Kernel Testing as an Alternative to χ^2 Analysis for Investigating the Distribution of Quantitative Traits

Jeffrey D. Hart, Anna Hale, J. Creighton Miller, Jr.*

Summary. Chi-square analysis is a popular means of analyzing data from inheritance studies. Such an analysis, though, can be inappropriate when the phenotypic characteristic under study is a continuous variable. We argue that arbitrary categorization of continuous variables can either seriously invalidate or make powerless a chi-square goodness-of-fit test. We propose an alternative test based on kernel density estimates. The test makes no assumption about the underlying distribution of the phenotypic variables. Furthermore, kernel density estimates provide an appealing means of describing the distribution of continuous variables.

Key words. Bootstrap; Chi-squared test; Cross-validation; Genotype; Inheritance studies; Kernel density estimator; Phenotype.

*Jeffrey D. Hart is Professor, Department of Statistics, Texas A&M University, College Station, TX, 77843. Anna Hale is Research Geneticist, USDA-ARS-Sugarcane Research Laboratory, 5883 USDA Road, Houma, LA, 70360. J. Creighton Miller, Jr. is Professor, Department of Horticultural Sciences, Texas A&M University, College Station, TX, 77843. The research of Professor Hart was supported in part by NSF Grant DMS-0604801.

1 Introduction

Data from inheritance studies are often analyzed using a chi-square analysis. However, a categorical analysis is not always appropriate for the phenotype being evaluated. Qualitative traits are considered to be less influenced by environment than quantitative ones, but environment can still influence the phenotypic expression of qualitative traits, resulting in misclassification in genetic studies. In addition, traits that are continuously expressed are frequently artificially categorized so that a chi-square analysis can be used. For example, in numerous studies on the inheritance of resistance to iron deficiency chlorosis in dry bean (*Phaseolus vulgaris*), there was a continuous distribution in symptoms ranging from highly resistant to susceptible. In order to analyze these data, a subjective rating was given to each, and they were artificially grouped into categories for analysis (Coyne, et al. 1982 and Zaiter, Coyne and Clark 1988). In a study on the inheritance of stem color and dwarf growth habit in alfalfa (*Medicago polymorpha*), there were difficult-to-classify intermediate plants; however, a chi-square analysis was used to analyze the data (De Haan and Barnes 1998).

The incorrect specification of a plant's genotype will be referred to as a *classification error*. Such an error can occur when one attempts to determine genotype by observing phenotypic expression of characteristics that are continuous and hence not definitive. Classification errors can have a serious adverse effect on the results of a chi-square analysis. An incorrect categorization of a trait displaying continuous variation can make results fit a genetic theory too well or indicate a deviation from a theory when none exists.

We propose two methods for overcoming the shortcomings inherent in arbitrary choice of categories. Both methods are applicable as long as sufficient information is available concerning plant genotype and phenotypic variation. In the first, a

chi-square analysis is used with an *objective* choice of categories. Such a choice ensures that the probability of a type I error in testing the fit of a genetic model is equal to the nominal level. A second approach is based on kernel density estimates; see Parzen (1962) and Silverman (1986). We prefer this latter approach for at least two reasons. First, it provides an elegant means of describing the distribution of a phenotypic characteristic in a set of observed data. Secondly, it requires no categorization of continuous variables, which makes for a more convenient method and leads, in certain situations, to a test more powerful than the chi-square test. An inheritance study on the genetics of resistance to iron deficiency chlorosis in cowpea provides a good example of how a genetic theory can be tested by comparing kernel density estimates. This example will be used subsequently to illustrate the proposed methodology.

2 A motivating problem

Plant material iron deficiency chlorosis is a problem in cowpea (*Vigna unguiculata*) because it affects the ability of the plant to produce chlorophyll. Pinkeye Purple Hull (PEPH) a susceptible cultivar, and Texas Pinkeye Purple Hull (TXPE), a resistant cultivar, were crossed and allowed to self for one generation. The F1's and reciprocals were backcrossed to the parents, and all populations were planted in the field in a randomized block design. SPAD readings were taken on each population. SPAD measures the transmission of light through the leaves at a wavelength where chlorophyll absorbs and a wavelength where it does not. The SPAD value is calculated based on a ratio of these two numbers. Thus, the SPAD value is unitless and is an indication of the relative amount of chlorophyll present in the leaf. High SPAD readings correspond to dark green leaves, and low SPAD readings correspond to yellow leaves. Since the cultivars (pure lines) used are completely inbred, there

should be no genetic variation within the parental and the F1 populations. Thus, the parental means can be used as an estimate of the two alternative homozygous conditions, and the F1 means can be used as an estimate of the heterozygous condition. A weighted average of the variation in these three populations can be used to estimate environmental variation. If iron deficiency chlorosis is conditioned by a single gene, the F2 population should segregate in a 1:2:1 ratio of homozygous susceptible to heterozygous to homozygous resistant. The single gene theory can thus be tested by evaluating the F2 plants for the expected segregation ratios. Unfortunately, because of environmental variation, none of the plants was purely green, yellow-green or yellow, which made it difficult to differentiate between plants that were heterozygotes and those that were homozygous for the proposed resistance allele or for the proposed susceptible allele.

