Department of Economics Finance & Accounting

Working Paper N274-16

Adjusted p-values for genome-wide regression analysis with non-normally distributed quantitative phenotypes

Gregory Connor
Department of Economics, Finance & Accounting
Maynooth University

September 2016

Adjusted p-values for genome-wide regression analysis with non-normally distributed quantitative phenotypes

Gregory Connor
Department of Economics, Finance and Accounting
Maynooth University

September 27, 2016

Abstract

This paper provides a small-sample adjustment for Bonferonni-corrected p-values in multiple univariate regressions of a quantitative phenotype (such as a social trait) on individual genome markers. The p-value estimator conventionally used in existing genome-wide association (GWA) regressions assumes a normally-distributed dependent variable, or relies on a central limit theorem based approximation. We show that the central limit theorem approximation is unreliable for GWA regression Bonferonni-corrected p-values except in very large samples. We note that measured phenotypes (particularly in the case of social traits) often have markedly non-normal distributions. We propose a mixed normal distribution to better fit observed phenotypic variables, and derive exact small-sample p-values for the standard GWA regression under this distributional assumption.

1 Introduction

This paper provides an alternative formula for the p-values in genome-wide association (GWA) univariate regressions of a phenotypic variable on single-nucleotide polymorphisms (SNPs). The formula is easy to apply, and can

provide substantially more accurate p-values if the phenotypic variable under consideration is non-normally distributed and the number of observations of the phenotypic variable is not very large (for example, less than ten thousand). For a normally distributed phenotypic variable, or with a very large sample, the adjustment is not necessary. The magnitude of the p-value adjustment depends upon the size of the sample, the non-normal features of the phenotypic variable including skewness and kurtosis, and the minor allele frequency of the SNP.

GWA regression is a key exploratory tool in genetic research on the heritability mechanisms of phenotypic traits, with the goal of identifying individual SNPs associated with the trait. GWA regressions involve a million or more individual univariate regressions (one per SNP), in the search for SNPs with significant univariate effects on an observed phenotypic variable. To account for the multiple comparisons problem, analysts use Bonferonni-corrected p-values, so that an adjusted 5% p-value with one million independent regressions requires an uncorrected univariate regression coefficient p-value (for a two-sided test) of $0.025x10^{-6}$.

In the estimation of Bonferonni-corrected p-values, analysts rely on the assumption that the estimated regression coefficient is normally distributed. This holds exactly if the phenotypic variable has a normal distribution, and approximately (for sufficiently large samples) if the phenotypic variable has any reasonably well-behaved distribution, by the central limit theorem. The quality of the central limit theorem based approximation depends upon the size of the sample, the distributional characteristics both of the observed phenotypic variable and the SNP, and (crucially in this application) on the magnitude of the p-value.

The central limit theorem guarantees uniform convergence of the true cumulative distribution to the normal distribution (see White (1984) for a review). An approximate p-value in the region of 0.025, accurate to within ± 0.0001, can be entirely adequate; an approximate p-value in the region of $0.025x10^{-6} = 0.000000025$ which is similarly accurate to within ± 0.0001 is effectively worthless. GWA regressions involve very large-number multiple tests and therefore extremely low p-value thresholds, with the conventional critical value set at $0.025x10^{-6}$. Invocation of the central limit theorem is problematic in this context.

In this paper, we develop an alternative approach based on fitting a mixed Bernoulli-normal distribution to the phenotypic variable. As we show, since a GWA regression has a trinomial explanatory variable (the three states of the

SNP) and the Bernoulli-normal mixture is a combination of a binomial and a normal, the resulting regression coefficient p-value is a multinomial-based linear combination of independent normals, with a closed-form expression in terms of the standard normal distribution.

We compare conventional and adjusted p-value for two common phenotypic variables: years of education (which has a notably non-normal distribution) and adult height (with a distribution that is close to normal). Empirically, the p-value adjustment can be quite large, and can increase or decrease estimated p-values relative to the conventional approach.

2 Exact small-sample p-values for GWA regression under a Bernoulli-normal mixture distribution

This section presents the new methodological result. We derive the exact p-values of the GWA regression under an assumed Bernoulli-normal mixture distribution. This is a reasonably straightforward exercise, combining the Bernoulli-normal mixture for the dependent variable with the three-valued explanatory variable of a GWA regression, and then rearranging, manipulating, and simplifying the expressions.

2.1 The GWA regression framework

The analyst has observations on $i = 1, n$ individuals with the data consisting of a phenotypic variable (such as income, years of education, life satisfaction rating, etc.) and a very long (we assume 10^6 for notational simplicity) string of genetic markers. The genetic markers are single nucleotide polymorphisms which have three potential states: major allele homozygot, minor allele homozygot, and heterozygot. Let SNP_{ij} be the trinomial explanatory variable, set equal to 0 if individual i is a major allele homozygot for the j^{th} genetic marker, 1 is he/she is a heterozygote, and 2 if he/she is a minor allele homozygot. To simplify notation we assume that the phenotype variable y is standardized and so has zero mean and unit variance.

The underlying theoretical model is that $SNPs$ have linear, additive impacts on the phenotypic variable, explaining the phenotype along with non-genetic (environmental) factors and genetic factors missing from the mea-

sured SNP string, all of which are included in the residual term:

$$y = a + \sum_{j=1}^{10^6} \beta_j SNP_j + \eta. \tag{1}$$

The formal null hypothesis is that there are no genetic effects: $\beta_j = 0$ for all j. The alternative hypothesis is that $\beta_j \neq 0$ for at least one j. We wish to test the null against the alternative, and also identify some j with $\beta_j \neq 0$.

The full linear model (1) is not directly estimable by multivariate ordinary least squares since the number of explanatory variables, $10^6 + 1$, greatly exceeds the number of phenotypic observations, n. However the assumed independence across the SNP markers allows us to decompose (1) into a set of 10^6 univariate regressions, each using one genetic marker:

$$y = \alpha_{j^*} + \beta_{j^*} SNP_{j^*} + \varepsilon_{j^*}, \tag{2}$$

where

$$\varepsilon_{j^*} = \sum_{\substack{j=1 \\ j \neq j^*}}^{10^6} \beta_j SNP_j + \eta.$$

The model (2) is estimated by ordinary least squares for each of the individual $SNPs$ and each $\widehat{\beta}_{j^*}$ is tested for significance. Let m_x and σ_x^2 denote the sample average and mean-square deviation of the explanatory variable in (2), SNP_{j^*}. The ordinary least squares regression coefficient from the GWA model (2) is:

$$\widehat{\beta}_{j^*} = \frac{1}{n\sigma_x^2} \sum_{i=1}^{n} y_i (SNP_{ij^*} - m_x). \tag{3}$$

Since (for convenience, without loss of generality) y is standardized, it follows from (3) that $Var[\widehat{\beta}_{j^*}] = \frac{1}{n}$.

