The Significance of Statistical Significance

Hal M. Switkay

Goldey-Beacom College, 4701 Limestone Road, Wilmington, DE 19808

Abstract

Hypothesis testing and decision rules are in the news as never before. The reproducibility of experiments, one of the touchstones of the scientific method, is uncertain, while some warn that most scientific results are wrong; see Ioannidis, as well as van der Laan.

At the heart of the controversy is the significance of statistical significance: specifically, the significance of p-values. Poorly crafted decision rules have led to a loss of confidence in p-values, with some proposing to ban this incredibly useful tool altogether. We reject this over-reaction.

We will discuss three aspects of p-values: 1) improving model specification, thereby reducing the probability of a type II error (false negative), by introducing new families of transformations to reduce skewness and excess kurtosis; 2) setting significance level as a decreasing function of sample size, thereby reducing the probability of a type I error (false positive), thus compromising between a fixed significance level and a fixed meaningful effect size; 3) continuous decision rules that assign plausibility levels to the null hypothesis and alternative hypothesis.

Keywords: transformation, p-value, significance

1. Introduction

This article represents a response to the ASA's recent Statement on p-Values, Wasserstein et al (2016). The statement elicited many replies at the ASA's subsequent Symposium on Statistical Inference in October 2017, in Bethesda, Maryland. Those responses often expressed wishful thinking that replacing the arbitrary constant 0.05 with the equally arbitrary constant 0.005 would solve the problems with statistical practice that led to the original statement. Other respondents proposed arbitrary cutoffs in the interpretation of likelihood ratios or Bayes factors. In contrast, we take an approach that attempts to fill in gaps in the classical approach to statistical inference.

Given a dataset, an investigator wants to explore its basic descriptive properties. Beyond that, an investigator often wants to know whether there is anything particularly noteworthy about this dataset. One of the most common examples of noteworthiness would be a discovery that there is an unexpected association among two or more variables in the dataset. This includes the conclusion that two or more populations have different means or proportions.

Noteworthiness is a scarce resource, and we wish to spend it wisely. Our first goal is to uncover hidden regularity within individual random variables and between random variables, using the power of elementary transformations to discover approximately linear relations among approximately normal variables.

Once we have maximized the possibility of finding noteworthy relations among variables in a classical model, our second goal is to ensure that the newfound relations meet a criterion of noteworthiness that becomes more stringent as sample size increases. Thus, we make it more difficult to engage in p-hacking.

Finally, our third goal is to provide formulas quantifying our degree of belief in a null hypothesis or in an alternative hypothesis respectively, as a function of the significance level of the experiment (itself a function of sample size) and of the computed p-value.

2. Model Specification

The goal of this section is to leverage the highest possible predictive power (large R^2 and low p-value) for a model at the lowest possible cost in degrees of freedom.

Most of the landscape of statistical practice lies within a framework in which some predictor or explanatory variables predict or explain some response variables. If the number of response variables is zero, we are doing descriptive statistics of a sample. If the number of explanatory variables is zero, we are doing simple estimation of population parameters. Otherwise, we are in the realm of predictive modeling.

The core framework for predictive modeling among quantitative variables is the general linear model $\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{U}$, where $\boldsymbol{X}$ is a matrix of observations of the explanatory variables, $\boldsymbol{Y}$ is a matrix of observations of the response variables, $\boldsymbol{B}$ is a matrix of model parameters, and $\boldsymbol{U}$ is a matrix of errors. This model unites linear regression, ANOVA, and related tests. There are well-known assumptions required for the model's validity, but the meaning of these assumptions is not completely appreciated.

The basic assumption underlying the general linear model is that the relation, or signal, between predictors and responses is essentially linear, if we overlook some noise. But one of the most basic facts about a non-constant linear function is that both its domain and range must be the full real number line, $(-\infty, \infty)$. In the context of a statistical model involving interval variables, that means that the components of both $\boldsymbol{X}$ and $\boldsymbol{Y}$ must have support equal to the full real number line. In other words, those components must be capable, at least in theory, if not in actual observation, of achieving any real value. Often, however, random variables encountered in practice are supported instead on a semi-infinite interval such as $(0, \infty)$, variables such as measurements of a physical quantity like length/distance, area, volume, mass/weight, age/time/duration, or other positive unbounded quantities like income; or they may be supported on a bounded interval such as $(0,1)$, variables such as proportion.

Another key assumption is that $\boldsymbol{U}$ be multivariate normal with mean equal to the zero vector; but it may not be so in fact, due, among other reasons, to the support of $\boldsymbol{Y}$. Shape issues like this and the linearity assumption are usually treated by applying a Box-Cox transformation to the dependent variable. However, this approach has problems.

1) This does not address the issue of the support of the independent variables.

2) the range of the Box-Cox transformation cannot be a symmetrically distributed random variable for any value of the Box-Cox exponent λ other than zero (the logarithmic transformation).

3) The Box-Cox transformations are employed most often to reduce skewness, but leave problems with kurtosis.

4) Substantial deviation from normality in the image variable often causes researchers to give up on parametric methods, relying on less powerful non-parametric methods instead.

Generalized linear models form only a partial answer to violations of the assumptions of the general linear model. The general linear model implies that $E(\boldsymbol{Y}) = \boldsymbol{XB}$, where the left side of the equation represents expected value, or mean. Then the generalized linear model is $g(E(\boldsymbol{Y})) = \boldsymbol{XB}$, where g is a link function. This more general model, including the original general linear model above together with logistic regression and Poisson regression, helps with the support of $\boldsymbol{Y}$, but not with the support of $\boldsymbol{X}$. We propose an algorithm to address the support of all quantitative variables in the model, and then to address their normality.