Let X_{P1} be the statistical designation for the SPAD reading of a randomly selected plant of susceptible Pinkeye Purple Hull (P1), and use a similar notation for the other three populations. Due to the aforementioned environmental variation, values of X_{P1} will vary about some average SPAD reading. We shall model this variation by assuming that X_{P1} has probability density function g_{P1}; likewise, resistant Texas Pinkeye Purple Hull (P2), F1, and F2 have densities g_{P2}, g_{F1} and g_{F2}, respectively. If the single gene theory is true, then

$$g_{F2} = .5g_{F1} + .25(g_{P1} + g_{P2}). \tag{1}$$

3 Chi-square methodology

When environmental variation is sufficiently small relative to genetic variation, a chi-square analysis seems to be a reasonable means of testing the single gene theory. To illustrate this idea, suppose that X_{P1}, X_{P2}, and X_{F1} are each normally distributed with the same variance σ^2 but different means.

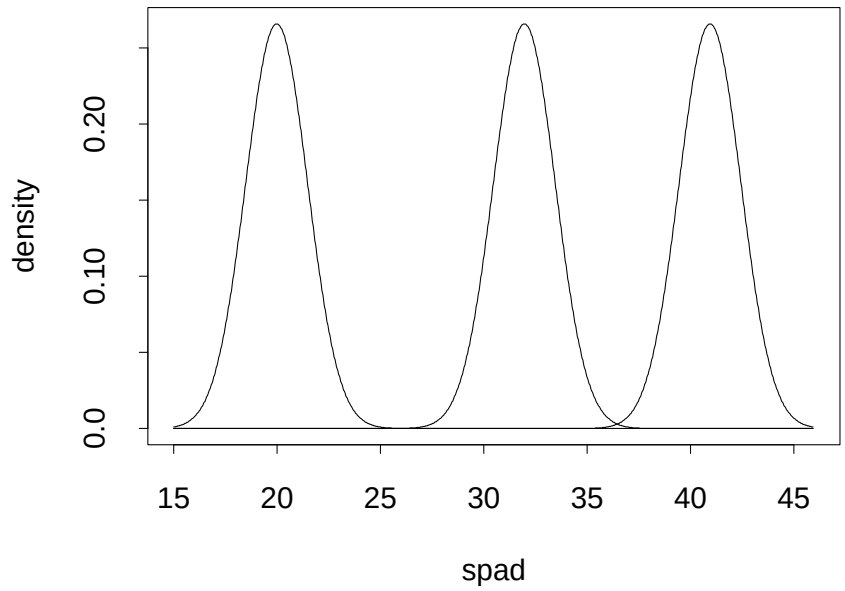

Figure 1: *Hypothetical SPAD distributions. From left to right are hypothetical SPAD distributions for P1, F1, and P2 populations, respectively. Here, the environmental variation is small enough that cowpea genotype can be effectively determined by observing a SPAD reading.*

If σ^2 is sufficiently small relative to the differences between means, then for all practical purposes, genotype can be determined by observing a SPAD reading. This situation is illustrated in Figure 1, where the respective means are 20, 32 and 41.

Suppose the single gene theory is true in the case just described. Then in the F2 generation we expect 25% of SPAD readings to be less than 26, about 50% between 26 and 36.5 and 25% greater than 36.5. A valid test of the single gene theory is one based on the chi-square statistic

$$\chi^2 = \frac{(O_1 - .25n)^2}{.25n} + \frac{(O_2 - .50n)^2}{.50n} + \frac{(O_3 - .25n)^2}{.25n}, \tag{2}$$

where O_1, O_2 and O_3 are the observed numbers of F2 plants in the three categories and n $(= O_1 + O_2 + O_3)$ is the total number of F2 plants observed.

In many cases, one encounters a much more ambiguous situation than that depicted in Figure 1. In experiments conducted at Texas A&M University, data were collected on 36 susceptible Pinkeye Purple Hull plants (P1), 35 resistant Texas Pinkeye Purple Hull plants (P2), and 68 F1 plants. The respective SPAD means for the P1, F1, and P2 plants were 19.5, 32.0 and 41.0 and the standard deviations were 8.3, 8.3 and 6.1. Thus there was a great deal of overlap between the three sets of readings. Attempting to determine genotype by SPAD readings alone would lead to numerous classification errors, regardless of the classification scheme used.

