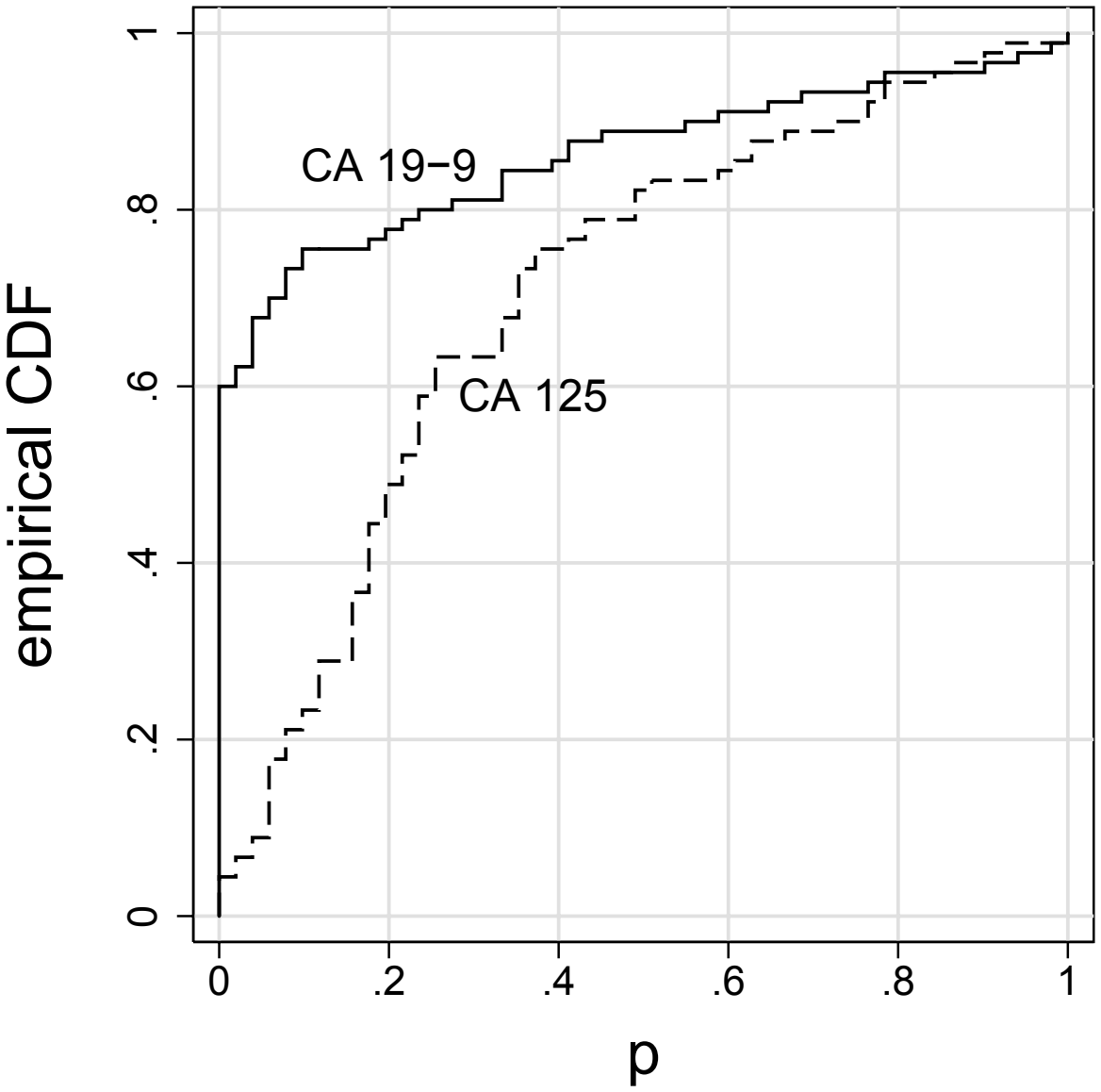

Figure 3c