Let us begin by considering a random variable supported on an interval. If the variable is associated with some well-known process (for example, exponential or uniform), there is no need for further transformation. In practice, however, we often deal with unknown variables from which various observations have been drawn, and it is our job to discover the identity of those variables. Ideally, we hope to fit to the observations a random variable taken from a family of probability distributions with very few parameters to be estimated.

Another approach is to transform the sought-after variable to approximate normality, using the fewest and simplest transformations possible, to enable the use of parametric methods for estimation and prediction. Our first challenge is to convert the support of an interval variable x to a variable supported on $(-\infty, \infty)$. These transformations $f(x)$ and their inverses $f^{-1}(x)$ are given in Table 1, corresponding to the original support of x. After applying transformations of this form are performed, all interval variables in the dataset should be supported on $(-\infty, \infty)$.

Table 1. Support transformations and their inverses

support of x	$f(x)$	$f^{-1}(x)$
$(-\infty, \infty)$	x	x
(a, ∞)	$\ln(x-a)$	$a + e^x$
$(-\infty, b)$	$-\ln(b-x)$	$b - e^{-x}$
(a, b)	$\ln\left(\frac{x-a}{b-x}\right)$	$\frac{b\,e^x + a}{e^x + 1}$

All transformations in the four classes above have many common features: they are increasing, smooth, elementary functions with increasing, smooth, elementary inverses, and with range equal to $(-\infty, \infty)$. If we permit linear transformations of the range, we can standardize the functions in all four classes so that $f(0) = 0$, $f'(0) = 1$, and then the functions in the first three classes are seen to be limits of the functions in the fourth class, allowing a to decrease without bound and/or b to increase without bound, as necessary. Assume that $a < 0 < b$. Then to map the interval (a, b) to $(-\infty, \infty)$, the general transformation we seek has the following formula:

$$f(x) = \frac{1}{1/b - 1/a} \ln\left(\frac{1 - x/a}{1 - x/b}\right)$$

Its inverse is:

$$f^{-1}(x) = \frac{e^{x/b} - e^{x/a}}{e^{x/b}/b - e^{x/a}/a}$$

In the special case where $a = -1$ and $b = 1$, $f(x)$ is Fisher's z-transformation, and its inverse is $\tanh(x)$, the hyperbolic tangent. In the special case where $a = -1$ and b increases without bound, in the limit, $f(x) = \ln(1 + x)$, the shifted logarithm transformation for variables supported on $(-1, \infty)$, such as interest or return.

Figure 1 depicts the transformation $f(x)$ above in the cases where (a, b) is $(-1,1)$ (green), $(-1,2)$ (red), $(-1, \infty)$ (blue), $(-2, \infty)$ (orange), and $(-\infty, \infty)$ (black) respectively.

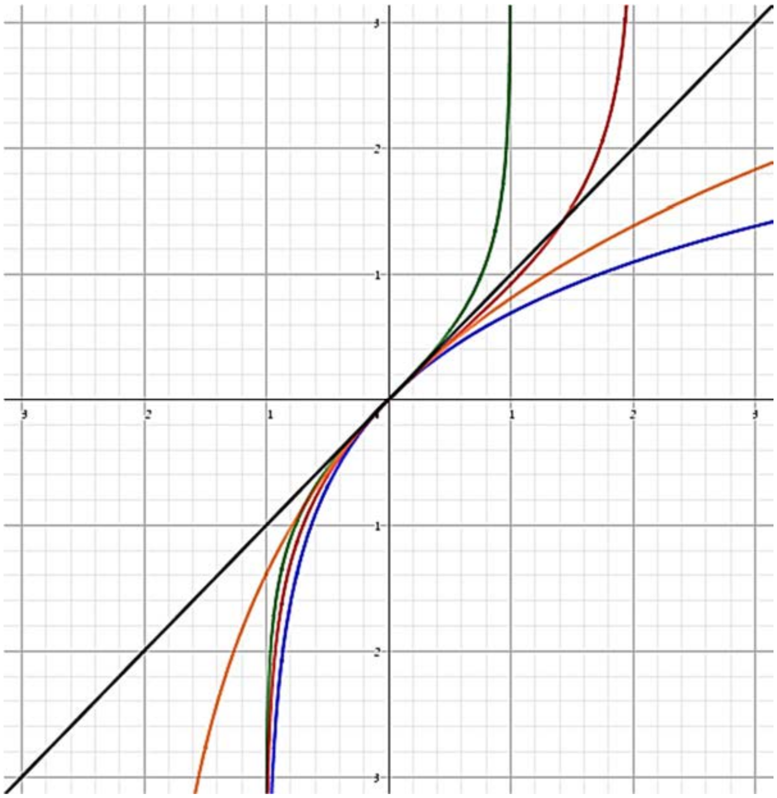

Figure 1. Support transformations

We now have the tools to transform any interval variable to one supported on $(-\infty, \infty)$. The other commonly encountered type of random variable is a discrete variable whose support equals the set of integers within an interval. If this set is bounded, it has the form $\{M, \ldots, N\}$, so one should employ the transformation $f(x) = \ln((x - a)/(b - x))$, with $a = M - 1$ and $b = N + 1$. This choice of (a, b) represents the largest open interval that does not contain any integers outside the set $\{M, \ldots, N\}$. In the popular case where $M = 0$, such as binomial variables, then $f(x) = \ln((x + 1)/(N + 1 - x))$ and $f^{-1}(x) = ((N + 1)e^x - 1)/(e^x + 1)$. If the support of the discrete random variable has the form $\{0,1, \ldots\}$, such as Poisson or negative binomial variables, then one should employ the transformation $f(x) = \ln(1 + x)$; here, $f^{-1}(x) = e^x - 1$. The transformations given above mimic the link functions that are used in generalized linear models, and the inverses of such functions.