With $n = 10^6$ independent tests, $H_0 : \beta_{j^*} = 0$ with $j^* = 1, n$ each tested separately, it is crucial to apply a Bonferonni correction to the individual test p-values. With 10^6 independent tests, and choosing a 95% confidence level, the two-tailed critical values for Bonferonni-corrected multiple-test significance of each coefficient uses a cumulative probability of 0.025×10^{-6} for testing a negative estimated coefficient and $1 - .025 \times 10^{-6}$ for a positive estimated coefficient.

2.2 Fitting a Bernoulli-Normal Mixture Distribution to a Phenotypic Variable

As we will demonstrate later, the central limit theorem does not always provide a reliable approximation for Bonferonni-corrected p-values with large-number multiple tests. We need an alternative estimator of GWA regression p-values in the case of a non-normally distributed phenotypic variable. We need a reasonable distributional assumption that, first, better fits the phenotypic variable and, second, allows for the feasible computation of small-sample p-values that do not rely on the central limit theorem approximation. In this section we propose a Bernoulli-normal mixture distribution.

The Bernoulli-normal mixture distribution is a flexible family of distributions with good fit in many applications, and convenient analytical properties in our model.

Let $z_1 \sim N(\mu_1, \sigma_1^2)$, $z_2 \sim N(\mu_2, \sigma_2^2)$, and λ a Bernoulli distributed random variable, $\lambda = 1$ with probability p; all three random variables assumed independent. The mixed Bernoulli-normal y is the random variable:

$$y = \lambda z_1 + (1 - \lambda) z_2,$$

which has five parameters: $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, p$. The first two moments are:

$$E[y] = p\mu_1 + (1-p)\mu_2 \tag{4}$$
$$Var[y] = p(\sigma_1^2 + \mu_1^2) + (1-p)(\sigma_2^2 + \mu_2)^2 - E[y^*]^2 \tag{5}$$

The distribution can be fitted via EM-maximum likelihood; see McLachlan and Peel (2000) for an overview of mixture distributions and estimation methods.

3 The GWA regression coefficient as a linear combination of independent normals

Returning to our GWA regression model (2) we now use the assumption that y has a Bernoulli-normal mixture distribution to derive the exact small-sample p-values of $\widehat{\beta}_{j^*}$. Since we now look at one particular j^* only, we simplify notation and drop the j^* subscript. To implement our technique, the analyst counts the number of major allele observations, heterozygot observations and minor allele observations in each regression sample. Let $\{n_0, n_1, n_2\}$

denote these three integer values, with $n_0+n_1+n_2=n$. The sample average and mean-square deviation of the explanatory variable have simple forms, since SNP_i only takes the three values $0, 1, 2$:

$$m_x = \frac{1}{n}\sum_{i=1}^{n} SNP_i = \frac{n_1+2n_2}{n}$$
$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(SNP_i - m_x)^2 = \frac{n_0 m_x^2 + n_1(1-m_x)^2 + n_2(2-m_x)^2}{n}.$$

The cumulative distribution of an estimate $\widehat{\beta}$ at a value $\overline{\beta}$ under the null hypothesis is:

$$\Pr[\widehat{\beta} \leq \overline{\beta}] = \Pr[\frac{1}{n\sigma_x^2}\sum_{i=1}^{n} y_i(SNP_i - m_x) \leq \overline{\beta}]. \tag{6}$$

For notational convenience, we re-order the observations index so that the major allele observations are listed first, then the heterozygot observations, and then the minor allele observations. Let i^* denote the re-ordered index:

$$SNP_{i^*} = 0;\ i^* = 1, n_0$$
$$SNP_{i^*} = 1;\ i^* = n_0+1, n_0+n_1$$
$$SNP_{i^*} = 2;\ i^* = n_0+n_1+1, n$$

Writing out the coefficient formula (6) using the observed values n_0, n_1, n_2:

$$\Pr[\widehat{\beta} \leq \overline{\beta}] = \Pr[\frac{1}{\sigma_x^2}(\sum_{i^*=1}^{n_1}(-m_x)y_{i^*} + \sum_{i^*=n_1+1}^{n_1+n_2}(1-m_x)y_{i^*} + \sum_{i^*=n_1+n_2+1}^{n}(2-m_x)y_{i^*}) \leq \overline{\beta}].$$

Under our distributional assumption on y, each of the three integers n_0, n_1 and n_2, in turn decomposes into two (unobserved) integers: the number of realizations of the dependent variable y_{i^*} for which the Bernoulli random variable λ equals zero or one. For notational convenience also define the remainders $n_{0,2} = n_0 - n_{0,1}$, $n_{1,2} = n_1 - n_{1,1}$ and $n_{2,2} = n_2 - n_{1,1}$. Let $\{n_{0,1}, n_{1,1}, n_{2,1}\}_h, h = 1, m$ denote the finite set of all integer combinations with $0 \leq n_{0,1} \leq n_0$, $0 \leq n_{1,1} \leq n_1$, $0 \leq n_{2,1} \leq n_2$. It is easy to show that $m = (n_0+1)(n_1+1)(n_2+1)$.