As mentioned above, it is not an uncommon practice to test the single gene theory using a chi-square statistic of the form (2) based on category cutoff points that are more or less arbitrarily chosen. When the phenotypic expression of the trait under study is sufficiently variable, this practice can seriously invalidate the usual chi-square test. To see why, first suppose that the single gene theory is true, the phenotypic trait is continuous, and there is overlap among the P1, P2, and F2 distributions. Then there is one and only one choice of cutoff points, call them c_1 and c_2, such that the large sample probability distribution of χ^2 will be χ^2 with 2 degrees of freedom. Any other choice of cutoff points will induce the type I error probability of the usual chi-square test to approach 1 as the F2 sample size becomes larger and larger. In other words, a seemingly small error in choosing category cutoffs can virtually guarantee significance of a chi-square test even though the single gene theory is true.

What are the "correct" cutoff points for a chi-square test? These are the values c_1 and c_2 such that 1/4 of the mixture distribution (as defined in (1)) is less than c_1 and 1/4 is larger than c_2 (meaning that 1/2 of the distribution is between c_1 and c_2). These values ensure that the large sample type I error probability is equal to the

stated nominal level. When phenotypic data are available for P1, P2 and F1, these cutoff points may be estimated from the observed data. This leads to an objective χ^2 test that is approximately valid even in settings with large phenotypic variation. This objective method will be illustrated in Section 5 using our cowpea data.

Suppose the single gene theory is *not* true, in which case one wishes to have a test with high power. We will assume that c_1 and c_2 are chosen so that the type I error probability of a chi-square test is some desired value, say 0.05. The problem here is that the proportion of the F2 distribution lying below c_1 and the proportion above c_2 do not uniquely determine that distribution. The F2 distribution could deviate dramatically from the mixture (1) and yet the proportion of F2 SPAD readings in $(0, c_1]$, $(c_1, c_2]$ and (c_2, ∞) could be 1/4, 1/2 and 1/4, respectively. If such were the case the power of the chi-square test would be only 0.05 even though the single gene theory is decidedly false. Figure 2 provides an example of such a troublesome case.

We have argued that arbitrarily chosen categories can lead to seriously invalid tests of a single gene theory. An objective method of choosing categories is proposed that avoids this problem. However, even when this objective method is used, a chi-square test can suffer from extremely low power in certain cases where the single gene theory is not true. The testing procedure described in the next section overcomes this problem by using essentially all the information available in the data concerning the F2 distribution.

4 Kernel methodology

A classical means of testing hypothesis (1) is to assume a parametric model for each of the four distributions and then perform a likelihood ratio test. There are at least two problems with this approach: (i) using an incorrect parametric model can invalidate the test, and (ii) it is not always clear how to parametrize the model to

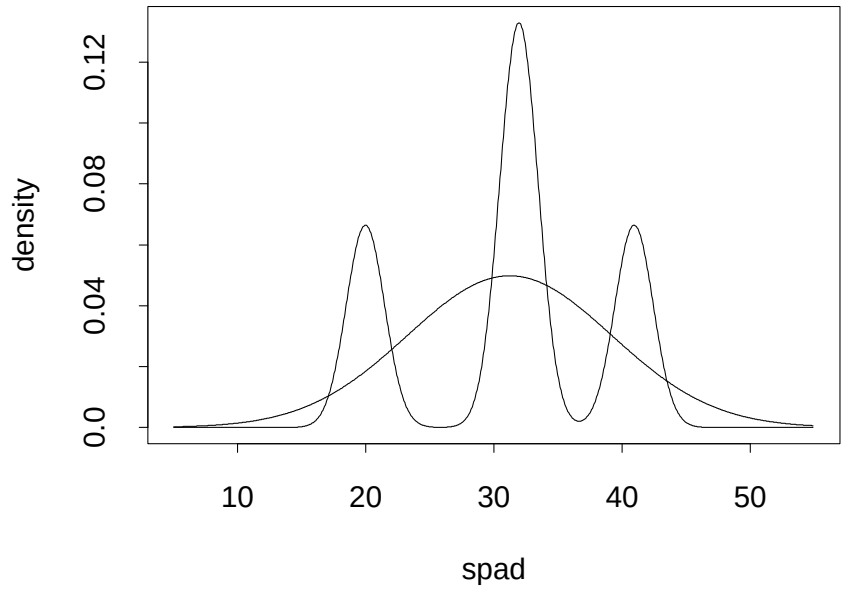

Figure 2: *SPAD distributions for which a χ^2 test has poor power. The trimodal density is a 1/4, 1/2, 1/4 mixture of the densities in Figure 1 and represents a hypothetical F2 distribution matching the single gene theory. The other density is a normal distribution with the same quartiles as the mixture. Were the typical 0.05 level chi-square test used to compare these two distributions, they would be judged insignificantly different with probability 0.95.*

reflect deviations from the null hypothesis. The second problem has an impact on power of the test.

A traditional nonparametric method for comparing distributions is to use a Kolmogorov-Smirnov, a Cramér-von Mises or some other related test based on empirical distribution functions. Such tests often suffer from poor power, as discussed in Eubank, Hart and LaRiccia (1993).