As an example, a bivariate dataset containing (2,4), along with a few points to the lower-left and a few points to the upper-right, might be described by the models $y = x + 2$, or $y = 2x$, or $y = x^2$, or $y = 2^x$. The researcher must have a deep acquaintance with the subject matter, or consult another expert with such deep acquaintance, to determine the range of theoretical variation (that is, the support) of the variables in the dataset.

We illustrate with data from the statistics of income, from the Internal Revenue Service; see citation. This dataset reports the cumulative proportion of individual income tax returns with adjusted gross income not exceeding a given level, for the tax year 2014. Income is supported on $(0, \infty)$, while proportion is supported on $(0,1)$. If we attempt to predict (cumulative) proportion as an approximately linear function of income, the positive association cannot mask the poor fit ($R^2 = 0.2349$); but logit proportion is predicted well as a linear function of log income ($R^2 = 0.9850$); see figure 2.

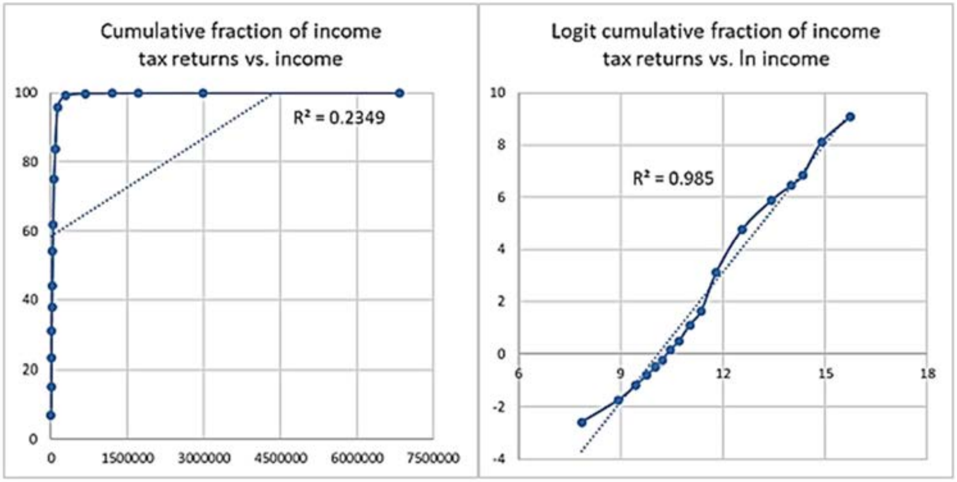

Figure 2. Fitting linear models to data before and after support transformation

Once we have performed the support transformations on all random variables in our model, $\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{U}$ is at least a theoretical possibility. But the distributions of errors are still a source of concern. We must consider the possible necessity of further transformations of our variables towards normality.

It should be well-known that to reduce multicollinearity in linear models, one tool is standardization of variables. This consists of two steps: centering (setting the mean equal to zero by subtracting the current mean), followed by scaling (setting the standard deviation to one by dividing by the current standard deviation).

Unfortunately, there is still confusion regarding the effectiveness of standardization of variables in reducing multicollinearity. Some practitioners point out that linear transformations of variables do not affect their correlations. However, we also need to be concerned about the stability of parameter estimates for models. Multicollinearity may be measured most simply by the condition number of the covariance matrix of the parameter estimates. And indeed, the condition number can be reduced by standardization of variables; see Kleinbaum et al, section 14.5.2, and Kutner et al, section 11.2.

Centering and scaling are the transformations that would be employed in the first two steps of a potentially infinite process to set the moments of a random variable equal to the moments of the standard normal random variable.

Substantial deviation from normality on the part of the residuals of a linear model can invalidate inferences made using that model. The Box-Cox family of transformations is not a satisfactory solution to the problem of non-normality, for reasons that were discussed above; and non-parametric methods yield power too easily, raising the risk of a false negative conclusion (type II error).

Instead, we suggest continuing the process of normalization that was begun by standardizing: that is, centering plus scaling. For each of these steps, we applied an appropriate member of a one-parameter family of functions: $x - \mu$ for centering, and x/σ for scaling. These two families have several important properties in common. 1) They are smooth, elementary functions of x and the parameter. 2) For each value of the parameter, the function of x is an increasing function with domain and range equal to $(-\infty, \infty)$. 3) The identity function of x is a member of the family, for some value of the parameter. 4) For each value of the parameter, the inverse function of x is a member of the same family, with some other value of the parameter.

For the first family, given any random variable with a finite mean, there is a unique value of the parameter such that the mean of the transformed variable equals zero. For the second family, given any random variable with a finite variance, there is a unique value of the parameter such that the variance of the transformed variable equals one. It is reasonable to speculate about whether there exist analogous families of functions capable of transforming skewness and excess kurtosis to zero.