Each of the integers $n_{0,1}, n_{1,1}, n_{2,1}$ has an independent binomial distribution. The probabilities of all the potential outcomes $\{n_{0,1}, n_{1,1}, n_{2,1}\}_h, h =$

$1, m$ can be found from the binomial formula (for three independent binomials):

$$P_h = \Pr[\{n_{0,1}, n_{1,1}, n_{2,1}\}_h] = \tag{7}$$
$$(\frac{n_0!}{n_{0,1}!n_{0,2}!})(\frac{n_1!}{n_{1,1}!n_{1,2}!})(\frac{n_2!}{n_{2,1}!n_{2,2}!})\times$$
$$p^{(n_{0,1}+n_{1,1}+n_{2,1})}(1-p)^{(n_{0,2}+n_{1,2}+n_{2,2})}.$$

The cumulative probability of $\widehat{\beta}$ is the conditional cumulative probability for each of the potential outcomes $h = 1, m$ times the probability of each outcome:

$$\Pr[\widehat{\beta} \leq \overline{\beta}] = \sum_{h=1}^{m} (\Pr[\widehat{\beta} \leq \overline{\beta}|h]) \times P_h. \tag{8}$$

The last step is to calculate

$$\Pr[\widehat{\beta} \leq \overline{\beta}|h] = \Pr[\frac{1}{n\sigma_x^2} \sum_{i^*=1}^{n} y_{i^*} SNP_{i^*} \leq \overline{\beta}|h];$$

this is the sum of n independent normals, consisting of $n_{0,1}$ draws of $(-m_x/n\sigma_x^2)z_1$, plus $n_{0,2}$ draws of $(-m_x/n\sigma_x^2)z_2$, plus $n_{1,1}$ draws of $((1-m_x)/n\sigma_x^2)z_1$, plus $n_{1,2}$ draws of $((1-m_x)/n\sigma_x^2)dz_2$, plus $n_{2,1}$ draws of $((2-m_x)/n\sigma_x^2)z_1$, plus $n_{2,2}$ draws of $((2-m_x)/n\sigma_x^2)dz_2$. A linear combination of independent normals has a normal distribution, and in particular:

$$\Pr[\frac{1}{n\sigma_x^2} \sum_{i^*=1}^{n} y_{i^*} SNP_{i^*} \leq \overline{\beta}|h] = \Pr[z_0 \leq \overline{\beta} -$$
$$\frac{1}{n\sigma_x^2}(-m_x n_{0,1} + (1-m_x)n_{1,1} + (2-m_x)n_{2,1})\mu_1 -$$
$$(-m_x n_{0,2} + (1-m_x)n_{1,2} + (2-m_x)n_{2,2})\mu_2)\times$$
$$(\frac{1}{n\sigma_x^2}((m_x^2 n_{0,1} + (1-m_x)^2 n_{1,1} + (2-m_x)^2 n_{2,1})\sigma_1^2 +$$
$$(m_x^2 n_{0,1} + (1-m_x)^2 n_{1,1} + (2-m_x)^2 n_{2,1})\sigma_2^2)^{-\frac{1}{2}} \tag{9}$$

where z_0 denotes a standard normal random variable. Combining (7), (8) and (9) gives a computable formula for the exact small-sample p-value of the GWA regression coefficient for any sample size.

3.1 Computationally efficient implementation of the estimator

The p-value formula (8) requires a sum over the set of outcomes from three independent binomials with n_0, n_1 and n_2 draws, giving a total of $(n_0+1)(n_1+1)(n_2+1)$ terms. Even with $n = 10,000$ this is not computationally difficult since the vast majority of the random outcomes can be dropped from the sum, without any discernible effect on the quality of the estimate. Suppose for example that the regression uses an SNP of 10000 total observations with $n_0 = 8100, n_1 = 1800$, $n_2 = 100$, and that $p = 0.7$. The complete sum (8) has a total of $1,473,580,001$ terms in this case. However, using the binomial distribution, the cumulative probability that the number of major allele homozygots with $\lambda = 1$, $n_{0,1}$, is less than 5376 or more than 5956 is less than 10^{-12}. All of these very low cumulative probability random outcomes can be dropped before calculating (7); along with the outcomes where the number of heterozygots with $\lambda = 1$, n_{11}, is less than 1118 or more than 1393; and those where n_{21} is less than 35 or more than 97. This leaves a manageable $10,102,428$ terms in the (trimmed) sum, without noticeably impacting the accuracy of the estimate.

4 Illustration of the magnitude of the p-value adjustment using two common phenotypic variables

This section examines the magnitude of the adjustment arising from our small sample p-values compared to using large-sample approximate p-values based on the central limit theorem. We illustrate the adjustment with two commonly used phenotypic variables: years of education, which is a social trait with a strongly non-normal distribution, and adult height, which is a physical trait with a near-normal distribution.

Given the parameters of the mixture distribution, our p-value formula (8) is exact; it does not require any simulation. The only inputs needed are the number of major allele homozygot, heterozygot, and minor allele homozygot observations in the regression sample, (n_0, n_1, n_2), the estimated regression coefficient, $\widehat{\beta}$, and the five parameters of the mixture distribution, $(p, \mu_1, \mu_2, \sigma_1, \sigma_2)$.

For the purposes of this comparison, we use five sample sizes, $n = 100$, 500, 1000, 5000, 10000. For each sample size we fit n_0, n_1, n_2 from the range of values typically encountered in GWA regression tests. Let MAF denote the minor allele frequency of the SNP; we chose four representative values: $MAF = 0.5\%, 1\%, 5\%, 10\%$. To choose the observation numbers n_0, n_1, n_2 we assume that the SNP is in Hardy-Weinberg equilibrium, which implies that $n_0 = n(1 - MAF)^2$, $n_1 = 2nMAF(1 - MAF)$ and $n_2 = nMAF^2$. The numbers of observations n_0, n_1, n_2 must be integers, so for fractional values we stick the "leftover" one or two observation(s) in the heterozygot category.

Note that the relative numbers of explanatory variables across the three categories, n_0, n_1, n_2, can affect the quality of the central limit theorem approximation. For example, with $MAF = 1\%$, only 0.01% of SNP observations take the value +1; 1.98% take the value 0, and for 98.01% of the regression sample, $SNP = -1$. This unbalanced distribution impacts the speed at which the central limit theorem acts upon the probability distribution of the coefficient estimate, and the asymmetry (right-tail probability versus left-tail probability) of its finite-sample distribution, unless the dependent variable is exactly normal. This will become clear in the tables below.

For comparative purposes, we assume $\widehat{\beta}$ values which have cumulative probability $2.5\%, 0.5\%$, and $2.5\% \times 10^{-6}$ under normality. These are:

$$\begin{aligned} Pr(\frac{\widehat{\beta}}{\sqrt{n}} \leq -1.96) &= 2.5\% \\ Pr(\frac{\widehat{\beta}}{\sqrt{n}} \leq -2.58) &= 0.5\% \\ Pr(\frac{\widehat{\beta}}{\sqrt{n}} \leq -5.45) &= 2.5\% \times 10^{-6}; \end{aligned}$$

the upper-tail tests are analogous,with a positive sign. Multiplying through by $\sqrt{n}$ gives:

$$\begin{aligned} \overline{\beta}_{.025} &= -1.96 \times \sqrt{n} \\ \overline{\beta}_{.025} &= -2.58 \times \sqrt{n} \\ \overline{\beta}_{10^{-6} \times .025} &= -5.45 \times \sqrt{n}, \end{aligned} \tag{10}$$

and the positive tail values are analogous. In the tables below we take each of the three $\overline{\beta}$ values in (10) and find the small-sample p-value under the

mixture distribution, which we can then compare to the normality-based p-values, $2.5\%, 0.5\%$, and $2.5\% \times 10^{-6}$.