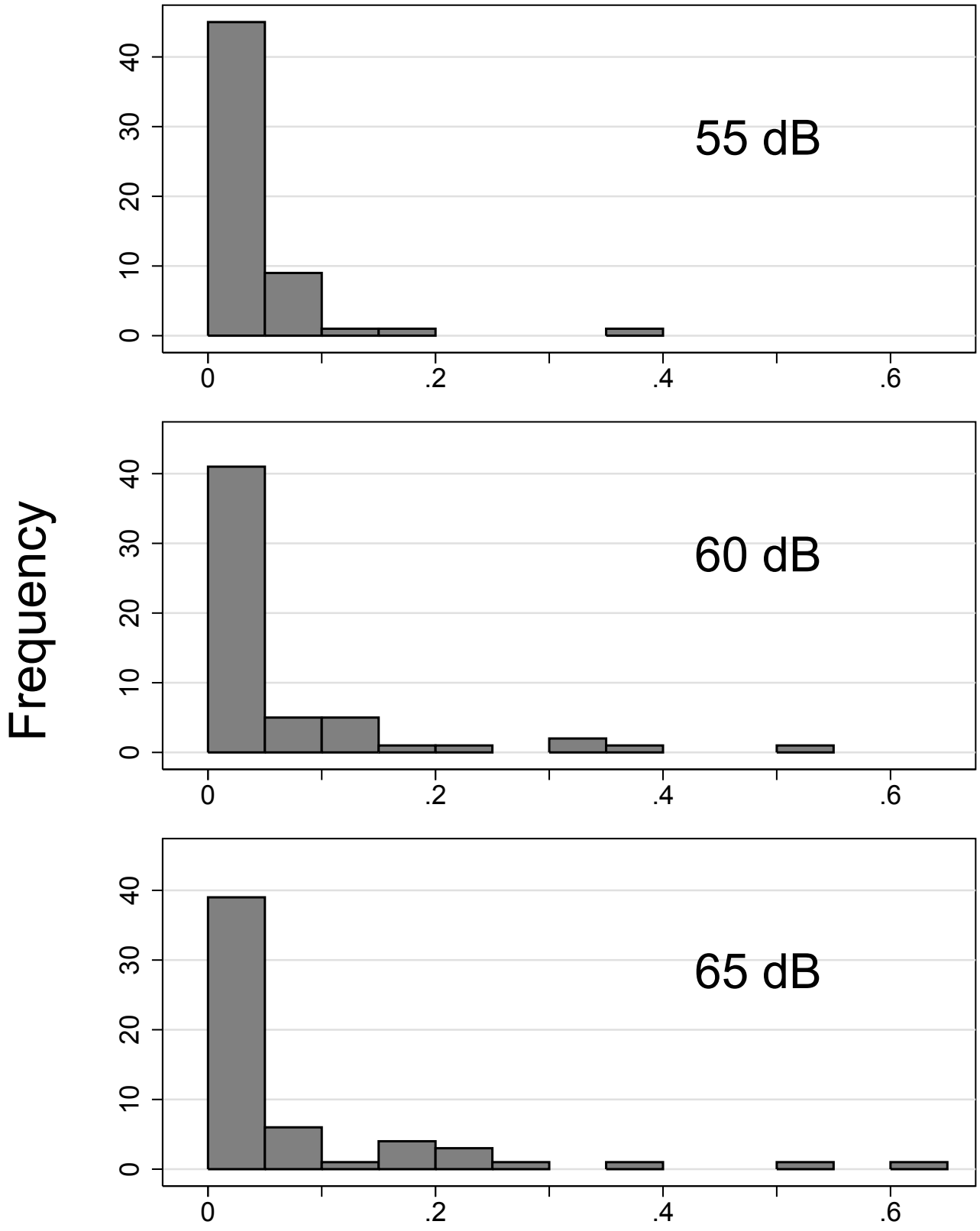

Figure 4a