To minimize excess technical detail in this presentation, we do not comment on additional properties satisfied by the families of centering and scaling functions. These properties enabled the discovery of two analogous families of functions, one intended to remove skewness and the other intended to remove excess kurtosis. Rather, we will simply present the results. For both new families, it is assumed that the original data has already had its support transformed to $(-\infty, \infty)$, and has then been standardized (centered and scaled). Importantly, data should be centered by subtracting its median, not its mean, to set the median to zero, as will be discussed below.

The de-skewing functions have the form:

$$f(x) = \frac{(x(1+k^2)+k^2-1) \pm \sqrt{(x(1+k^2)+k^2-1)^2 - 4kx(kx+k^2-1)}}{2k}$$

Here, k is a parameter that ranges over $(0, \infty)$. The graph is a branch of a hyperbola; the selected branch passes through the origin with slope one. The + branch passes through the origin when $0 < k < 1$, and is used to de-skew negatively skewed random variables; the – branch passes through the origin when $k > 1$, and is used to de-skew positively skewed random variables. The inverse transformation is in the same family, with parameter $1/k$.

Figure 3 depicts the de-skewing function $f(x)$ above for the cases $k = 1/2$ (red), $k = 1$ (green, for comparison), and $k = 2$ (blue). As k decreases but remains positive, the graph of $f(x)$ is bowed more sharply upward. As k increases, the graph of $f(x)$ is bowed more sharply downward.

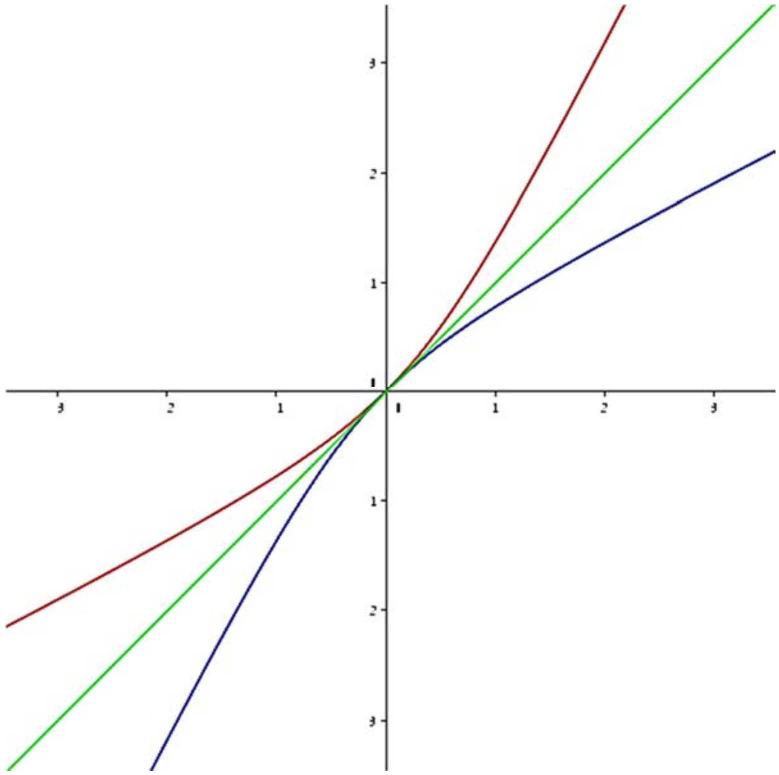

Figure 3. De-skewing functions

The de-skewing functions are order-preserving, and fix zero. If the median of the standardized data is zero, the median of the de-skewed data will be zero, implying that the mean will be zero or close to zero. This is the reason why we centered by subtracting the median rather than the mean of the original data.

There are two related families of functions that are used to address issues of kurtosis, assuming the data has already been standardized and de-skewed. Random variables with low excess kurtosis (light-tailed) are treated with $f(x) = \text{sinh}\ (kx)/k$, where $k > 0$, and sinh represents the hyperbolic sine; random variables with high excess kurtosis (heavy-tailed) are treated with $f(x) = \text{asinh}\ (kx)/k$, where $k > 0$, and asinh represents the inverse hyperbolic sine. In both cases, the functions approach $f(x) = x$, the identity function, as k approaches zero.

Figure 4 depicts instances of the functions above. Functions to increase low excess kurtosis are shown for $k = 1$ (green) and $k = 2$ (red); functions to decrease high excess kurtosis are shown for $k = 1$ (blue) and $k = 2$ (orange); the limiting case as k approaches zero is shown for comparison, in black.