The testing approach proposed here is to construct kernel-type estimates of the four component densities, and to then compare the appropriate mixture of the P1, P2, and F1 estimates with the estimated F2 density. One advantage of this approach is that it often has better power than tests based on empirical distribution functions. Another attractive feature of the test is that it supplements an informative graphical description of the data. Histograms and/or kernel estimates are arguably better means of describing "central" features of data distributions than are cumulative distributions. For example, multimodality shows up much more readily in density estimates. The last point is particularly important in the genetics problem addressed herein, since trimodality of the F2 distribution is a natural consequence of the single gene theory.

The null hypothesis does not restrict the P1, P2 and F1 distributions. One expects the three populations of cowpeas to have different mean SPAD readings, whereas large differences in distribution *shapes* are not expected. A simple model would be to assume that each population is normally distributed, with all three having the same variance but possibly different means. This would leave four parameters to be estimated on the assumption that hypothesis (1) is true.

The nonparametric testing method to be proposed here can accommodate a variety of assumptions on the distributions of interest. However, to illustrate the method by example, we will use its most general form in which all four distributions are allowed to be arbitrary. Given a set of observations $X_1, \ldots, X_n$, a kernel density

estimate, $\hat{g}_h$, is defined as

$$\hat{g}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \quad \text{for each } x,$$

where K is usually a probability density symmetric about 0, and h is a positive number called the *bandwidth*. A popular choice for K is the standard normal density. The bandwidth controls the smoothness of the density estimate; small values of h produce rough estimates and large values of h very smooth estimates. A kernel estimate may be regarded as a "moving" histogram that centers a bin at each x and counts the number of data lying within this bin. The bandwidth plays the same role as does the bin width for a histogram.

Kernel density estimates for the four sets of cowpea SPAD readings are shown in Figure 3. A standard normal kernel and a bandwidth of 3.5 were used to obtain each estimate. (Bandwidth selection will be discussed later.) The four kernel estimates will be denoted $\hat{g}_{F1}$, $\hat{g}_{P1}$, $\hat{g}_{P2}$ and $\hat{g}_{F2}$. When the null hypothesis (1) is true, we would expect $\hat{g}_{F2}$ to be relatively close to $\hat{g}_{mix}$, where

$$\hat{g}_{mix}(x) = .5\hat{g}_{F1}(x) + .25\left(\hat{g}_{P1}(x) + \hat{g}_{P2}(x)\right).$$

We can measure the overall discrepancy between $\hat{g}_{F2}$ and $\hat{g}_{mix}$ by using the following *Kullback-Leibler* type discrepancy:

$$T = \int_{-\infty}^{\infty} \log\left(\frac{\hat{g}_{mix}(x)}{\hat{g}_{F2}(x)}\right) d\hat{G}_{mix}(x),$$

where $\hat{G}_{mix}$ is the appropriate mixture of the empirical distribution functions for the F1, P1 and P2 data. The quantity T will be our test statistic for testing the null hypothesis (1). An advantage of T is that it has a particularly simple computational form that involves no integration. We have

$$T = \frac{1}{2n_{F1}}\sum_{i=1}^{n_{F1}} \log\left[\hat{g}_{mix}(X_{F1,i})/\hat{g}_{F2}(X_{F1,i})\right] +$$

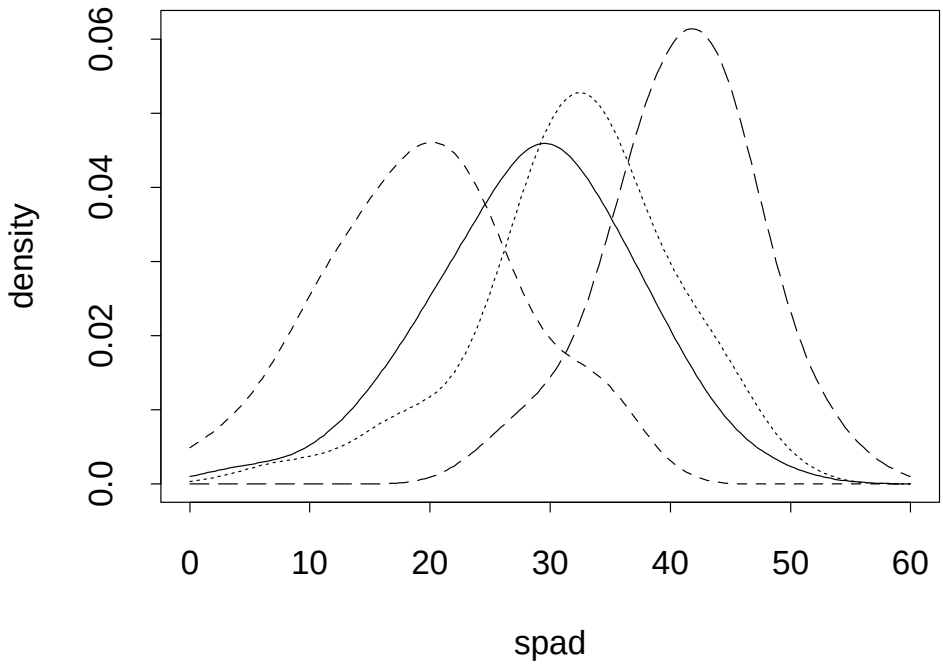

Figure 3: *Density estimates for cowpea data. The F2 estimate is the solid line, F1 is the dotted line, and P1 and P2 are the two dashed lines.*