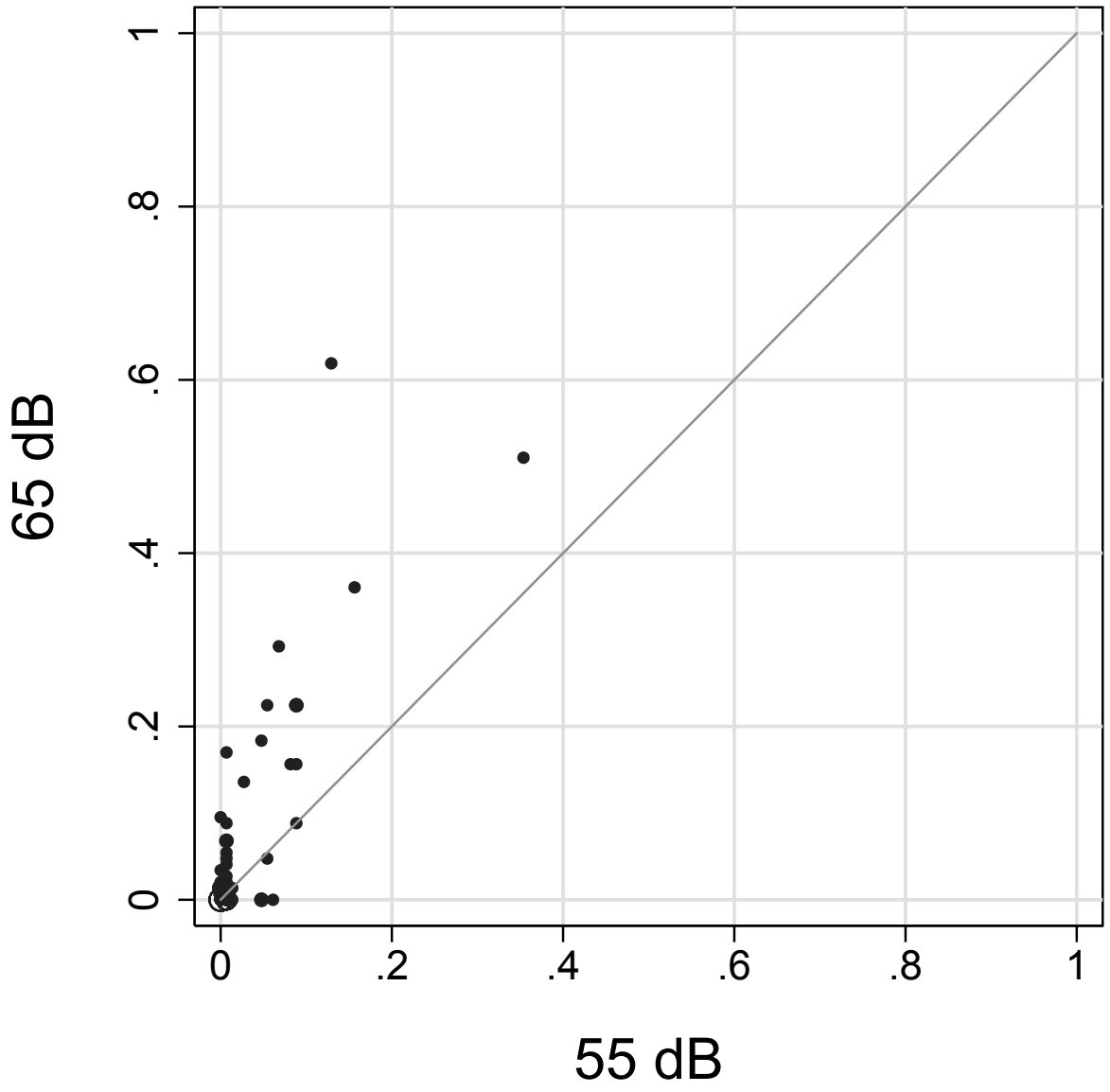

Figure 4b