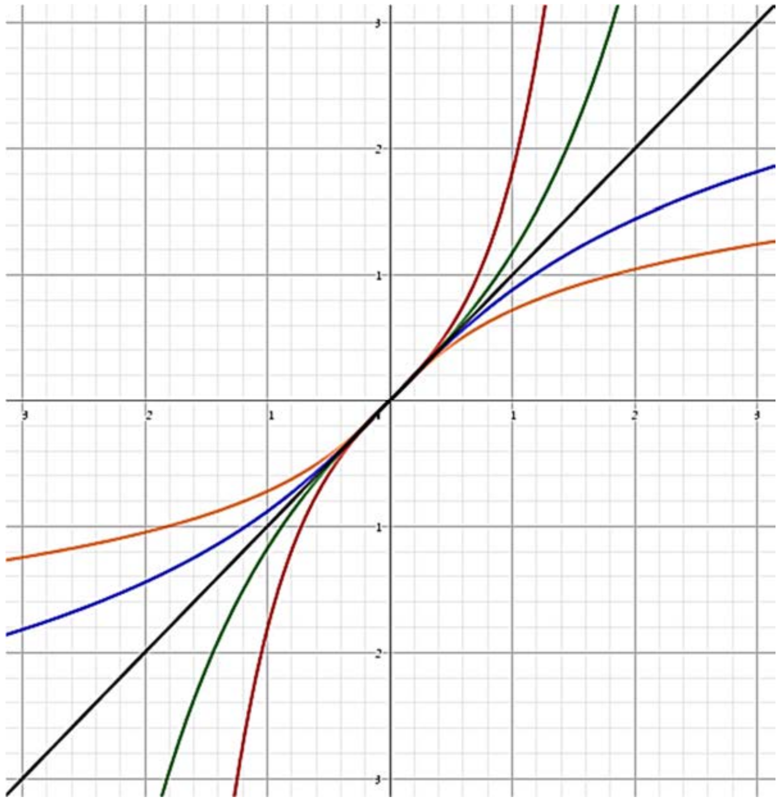

Figure 4. Functions to transform excess kurtosis to zero

We emphasize that the functions used to eliminate skewness and excess kurtosis are ineffective unless the variable has been transformed previously to have support $(-\infty, \infty)$.

To test the power of these new families of transformations, we tested them in R on large random data sets drawn from various members of the Pearson distribution family, as well as on daily return data from a stock market index. We generated random samples of size 2^{24} (about 16 million) drawn from the following distributions: 1) beta with parameters one-half and one-half (the arcsine distribution); 2) uniform (that is, beta with parameters one and one); 3) chi-squared with 1 degree of freedom (the most skewed chi-squared distribution); 4) Student's t with 5 degrees of freedom (the heaviest-tailed t distribution with finite kurtosis); 5) inverse chi-squared with 9 degrees of freedom (the most skewed inverse chi-squared distribution with finite kurtosis). These distributions are described in table 2.

Table 2 includes a column with an apparently new concept: surplus kurtosis, defined as the difference of excess kurtosis minus skewness squared. This quantity varies in the range $[-2, \infty)$ for any value of skewness, provided the distribution has finite kurtosis, and equals -2 for Bernoulli distributions and only for Bernoulli distributions.

Table 2. Random samples from these distributions transformed towards normality

distribution	skewness squared	excess kurtosis	surplus kurtosis = excess kurtosis – skewness squared
$\beta(1/2,1/2)$ (arcsine)	0	-1.5	-1.5
$\beta(1,1)$ (uniform)	0	-1.2	-1.2
$\chi^2(1)$	8	12	4
Student's $t(5)$	0	6	6
Inverse $\chi^2(9)$	$160/9 \approx 17.78$	92	≈ 74.22

Our financial dataset consisted of the daily percent change in the Wilshire 5000 stock market index, from December 3, 1979 to September 29, 2016, with a total of 8972 observations; this dataset exhibited negative skewness (mean less than median) and positive kurtosis (larger than normal fraction of extreme events), as is well-known among financial analysts. This last dataset was retrieved from FRED, the research division of the St. Louis Federal Reserve Bank, using the series identification WILL5000INDFC: see citation.

For each dataset, data were transformed, based on their theoretical support, to have support on $(-\infty,\infty)$; then data were standardized; then data were transformed using the functions above to remove skewness and excess kurtosis. For each dataset, the goal was to produce a transformed dataset with mean equal to zero, standard deviation equal to one, skewness and excess kurtosis equal to zero. By searching in the parameter space, it proved possible to locate rational values of the parameters allowing skewness and excess kurtosis to approach zero as closely as desired.

To test whether the transformed data are approximately normal, histograms and Q-Q plots with respect to normality were produced, and the Shapiro-Wilk normality test was applied. It should go without saying that the d'Agostino normality test, which looks for values of skewness and excess kurtosis to be close to zero, was passed quite easily! Furthermore, the fifth and sixth standard moments of the transformed data were computed, in the hope that these values would be close to 0 and 15, the values for the standard normal distribution. For each dataset and for each test, the results were very satisfactory.

We illustrate for the case of the stock market return data. Figure 5 depicts histograms and normal Q-Q plots for the original data; for the data transformed to have support on $(-\infty,\infty)$; and for the final transformation to eliminate skewness and then excess kurtosis. The final Q-Q plot is (mostly) pleasingly straight, implying a very good approximation to normality.