To calibrate the parameters of the mixture distribution, $(p, \mu_1, \mu_2, \sigma_1, \sigma_2)$, we run EM-maximum likelihood on the phenotypic variable; see below for details.

4.1 Application to a non-normal phenotype: Years-of-education

In this subsection we calibrate the Bernoulli-normal mixture using data on years of education from the U.S. Census Bureau Current Population Survey of Educational Attainment, 2015. See Rietveld, et al. (2013, 2015), Okbay et al. (2016), and references therein for details on the considerable number of GWA regression studies with years-of-education as the phenotypic variable.

Figure 1 shows a frequency distribution of the years-of-education data, along with fitted normal and Bernoulli-normal mixture distributions. See the Appendix for description of the U.S. census data. The mixture distribution picks up the high-peakedness and asymmetry in the data distribution, associated with the 76% frequency of data observations in the range $12-16$ years, and the secondary clump of observations in the $0-6$ years range with frequency 3.04%. These features are missed by the fitted normal. The data has skewness of -0.676781 and excess kurtosis of 2.126954, which both differ significantly from zero with 99% confidence. The Jarque-Bera test rejects normality with 99% confidence.

The 81,913 years-of-education data observations are fitted to a Bernoulli-normal mixture distribution using the *normalmixEM* command in the *mixtools* library of programming language *R;* see Benaglia et al. (2009) for details on the estimation routine. The estimated parameter values are $p = 0.9654$, $\mu_1 = 13.872$, $\mu_2 = 4.628$, $\sigma_1 = 2.588$, $\sigma_2 = 2.518$.

Table 1 Panel A considers a single-hypothesis, two-sided test with a 95% confidence limit. The table shows exact small-sample p-values under the mixture distributions for estimated regression coefficients with approximate p-values (using the central limit theorem) of 2.5%. The central limit theorem approximation gives quite accurate p-values in almost all cases, even with small sample sizes and low minor allele frequencies. The approximation error from invoking the central limit theorem to compute p-values is never severe.

Panel B of the table repeats the exercise for a 99% confidence test, so that

the p-value under normality is 0.5%. The central limit theorem approximation continues to work reasonably well, with the exception of small sample sizes (500 or less observations) with minor allele frequencies of 0.5% or 1%.

For 10^6 multiple test Bonferonni-corrected p-values, shown in Table 2, the approximation error from relying on the central limit theorem is severe. Convergence of the p-value toward its normality-derived value is much slower, and the small-sample asymmetry in the approximation error is more notable. For small to medium sample sizes, the true p-value for a negative-tail test is very substantially above 2.5%, the p-value for the positive-tail test is substantially below 2.5%, and the sum of the two tail probabilities (which should be 5%) is substantially higher. The central limit theorem approximation only works reasonably accurately for sample sizes of five or ten thousand, and only with relatively high minor allele frequency. In the other cases considered in Table 2, the small-sample adjustment is critically important.

Figures 2-4 illustrate the dependence of the central limit theorem approximation to p-values on the size of the regression sample and the minor allele frequency of the SNP. Figure 2 assumes a minor allele frequency of 1% and shows the true p-values, under the mixture distribution calibrated to the years-of-education variable, compared to the approximate p-values from the central limit theorem, for three regression sample sizes, 500, 1000, and 10000. As in the tables above, the p-value equals the cumulative distribution for $\widehat{\beta} < 0$ and one minus the cumulative distribution for $\widehat{\beta} \geq 0$. Figure 3 repeats this exercise for a minor allele frequency of 10%. Figure 4 holds the regression sample size fixed at 1000 and examines the quality of the central limit theorem approximation as a function of the minor allele frequency.

4.2 Application to a near-normal phenotype: Adult height

In this subsection we examine a phenotypic variable that is closer to normality than the years of education variable. Recall that the p-value adjustment in this paper is only relevant in the case of a non-normally distributed phenotype. If the phenotypic variable is exactly normal (which corresponds to $p = 0$ or $p = 1$, or $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_1$ in the mixture distribution) then the adjusted p-value exactly equals the conventional normality-based p-value. If the phenotypic variable is "close" to normal (as measured for example by its skewness and excess kurtosis) then the convergence of the estimated coeffi-

cient to normality is likely to be relatively rapid, and the p-value adjustment is likely to have limited value except for very small sample sizes.

Figure 5 shows the sample frequency distribution of height for U.S. white males age 25 and over, using 129,735 observations; see the Appendix for a description of the data source. A normal and fitted mixture distribution are overlaid on the sample frequency distribution in the Figure. The height sample has a mean of 70.35, standard deviation 2.83, skewness of -0.086 , and excess kurtosis of 0.721. We estimate a mixture distribution using mixtools as above, getting parameter estimates $p = 0.0386$, $\mu_1 = 69.52$, $\mu_2 = 70.38$, $\sigma_1 = 4.95$, $\sigma_2 = 2.71$. Visually the sample distribution looks quite close to normality.

Table 3 and 4 repeat the exercise done in Tables 1 and 2, using the mixture distribution for height in place of years of education. For p-values of 2.5% and 0.5% the p-value adjustment is negligible throughout, even for the smallest sample size and lowest minor allele frequency. In the case of $2.5\% \times 10^{-6}$ the p-value adjustment is not large, except for the smallest sample size and minor allele frequencies of 1% or less. Sample sizes of 5000 or more have reasonably accurate conventional p-values even with the low minor allele frequencies. With a near-normal phenotypic variable like height, the p-value adjustment adds limited empirical value.

5 Summary

This paper provides a new approach to estimating multiple-test Bonferonni-corrected p-values in genome-wide association (GWA) regressions of individual genetic markers on a phenotypic variable such as a social trait. The current standard approach in computing coefficient p-values is to assume a normal distribution for the phenotypic variable, or to invoke the central limit theorem to justify the approximate normality of the coefficient estimate. Many phenotypic variables, particularly for social traits like income and education levels, have distributions which are far from normality. The central limit theorem, as we show, does not give reliable p-values for the types of sample sizes (and multiple-test numbers) used in GWA regression studies with non-normally distributed phenotypic variables.