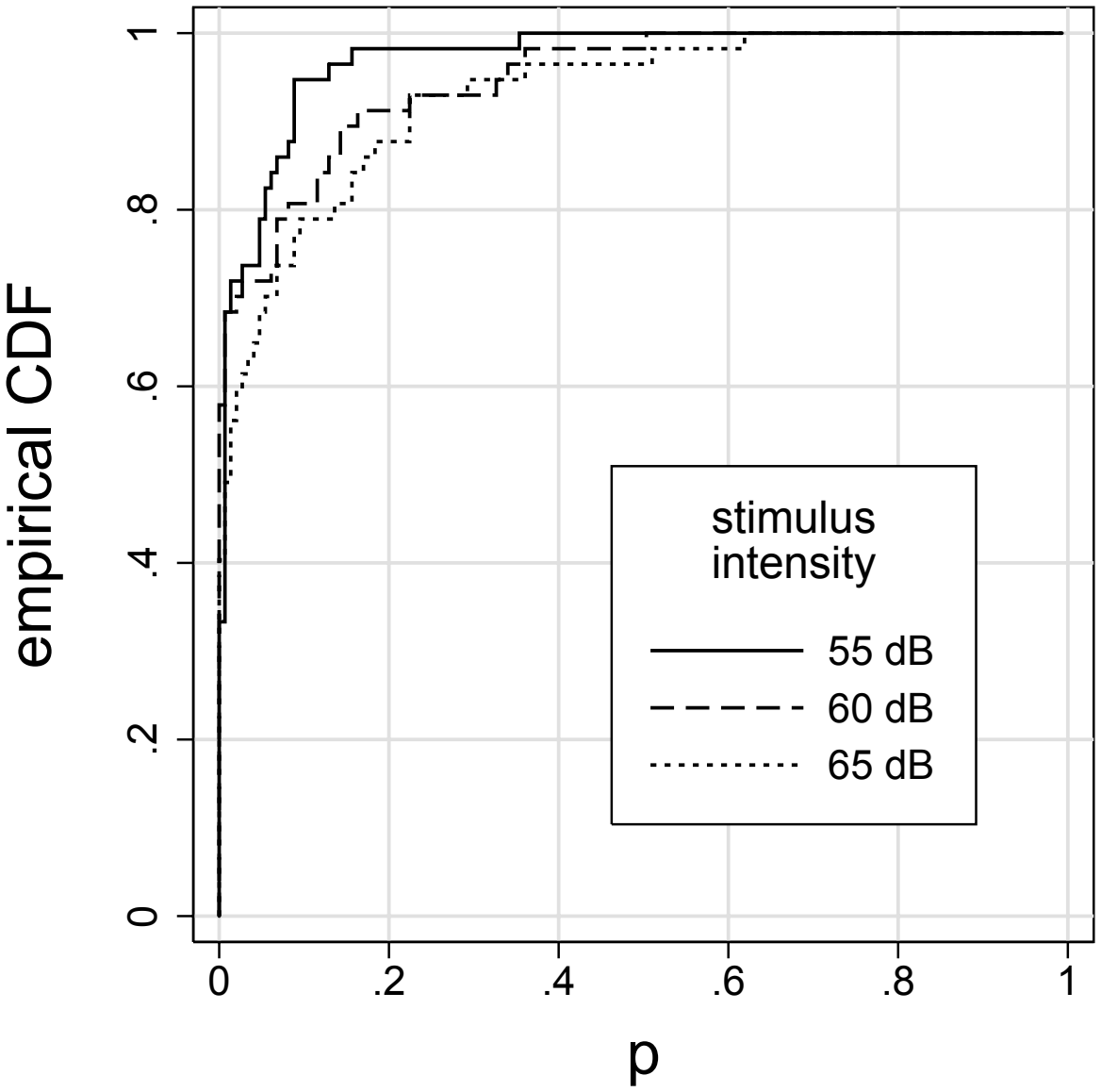

Figure 4c