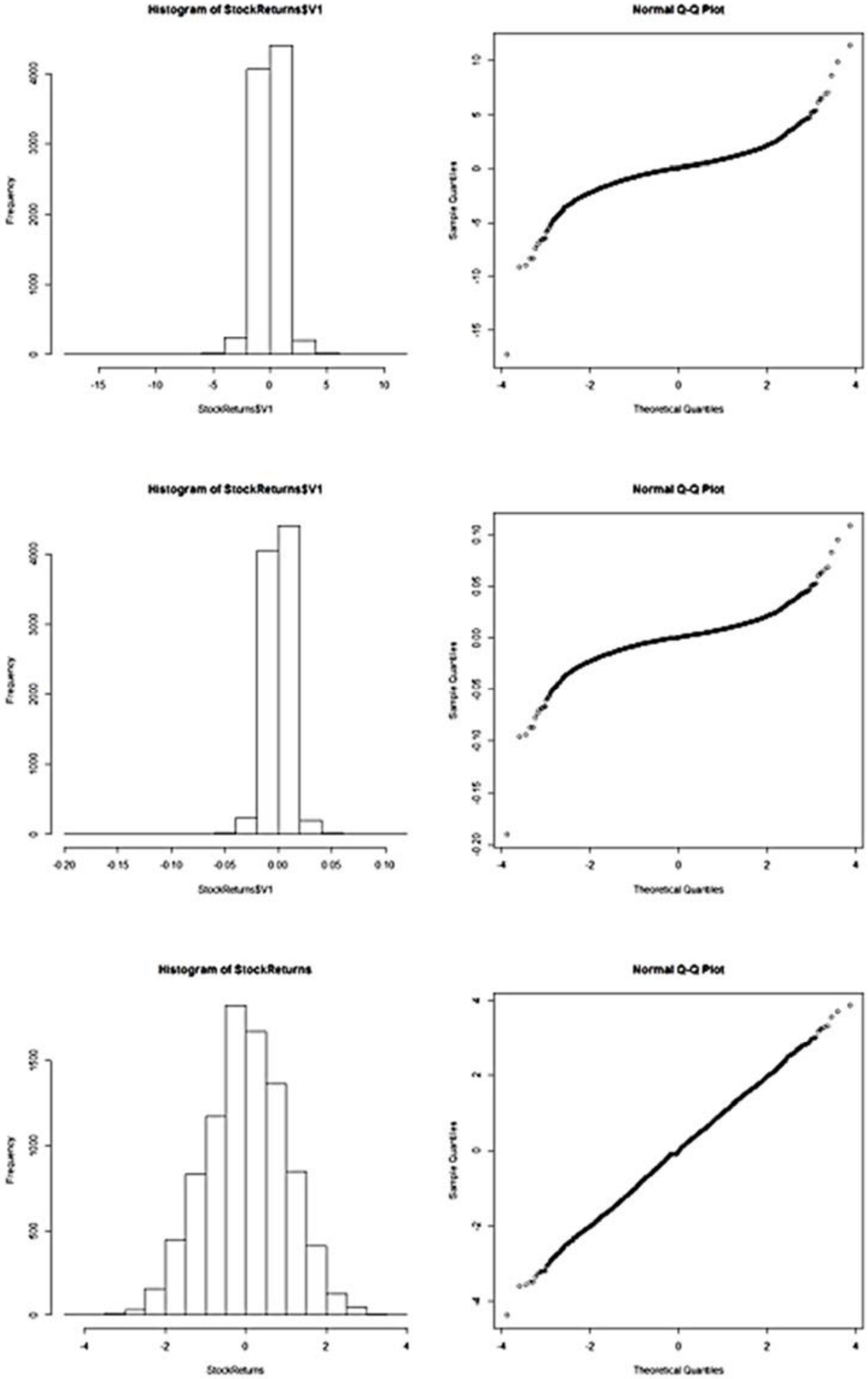

Figure 5. Histograms and Q-Q plots for stock return data before, during, and after transformation

For the record, in the original stock return data, skewness was -0.62, and excess kurtosis was 15.96. After transforming support to $(-\infty, \infty)$, skewness was -0.93, and excess kurtosis was 19.06. This new data was standardized by subtracting the median and dividing by the standard deviation. Applying the de-skewing function with parameter 55/64 resulted in skewness equal to -0.001. Excess kurtosis remained at 15.08. Finally, applying the function to reduce excess kurtosis with parameter 49/37 resulted in excess kurtosis equal to 0.001. Both skewness and excess kurtosis could have been brought closer to zero with a longer search. For this final version of the data, the standardized fifth and sixth moments were -0.28 and 15.25 respectively, surprisingly close to the values expected under full normality (0 and 15). For the transformed data, a random sample of size 720 gave a Shapiro-Wilk statistic of 0.99721, with p-value 0.2574.

The process described above is a method to transform a random variable to standard normality. One could reverse this process, starting with a standard normal distribution and then applying the inverse transformations in the reverse order: add or subtract kurtosis from a standard normal random variable; add or subtract skewness; apply a linear transformation to alter the scale and center of the distribution; apply shifted exponential or shifted logistic functions to transform support from $(-\infty, \infty)$ to a semi-infinite or bounded interval; and conclude with a final linear transformation. This yields an 8-parameter family of distributions that includes all families of the Johnson distributions (including the log-normal and logit-normal distributions), the Gumbel and logistic distributions, and apparently, if the experiments above are an indication, distributions well approximating all the Pearson distributions as well (such as arcsine, beta, beta prime, chi-squared, exponential, F, gamma, inverse chi-squared, inverse gamma, normal, Student's t, uniform); see Hahn and Shapiro, chapter 6. Parameter estimation would be challenging except for very large datasets, due to the increasingly large variance in the distribution of sample moments as the order increases.

Nevertheless, the purpose of this section was not to produce new families of distributions, as useful as they would be for prediction of univariate phenomena, but rather to find new elementary transformations to uncover normal behavior underlying variables supported on intervals. These techniques should maximize the usefulness of powerful parametric models, thus locating the signal amidst the noise, increasing R^2, decreasing the p-value, and reducing the probability of a type II error (false negative).

3. Significance Level and Sample Size

Skepticism about classical null hypothesis significance testing in recent years has tarnished unjustly the reputation of the p-value. In fact, all the skepticism directed at the p-value should have been directed instead at the unjustified assumption that the significance level α for most experiments should remain constant, at 0.05, or indeed, constant at any pre-specified level. The goal of this section is to argue that the significance level should decrease as a function of sample size. We tentatively propose a formula for α and a means to employ it in predictive modeling.