$$\frac{1}{4n_{P1}} \sum_{i=1}^{n_{P1}} \log\left[\hat{g}_{mix}(X_{P1,i})/\hat{g}_{F2}(X_{P1,i})\right] +$$
$$\frac{1}{4n_{P2}} \sum_{i=1}^{n_{P2}} \log\left[\hat{g}_{mix}(X_{P2,i})/\hat{g}_{F2}(X_{P2,i})\right],$$

where n_{F1}, n_{P1} and n_{P2} are the respective sample sizes for the F1, P1 and P2 data sets. When hypothesis (1) is true, T should be relatively close to 0, and when false T will be a relatively large positive number.

Two problems remain if we are to use the statistic T to test (1). First we must have a rule for selecting the bandwidths of the density estimates, and second we need to find critical values for the test and/or be able to approximate the P-value corresponding to an observed T. To deal with the first problem, we simplify matters by using the same bandwidth for each of the four kernel estimates on which T relies. Using a common bandwidth implies that the expected value of $\hat{g}_{F2}(x)$ is the same as that of $\hat{g}_{mix}(x)$ for each x under null hypothesis (1). The principle of using a common bandwidth when comparing curves has been advocated by Young and Bowman (1995).

We are still faced with the problem of choosing the common bandwidth of the four estimates. To do so, we will use a weighted form of cross-validation. For a kernel density estimate $\hat{g}_h$ based on data $X_1, \ldots, X_n$ and bandwidth h, define the cross-validation function $CV(\hat{g}_h)$ by

$$CV(\hat{g}_h) = \int_{-\infty}^{\infty} (\hat{g}_h(x))^2 \, dx - \frac{2}{n} \sum_{i=1}^{n} \hat{g}_h^i(X_i), \tag{3}$$

where $\hat{g}_h^i$ is exactly the same density estimate as $\hat{g}_h$ but calculated leaving out the data value X_i. Rudemo (1982) and Bowman (1984) have shown that $CV(\hat{g}_h)$ is an unbiased estimator of a squared error risk criterion. The cross-validation bandwidth is the one that minimizes $CV(\hat{g}_h)$. In our problem we have four sets of data and hence will use the following weighted cross-validation:

$$WCV(h) = w_{P1}CV(\hat{g}_{P1,h}) + w_{P2}CV(\hat{g}_{P2,h}) + w_{F1}CV(\hat{g}_{F1,h}) + w_{F2}CV(\hat{g}_{F2,h}), \tag{4}$$

where the weights (for example w_{P1}) are the ratios of sample size to total sample size. The bandwidth we shall use in comparing the density curves will be the minimizer of $WCV(h)$.

The second problem that must be dealt with is determining the sampling distribution of the test statistic T. Finding this distribution exactly and/or asymptotically (as sample sizes tend to ∞) is beyond the scope of this article. In any event, the distribution depends on the unknown densities g_{F1}, g_{P1} and g_{P2}, necessitating some form of empirical approximation to the sampling distribution. Here we shall use the bootstrap (Hall 1992) to approximate the distribution of T. We resample from the observed data in such a way that the null hypothesis (1) is true for the simulated experiment. In a single bootstrap replication, we thus resample from the original P1, P2, and F1 data sets to obtain bootstrap versions of those data, and then obtain bootstrap F2 data by resampling from the appropriate mixture of the original P1, P2, and F1 data. The statistic T is calculated for a set of bootstrap data in exactly the same way it was for the original data. This process is replicated a large number of times, usually at least 1000, and the resulting values of the test statistic allow an approximation of its sampling distribution. One detail worth mentioning here is that some version of the "smoothed" bootstrap (Hall 1992, pp. 157-58) is required, due to the fact that cross-validation is used to choose a bandwidth on each bootstrap sample. The repeat values obtained when using the ordinary, unsmoothed, bootstrap make it much too likely that cross-validation will choose the smallest bandwidth under consideration. Further details concerning our bootstrap method are given in Section 5.