We suggest a new approach, based on fitting a Bernoulli-normal mixture distribution to the phenotypic variable before running the GWA regressions. We derive the exact small-sample distribution of GWA regression coefficient

p-values under this more flexible distributional assumption. We illustrate the magnitude of the p-value adjustment from our approach (relative to the conventional approach) with sample data on a commonly-used phenotypic variable with a notably non-normal distribution: years of education. The derived p-values differ markedly from the conventional, normality-based, p-values. For comparison purposes we also consider a near-normal phenotypic variable, adult height, for which the difference between the adjusted and conventional p-values is much more limited.

Acknowledgments

I would like to thank Philipp Koellinger, Richard Linner, Brian O'Kelly and Donal O'Neill for helpful comments. I wish to acknowledge support from the Science Foundation of Ireland.

Appendix: Data Description

The first two columns of Table A-1 below reproduce two rows from *Table 1: Educational Attainment of the Population 18 Years and Over, by Age, Sex, Race, and Hispanic Origin: 2015* in Current Population Survey Data on Educational Attainment (U.S. Census Bureau (2015)). We choose the subsample "U.S. white males ages 25 and greater" from that data source, which is row 25 of their *Table 1.* The [white/male/age 25 and over] subsample has 81,913 observations. Row three of Table A-1 below transforms the qualitative categories into a quantitative variable. There are a few minor subjective judgements in transforming the survey categories into quantitative years-of-education. The final column shows the frequency distribution of the data.

The adult height variable comes from the publicly available database in *Behavioral Risk Factor Surveillance System,* U.S. Centers for Disease Control and Prevention (2010). The selected sample consists of white, male, 25 years or older respondents excluding those with a missing height data item, giving a total of 129,735 observations. The data item is recorded in whole inches, there are no fractional values.

References

[1] Benaglia T., D. Chauveau, DR. Hunter, D. Young (2009). "Mixtools: An R Package for Analyzing Finite Mixture Models," *Journal of Statistical Software*, 32(6):1-29.

[2] McLachlan, G., and D. Peel (2000). *Finite Mixture Models*, Wiley Series in Probability and Statistics, New York.

[3] Okbay, A., JP. Beauchamp, MA. Fontana, JJ. Lee, TH. Pers, CA. Rietveld, P. Turley, G. Chen, V. Emilsson, SFW. Meddens, S. Oskarsson, JK. Pickrell, K. Thom, P. Timshel, R. de Vlaming, A. Abdellaoui, TS. Ahluwalia, J. Bacelis, C. Baumbach, G. Bjornsdottir, JH. Brandsma, MP. Concas, J. Derringer, NA. Furlotte, TE. Galesloot, et al. (2016). "Genome-wide association study identifies 74 loci associated with educational attainment," *Nature,* 533: 539–542.

[4] Rietveld CA, , SE. Medland, J. Derringer, J. Yang, T. Esko, NW. Martin, HJ. Westra, et al. (2013). "GWAS of 126,559 individuals identifies genetic variants associated with educational attainment," *Science,* 340(6139): 1467–1471.

[5] Rietveld, CA., T. Esko, G. Davies, TH. Pers, P. Turley, B. Benyamin, CF. Chabris, V. Emilsson, AD. Johnson, J J. Lee, C. de Leeuw, RE. Marioni, SE. Medland, MB. Miller, O. Rostapshova, SJ. van der Lee, AAE. Vinkhuyzen, N. Amin, D. Conley, J. Derringer, CM. van Duijn, R. Fehrmann, L. Franke, EL. Glaeser, NK. Hansell, C. Hayward, WG. Iacono, C. Ibrahim-Verbaas, V. Jaddoe, J. Karjalainen, D. Laibson, P. Lichtenstein, DC. Liewald, PKE. Magnusson, NG. Martin, M. McGue, G. McMahon, NL. Pedersen, S. Pinker, DJ. Porteous, D. Posthuma, F. Rivadeneira, BH. Smith, JM. Starr, H. Tiemeier, NJ. Timpson, M. Trzaskowski, AG. Uitterlinden, FC. Verhulst, ME. Ward, MJ. Wright, GD. Smith, IJ. Deary, M. Johannesson, R. Plomin, PM. Visscher, DJ. Benjamin, D. Cesarini, and PD. Koellinger (2015). "Common genetic variants associated with cognitive performance identified using the proxy-phenotype method," *Publications of the National Association of Science,*

Psychological and Cognitive Sciences, Genetics, January, 12(15): 13790-13794.

[6] U.S. Census Bureau (2015). *Current Population Survey Data on Educational Attainment: 2015,* available at http:// www.census.gov/ hhes/ socdemo/ education/.

[7] U.S. Center for Disease Control and Prevention (2010). *Behavioral Risk Factor Surveillance System,* available at http://www.cdc.gov/brfss/ annual_data/annual_2010.htm.

[8] White, Halbert (1984). *Asymptotic theory for econometricians*, Academic Press, Inc., London.

Table 1

Comparison of adjusted/unadjusted single-test p-values for GWA regression coefficients under a mixture distribution: Years of education

Panel A: 95% two-tailed confidence test (conventional p-value=2.5%)

	Sample Size:	**100**	**500**	**1000**	**5000**	**10000**
Minor allele frequency	Sign of tested coefficient	p-value				
0.5%	Negative	3.68%	3.47%	3.19%	2.81%	2.72%
	Positive	1.31%	1.66%	1.84%	2.19%	2.28%
1%	Negative	4.00%	3.18%	2.98%	2.72%	2.66%
	Positive	1.45%	1.85%	2.02%	2.28%	2.34%
5%	Negative	3.10%	2.79%	2.71%	2.59%	2.57%
	Positive	1.94%	2.21%	2.30%	2.41%	2.43%
10%	Negative	2.93%	2.69%	2.64%	2.56%	2.54%
	Positive	2.11%	2.31%	2.37%	2.44%	2.46%

Panel B: 99% two-tailed confidence test (conventional p-value=0.5%)