For much of modern statistical history, null hypothesis significance testing has relied on the comparison of the p-value to a seemingly arbitrary value of 0.05, since approximately 5% of the data in a normal distribution is more than 2 standard deviations away from the mean, and 5% and 2 are nice, round numbers that are easy to remember. It should be self-evident that 0.005, the suggested replacement for 0.05, is just as arbitrary a number as 0.05.

However, common test statistics are functions of two quantities: observed effect size relative to variability, and sample size. Given two experiments that detect the same relative effect size, the one with the larger sample size will yield a more extreme value of the test statistic, and hence a smaller, more significant p-value. This is all as it should be; the more we can reproduce a non-trivial effect, in violation of the null hypothesis, the more significant the result appears to be.

This last observation points out the weakness of the argument that statistical tests should report only effect size, as important as that is. A knowledge of the sample size must be the inseparable companion of a knowledge of the effect size. Indeed, it can be argued that the sample size is the first, or perhaps the zeroth, in a series of numerical characteristics of a quantitative sample, a characteristic that is followed by the mean, variance, skewness, and kurtosis.

Since a larger sample size causes a fixed effect size to become more significant automatically, it follows that our problem is not with the p-value itself, but rather in the target for which we aim. That target is the significance level, α, which should itself decrease as a function of sample size. We expect that such a methodology would reduce the frequency of false positives and false negatives, compared with introducing a new target level for α, such as 0.005.

Series of observations taken across time provide yet another argument in support of decreasing significance levels as a function of sample size. Consider a series of annual observations of some quantity of interest, such as median income adjusted for inflation. Our research interest is to determine whether there has been a significant change in this quantity during the period of observation. There may be a slight trend in the annual data, a trend that is not significant in the traditional sense. However, if more frequent data becomes available, such as quarterly or monthly data, the trend suddenly becomes significant. The absolute value of the test statistic will increase automatically, because, as discussed below, the test statistic is approximately proportional to the square root of the sample size. Hence, the target significance level needs to decrease in response.

What should the target significance level be? We shall present heuristic arguments for several candidate functions, and choose among them.

We begin by recalling the central limit theorem: when sample sizes are sufficiently large, sample means taken from a population with finite variance are distributed approximately normally. Suppose we perform an experiment, consisting of drawing n paired observations from two normal populations with a common, known standard deviation. Our null hypothesis is that the means of the two populations are equal; thus, that the difference of the means is zero; thus, that the mean of the differences is zero.

If the null hypothesis is true, then $z = \bar{x}/\left(\sigma/\sqrt{n}\right) = \sqrt{n}(\bar{x}/\sigma)$ is normally distributed with mean zero and standard deviation one, where $\bar{x}$ is the sample mean difference, and σ is the standard deviation of the population of differences. The relative effect size is just $\bar{x}/\sigma$, but the test statistic incorporates the sample size, being $\sqrt{n}$ times as large. The p-value of the test statistic is the probability that a standard normal random variable has absolute value greater than that of z. For the sake of argument, our first candidate for the significance level function is the p-value of z above, where the relative effect size is one: in other words, the probability that a standard normal random variable has absolute value larger than $\sqrt{n}$.

In practice, if the population mean is not known, it is unlikely that the population standard deviation would be known. We approximate the latter with the sample standard deviation s. Then $t = \bar{x}/\left(s/\sqrt{n}\right) = \sqrt{n}(\bar{x}/s)$ is distributed as a Student's-t random variable, with $n - 1$ degrees of freedom. Our second candidate for the significance level is the p-value of t, where the relative effect size is one: in other words, the probability that a t-variable with $n-1$ degrees of freedom has absolute value larger than $\sqrt{n}$.

The pattern of test statistics from the Student's-t family (including normal random variables as a limiting case) being on the order of the square root of the sample size extends to another case: that of the sample correlation. Assuming data comes from a bivariate normal distribution with zero correlation, if we draw a sample of size n with correlation r, then $t=\sqrt{n-2}\left(r/\sqrt{1-r^2}\right)$ is distributed as a Student's-t random variable, with $n-2$ degrees of freedom. This observation will be used below.

We depict both significance level candidate functions in Figure 6, where the vertical axis gives probabilities on a logarithmic scale. The z tail probabilities are shown in orange, while the t tail probabilities are shown in gray. Both graphs are asymptotically linear, indicating that these functions decrease asymptotically exponentially as functions of n. The increasing gap between the two graphs shows that while the absolute difference between the two tail probabilities decreases towards zero as n increases, the relative difference increases without bound. (Incidentally, this demonstrates that we mislead our students when we teach them that the normal distribution is a satisfactory approximation to the Student's-t distribution when the number of degrees of freedom is at least 30; the approximation is not good in the tail.) The green line is the traditional level $\alpha=.05$. The two other functions in the graph will be described below.