As mentioned previously, the kernel method can accommodate various assumptions about the data distributions. In the approach just described, the P1, P2, and F1 distributions were allowed to be arbitrarily different. In some cases it might be reasonable to assume that the three distributions are identical except for having

different means. In this scenario, the distributions of $X_{P1} - \mu_{P1}$, $X_{P2} - \mu_{P2}$, and $X_{F1} - \mu_{F1}$ are the same, where μ denotes population mean. The common distribution of the three populations can be estimated by a single kernel estimate using the pooled data $X_{P1,i} - \bar{X}_{P1}$, $i = 1, \ldots, n_{P1}$, $X_{P2,i} - \bar{X}_{P2}$, $i = 1, \ldots, n_{P2}$, $X_{F1,i} - \bar{X}_{F1}$, $i = 1, \ldots, n_{F1}$, where $\bar{X}$ denotes sample mean. An individual density estimate is then obtained by simply translating this estimate so that its mean is equal to the appropriate sample mean. The main advantage of this approach is that, when the equality-of-distributions assumption is true, the common distribution is estimated more efficiently by pooling the data than by computing three separate kernel estimates. The pooling technique was applied to the P1, P2, and F1 cowpea data using a standard normal kernel and a bandwidth of 3.5; the resulting estimates are shown in Figure 4. The Figure 4 estimate of, for example, the P1 density has the following weighted average form:

$$w_{P1}\hat{g}_{P1,h}(x) + w_{P2}\hat{g}_{P2,h}(x + \bar{X}_{P2} - \bar{X}_{P1}) + w_{F1}\hat{g}_{F1,h}(x + \bar{X}_{F1} - \bar{X}_{P1}),$$

where the weights and the kernel estimates are as in (4).

5 Data analysis

We now illustrate the methodology described in Sections 3 and 4 using data from plant breeding experiments at Texas A&M University. Figure 5 provides a visual test of the single gene theory. The dashed line is $\hat{g}_{mix}$, the 1/4, 1/2, 1/4 mixture of the P1, F1 and P2 density estimates in Figure 3, and the solid curve is $\hat{g}_{F2}$. The same bandwidth was used for all four density estimates and was chosen by the weighted cross-validation method described in Section 4. Figure 5 casts considerable doubt on the truth of the single gene theory for these data. However, there is still the question of statistical significance. How likely is it that such a large difference in

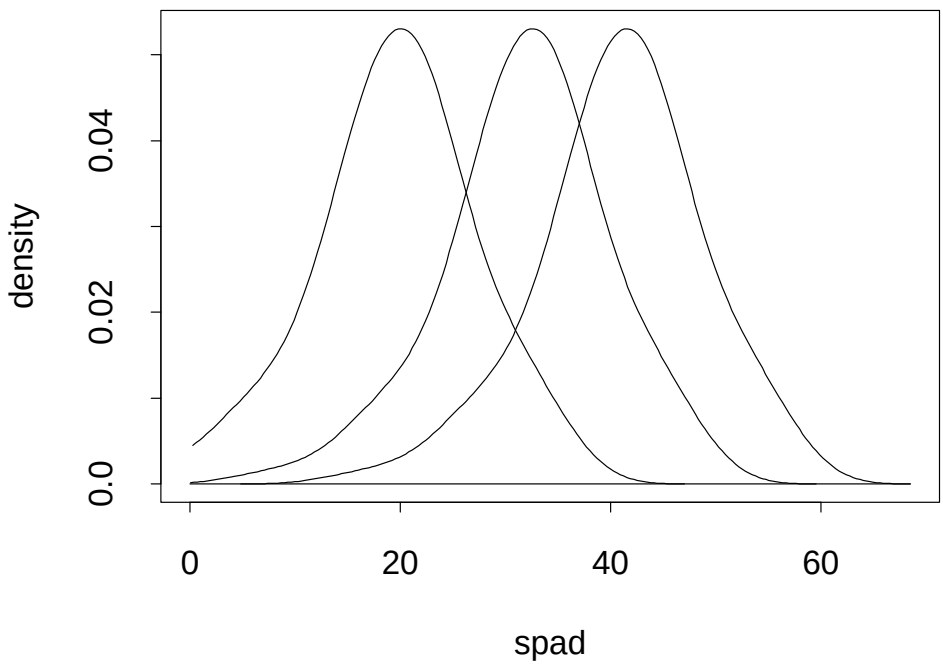

Figure 4: *Density estimates for cowpea data assuming equal shapes. From left to right the estimates correspond to P1, F1 and P2, respectively.*