	Sample Size:	**100**	**500**	**1000**	**5000**	**10000**
Minor allele frequency	Sign of tested coefficient	p-value				
0.5%	Negative	2.29%	1.12%	0.92%	0.67%	0.62%
	Positive	0.16%	0.22%	0.26%	0.37%	0.40%
1%	Negative	1.68%	0.91%	0.78%	0.62%	0.58%
	Positive	0.18%	0.26%	0.31%	0.40%	0.43%
5%	Negative	0.86%	0.66%	0.61%	0.55%	0.53%
	Positive	0.29%	0.38%	0.41%	0.46%	0.47%
10%	Negative	0.76%	0.61%	0.57%	0.53%	0.52%
	Positive	0.35%	0.42%	0.44%	0.47%	0.48%

Table 2

Comparison of adjusted/unadjusted 10^6 multiple-test p-values for GWA regression coefficients under a mixture distribution: Years of education

95% two-tailed confidence test (conventional p-valuex10^6 = 2.5%)

	Sample Size:	**100**	**500**	**1000**	**5000**	**10000**
Minor allele frequency	Sign of tested coefficient	p-valuex10^6				
0.5%	Negative	>100%	>100%	>100%	40.48%	19.56%
	Positive	0.01%	0.02%	0.04%	0.19%	0.35%
1%	Negative	>100%	>100%	>100%	19.08%	11.21%
	Positive	0.01%	0.04%	0.08%	0.36%	0.59%
5%	Negative	>100%	37.44%	18.21%	6.29%	4.81%
	Positive	0.08%	0.25%	0.43%	1.06%	1.35%
10%	Negative	>100%	17.66%	10.05%	4.61%	3.84%
	Positive	0.23%	0.56%	0.80%	1.45%	1.70%

Table 3
Comparison of adjusted/unadjusted single-test p-values for GWA regression coefficients under a mixture distribution: Adult height

Panel A: 95% two-tailed confidence test (conventional p-value=2.5%)

	Sample Size:	**100**	**500**	**1000**	**5000**	**10000**
Minor allele frequency	Sign of tested coefficient	p-value				
0.5%	Negative	2.54%	2.60%	2.57%	2.53%	2.52%
	Positive	2.38%	2.45%	2.46%	2.48%	2.48%
1%	Negative	2.60%	2.57%	2.55%	2.52%	2.52%
	Positive	2.43%	2.46%	2.47%	2.48%	2.49%
5%	Negative	2.56%	2.53%	2.52%	2.51%	2.51%
	Positive	2.46%	2.48%	2.48%	2.49%	2.49%
10%	Negative	2.54%	2.52%	2.51%	2.51%	2.50%
	Positive	2.47%	2.49%	2.49%	2.50%	2.50%

Panel B: 99% two-tailed confidence test (conventional p-value=0.5%)

	Sample Size:	**100**	**500**	**1000**	**5000**	**10000**
Minor allele frequency	Sign of tested coefficient	p-value				
0.5%	Negative	0.69%	0.60%	0.56%	0.52%	0.51%
	Positive	0.55%	0.51%	0.50%	0.49%	0.49%
1%	Negative	0.66%	0.56%	0.54%	0.51%	0.51%
	Positive	0.53%	0.50%	0.49%	0.49%	0.50%
5%	Negative	0.55%	0.52%	0.51%	0.51%	0.50%
	Positive	0.50%	0.49%	0.49%	0.50%	0.50%
10%	Negative	0.54%	0.51%	0.51%	0.50%	0.50%
	Positive	0.50%	0.50%	0.50%	0.50%	0.50%

Table 4
Comparison of adjusted/unadjusted 10^6 multiple-test p-values for GWA regression coefficients under a mixture distribution: Adult height

95% two-tailed confidence test (conventional p-value$\times 10^6$ = 2.5%)

	Sample Size:	**100**	**500**	**1000**	**5000**	**10000**
Minor allele frequency	Sign of tested coefficient	p-value$\times 10^6$				
0.5%	Negative	>100%	>100%	27.05%	4.84%	3.67%
	Positive	>100%	27.06%	7.87%	2.70%	2.46%
1%	Negative	>100%	26.24%	9.87%	3.65%	3.16%
	Positive	>100%	7.70%	3.87%	2.46%	2.40%
5%	Negative	20.44%	5.10%	3.70%	2.82%	2.70%
	Positive	6.53%	2.84%	2.51%	2.41%	2.42%
10%	Negative	16.59%	3.80%	3.16%	2.69%	2.62%
	Positive	6.16%	2.59%	2.46%	2.44%	2.46%

Figure 1: The frequency distribution of years-of-education (red bars) and fitted normal (blue line) and mixture (green line) probability densities