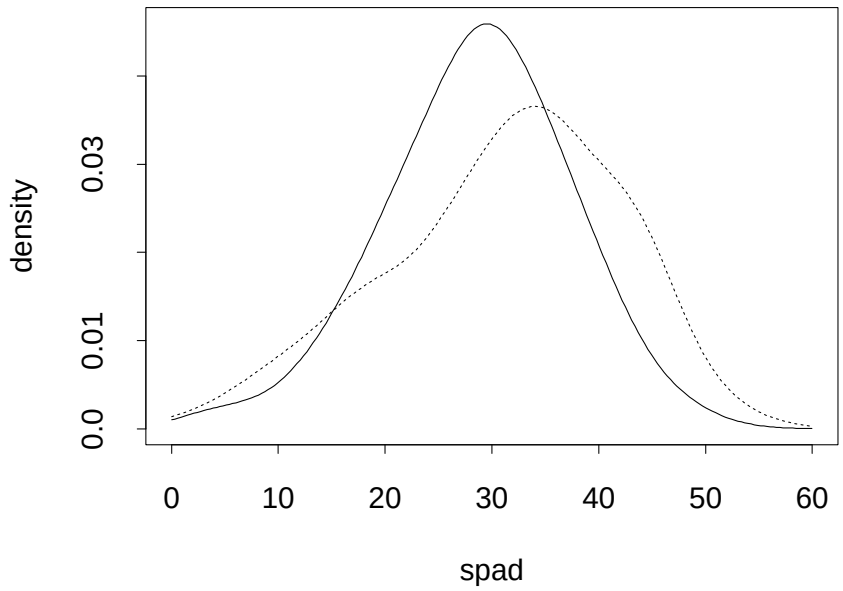

Figure 5: *Comparison of F2 and mixture densities. Solid line: F2 Dashed line: mixture of P1, P2 and F1*

density estimates would occur if in fact the single gene theory is correct? We may answer this question by computing the test statistic T and comparing it with the sampling distribution of T on the assumption that hypothesis (1) is true.

The value of T for the observed data is 0.0994, which is not particularly meaningful without knowledge of the sampling distribution of T. Two methods of simulating bootstrap data were used to approximate this distribution. In one method, it was assumed that the F1, P1, and P2 populations were normally distributed with means and variances equal to the sample means and variances of the three data sets. The second bootstrap method used was nonparametric in that data were resampled from the nonparametric density estimates $\hat{g}_{P1}$, $\hat{g}_{P2}$, and $\hat{g}_{F1}$. This process is facilitated using the fact that a kernel estimate is a convolution of an empirical distribution and

the kernel used to compute the estimate. In other words, to randomly select a value from, say, $\hat{g}_{P1}$, one randomly selects a value from the P1 data set and adds it to a randomly selected value from a normal distribution having mean 0 and standard deviation equal to the bandwidth of the kernel estimate.

The two bootstrap methods led to very similar distributions of bootstrapped test statistics, as shown in Figure 6. The first, or parametric, bootstrap method yielded an estimated P-value of .012, i.e., in only 12 of the 1000 bootstrap samples generated was the simulated value of T larger than 0.0994. So, the result pictured in Figure 5 is very unlikely on the assumption that the single gene theory is true, at least when the P1, P2, and F1 populations are assumed to be normally distributed. The nonparametric bootstrap yielded similar results. Only 11 of 1000 bootstrap samples produced a value of T larger than 0.0994, meaning that the estimated P-value is .011. Again, there is clear evidence that Figure 5 is inconsistent with the single gene theory.

When categories are chosen as described in Section 3, the chi-square test is in agreement with the kernel-based test. Two probability models were used to choose the cutoff points. In the first, the P1, F1, and P2 populations were assumed to be normally distributed with the same variance. The common variance was estimated by a weighted average of the three sample variances. The 25th and 75th percentiles of the 1/4, 1/2, 1/4 mixture of these three densities were 23.553 and 38.889, respectively. The number of F2 data values less than 23.553, between 23.553 and 38.889 and greater than 38.889 were 119, 260 and 53, respectively, which leads to a chi-square statistic of 37.5671 and corresponding P-value of $(6.96)10^{-9}$. Another means of determining cutoff points is to simply use the 25th and 75th percentiles of the mixture of the three sets of data, without resort to any parametric model. This method yields cutoffs of 24.29 and 38.72, corresponding chi-square statistic 45.61472, and P-value $(1.244)10^{-10}$, and hence the conclusion is the same with either method

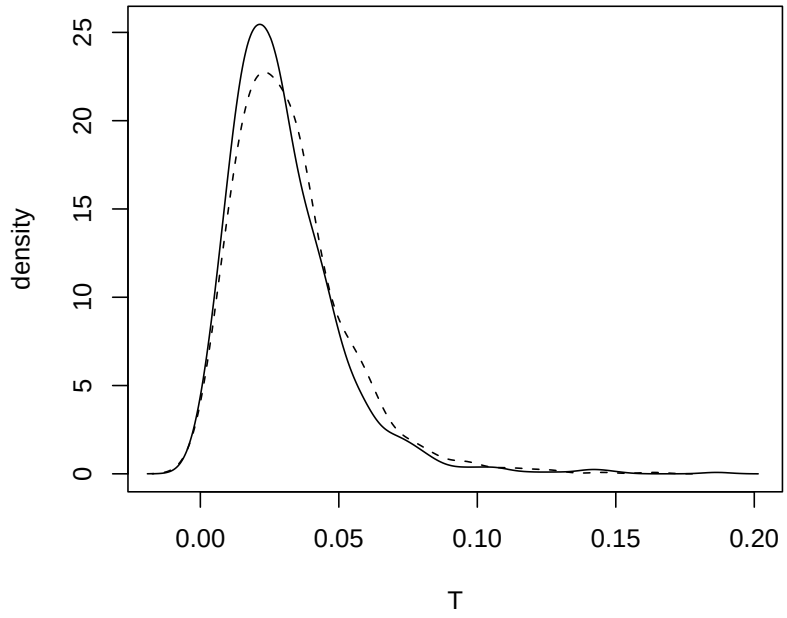

Figure 6: *Kernel density estimates for bootstrap test statistics. The solid and dashed lines correspond to the parametric and nonparametric bootstrap algorithms, respectively.*

of choosing cutoffs. The biggest discrepancy from what is expected is in the largest category, where the observed number of F2 responses is much less than predicted by the single gene theory. This can also be seen in Figure 5.