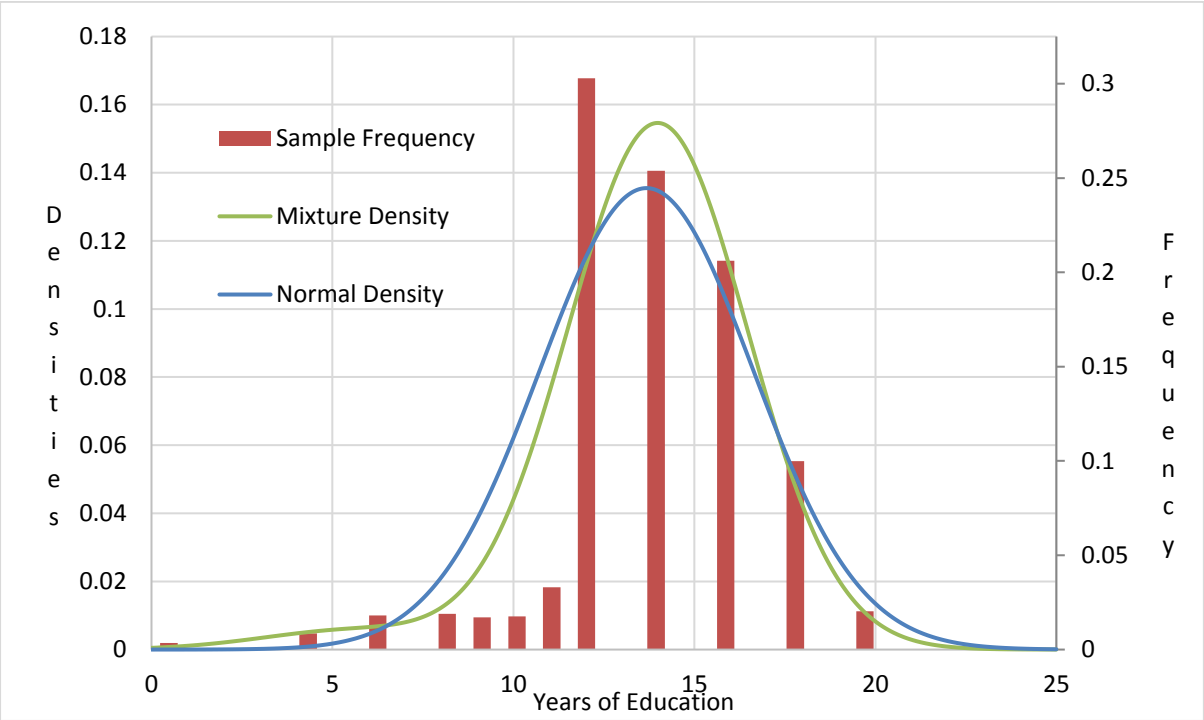

Figure 2: The central limit theorem p-value approximation (solid line) for sample sizes of n=500 (small dashes), n=1000 (medium dashes) and n=10000 (large dashes) using the years-of-education mixture distribution; 1% minor allele frequency

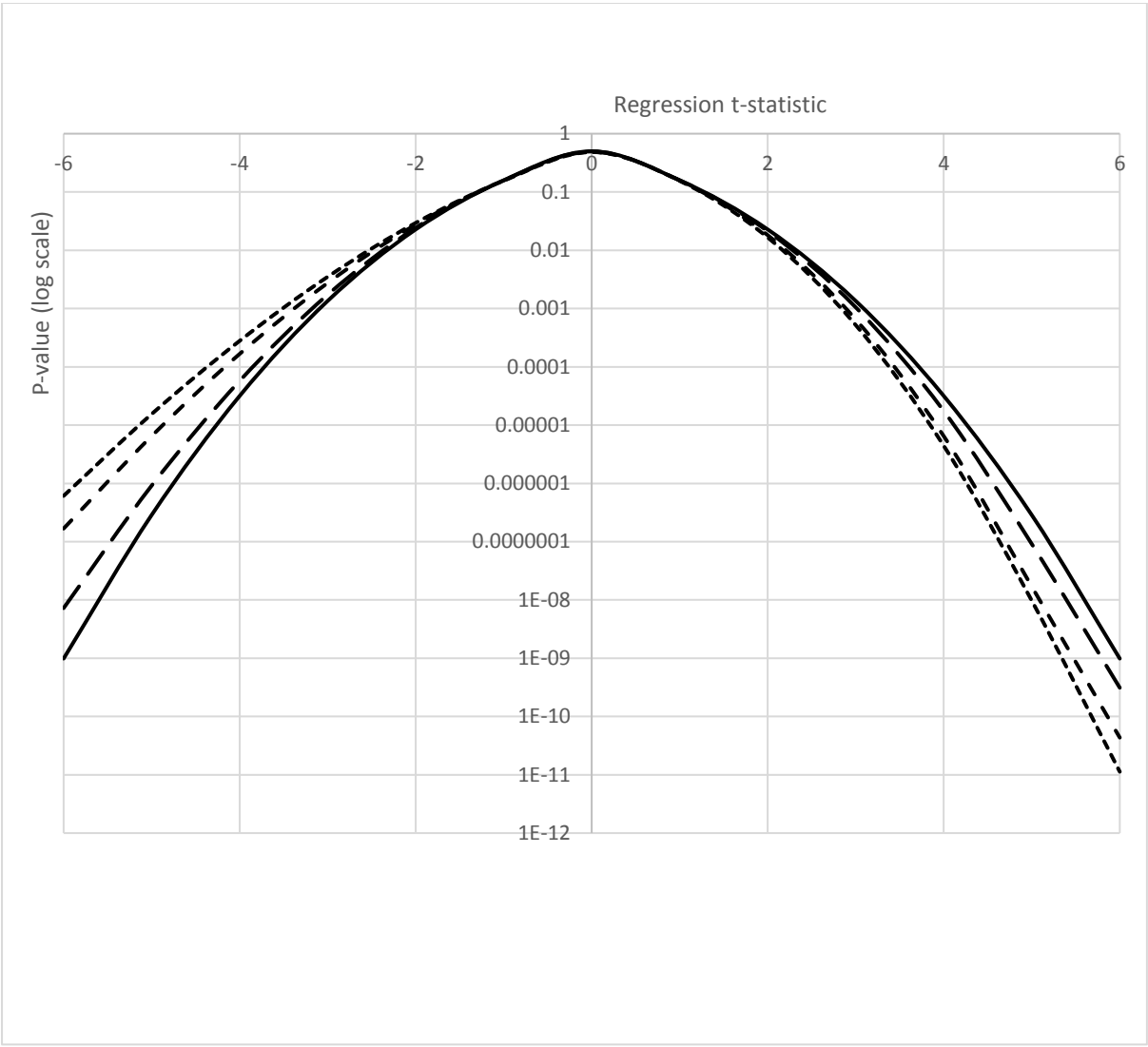

Figure 3: The central limit theorem p-value approximation (solid line) for sample sizes of n=500 (small dashes), n=1000 (medium dashes) and n=10000 (large dashes) using the years-of-education mixture distribution; 10% minor allele frequency