It is interesting to investigate how sensitive the χ^2 statistic is to choice of cutoffs. Suppose we use as cutoffs 24 and c, where the categories are $[0, 24]$, $(24, c]$ and (c, ∞). The resulting P-values as a function of c are shown in Figure 7 for the F2 data.

Note that if c were taken to be between 34 and 35.5, then the single gene theory would not be rejected at the .05 level of significance. It is not inconceivable that 24 and 35 would be chosen as cutoffs were the choice made subjectively, in which case a chi-square analysis would lead to an arguably incorrect conclusion for these data.

6 Concluding remarks

We have proposed the use of kernel density estimates in the specific context of testing whether a single gene is responsible for an inherited trait. Our method has none of the subjectivity inherent in a traditional χ^2 analysis, and is usually more reliable than such an analysis in terms of both validity and power. The methodology can be generalized to more complicated traits, such as those controlled by two or more genes. Even more generally, our testing methodology, or slight variations thereof, could be used in a number of different contexts where densities are to be compared.

Finally, we wish to dedicate this paper to Distinguished Professor Emanuel Parzen. His example as an intellectual and a scholar is inspiring, and his pioneering work in kernel density estimation, time series analysis, goodness-of-fit tests and quantile-based modeling has enlightened generations of scientists. Manny, congratulations on a fabulous career, and enjoy retirement!

REFERENCES

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71**, 353-360.

Coyne, D. P., Koroban, S. S., Knudsen, D. and Clark, R. B. (1982). Inheritance of iron deficiency in crosses of dry beans (*Phaseolus vulgaris L.*). *Journal of Plant Nutrition* **5**, 575-585.

De Haan, R. L. and Barnes, D. K. (1998). Inheritance of pod type, stem color, and dwarf growth habit in Medicago polymorpha. *Crop Science* **38**, 1558-1561.

Eubank, R. L., Hart, J. D. and LaRiccia, V. N. (1993). Testing goodness of fit via nonparametric function estimation techniques. *Communications in Statistics–Theory and Methods* **22**, 3327-3354.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* New York: Springer-Verlag.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics* **33**, 1065-1076.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics* **9**, 65-78.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman & Hall.

Young, S. G. and Bowman, A. (1995). Non-parametric analysis of covariance. *Biometrics* **51**, 920-931.

Zaiter, H. Z., Coyne, D. P. and Clark, R. B. (1988). Genetic variation, heritability, and selection response to iron deficiency chlorosis in dry beans. *Journal of Plant Nutrition* **11**, 739-764.

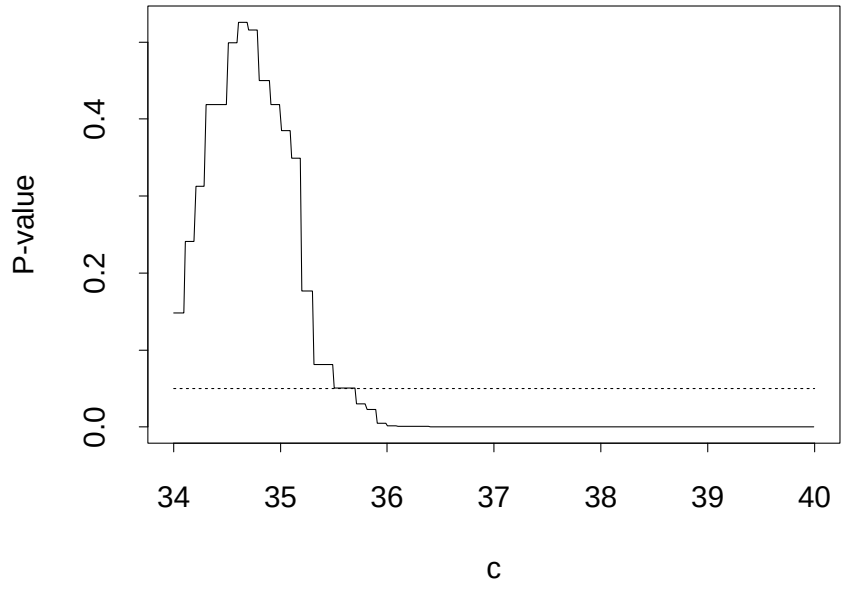

Figure 7: *P-value as a function of category cutoff. The χ^2 statistic for the F2 data was computed for cutoff points 24 and c, and the P-value computed for various values of c. The dashed line indicates 0.05.*