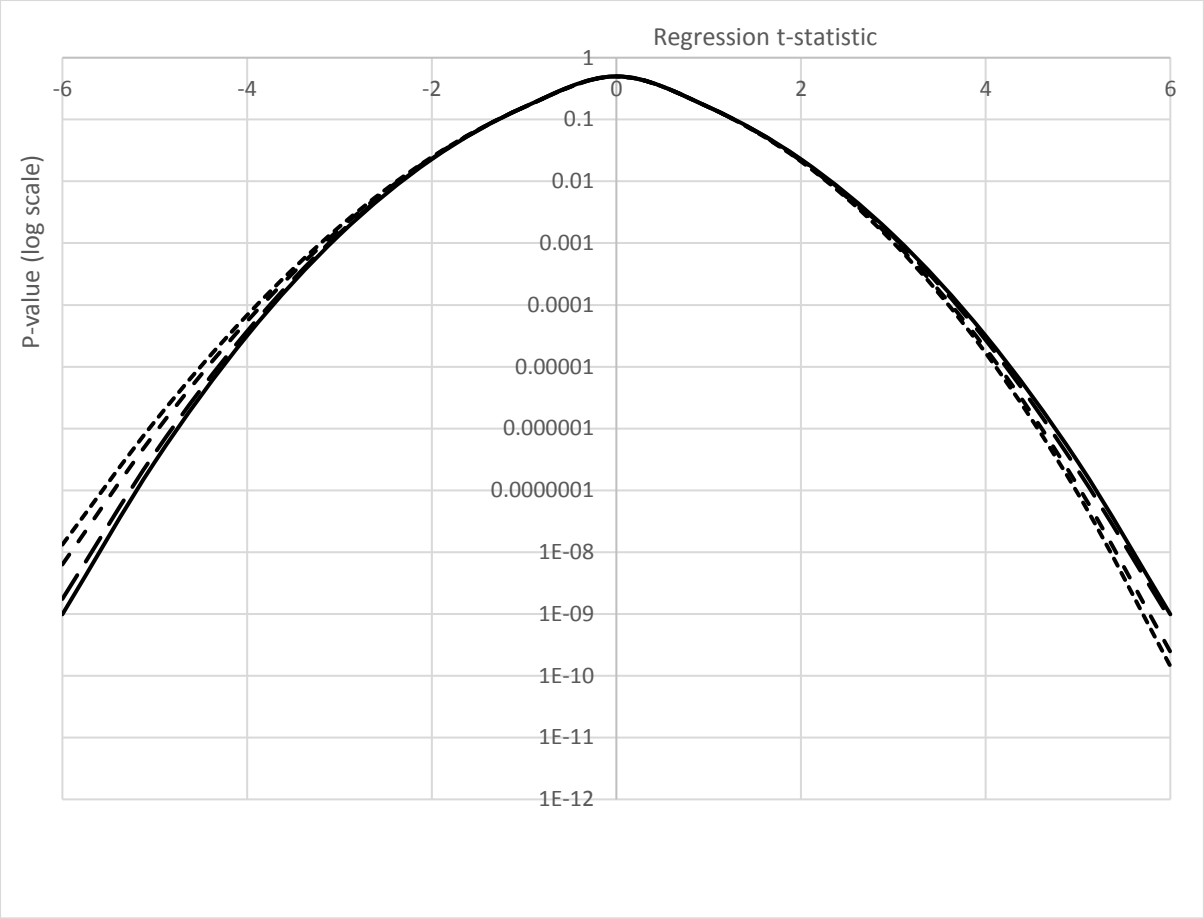

Figure 4: The central limit theorem p-value approximation (solid line) for minor allele frequency of 0.5% (dotted line), 1% (small dashes), 5% (medium dashes) and 10% (large dashes) using the years-of-education mixture distribution; 1000 sample size

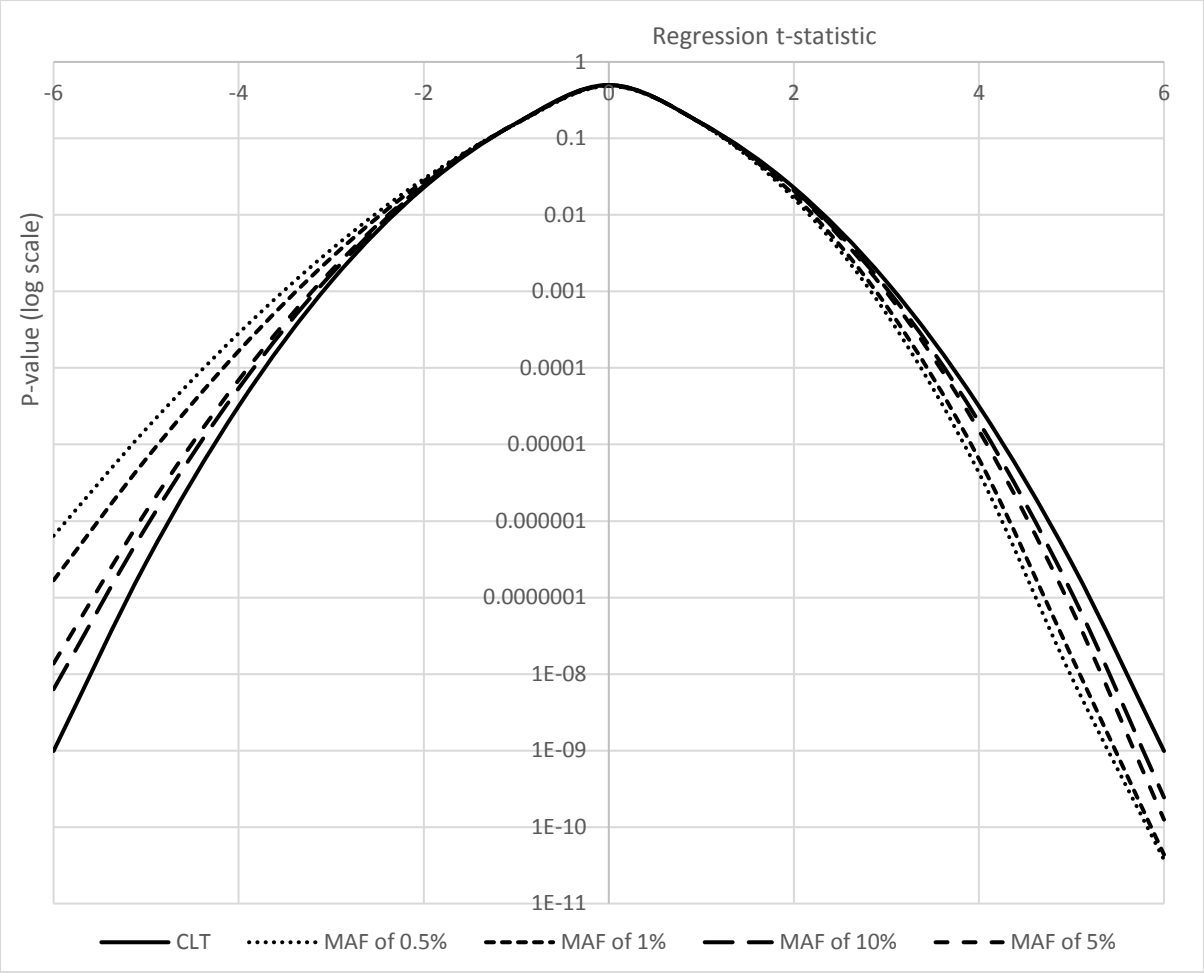

Figure 5: The frequency distribution of adult height (red bars) and fitted normal (blue line) and mixture (green line) probability densities