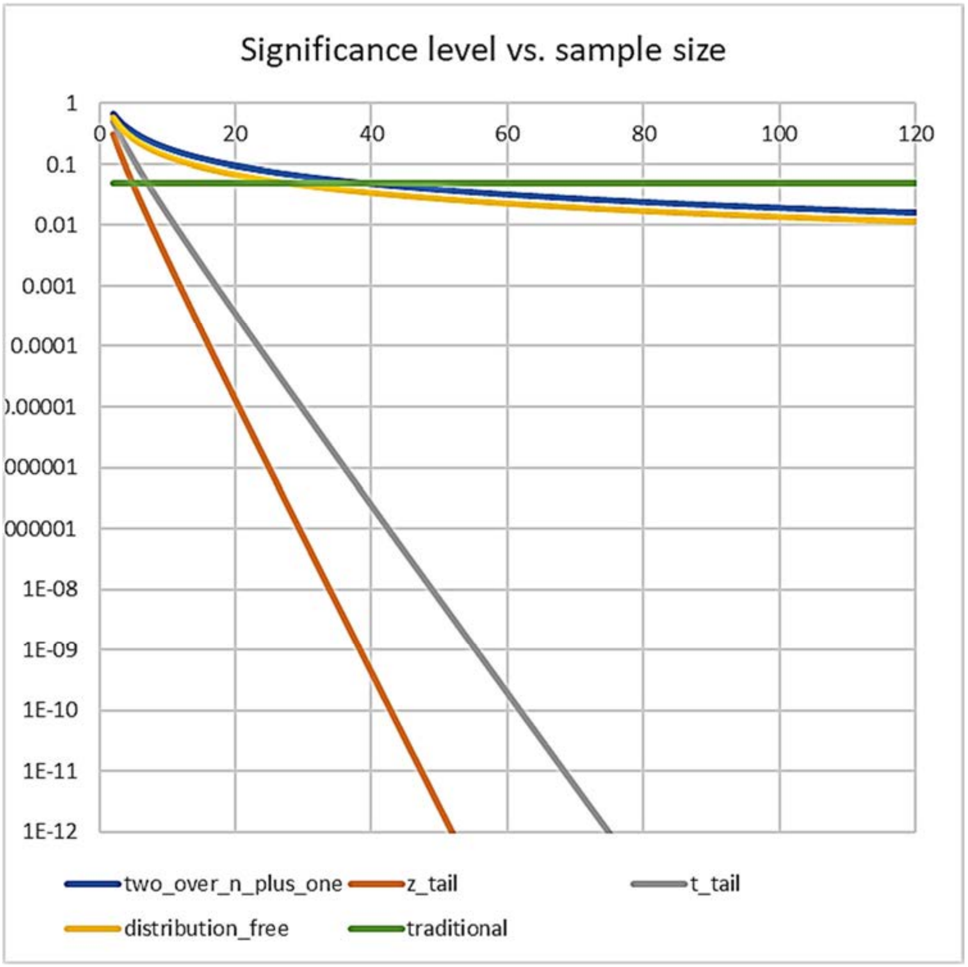

Figure 6. Candidate targets for significance level as a function of sample size

There is another interpretation for the values of these two significance level functions: namely, as sample mean and standard deviation remain fixed, or alternatively, as relative effect size remains fixed, these are the values of α for which confidence intervals maintain constant width as sample size changes.

The rapid fall-off in even the t tail probabilities implies that this significance level target is too strict; that is, it puts researchers at risk of type II (false negative) errors nearly all the time. This would be the case even if we chose a constant relative effect size less than one. We need an argument allowing for a more lenient standard, one that makes it somewhat easier (but not too easy!) to reject the null hypothesis. We consider such a heuristic argument below.

Consider a location on earth near the equator, where temperatures remain approximately constant throughout the year. At this location, we record the high temperature every day. Every time the newest daily high temperature reaches a record, either higher than any previous measurement or lower than any previous measurement, we sound an alarm indicating that a noteworthy event has taken place.

The daily high temperature is a continuous variable. We assume that our thermometer is capable of arbitrarily high precision. The probability that any two days have high temperatures that match exactly is zero. Suppose there are n observations so far, and we want to know how likely it is that the next observation will be a record. The $n + 1$ observations could fall in any of $(n + 1)!$ orderings, all equally likely, independent of the distribution of the temperatures themselves. The newest observation is the minimum of the set in $n!$ cases, and is the maximum of the set in $n!$ cases. Thus, the probability that the newest observation is the minimum of the set is $1/(n + 1)$, the probability that the newest observation is the maximum of the set is $1/(n + 1)$, and the probability that the newest observation is a record of either sort is $2/(n + 1)$.

Our third candidate for the significance level is $2/(n + 1)$, where n is the sample size. It appears in Figure 5 as well, as a blue curve. This new significance level coincides with the traditional alpha, 0.05, when n is 39. When n is less than 39, it is easier to reject the null hypothesis compared to the traditional standard, but as n increases beyond 39, it is very gradually increasingly difficult to reject the null hypothesis. This accords with our intuition that when a given effect size is observed and replicated in larger and larger samples, the results are increasingly significant. Thus, we should demand a greater level of surprise, and therefore a smaller p-value, when the sample size increases, to justify rejecting the null hypothesis.

This line of reasoning invites further comment. It would appear that the mere replication of an effect size in larger and larger samples demonstrates the very noteworthiness we seek in rejecting the null hypothesis, and thus that there is no need to have the significance level decrease with sample size. The best argument against this conclusion is our observation of statistical practice, wherein statistically significant conclusions are frequently found on opposing sides of an argument, to the point that the public loses faith in scientific announcements; see Bohannon. It is vital that the award of noteworthiness not be conferred upon the results of an experiment merely by virtue of the large sample size. Moreover, our approach slows the decrease in the minimum significant effect as a function of sample size.

In a private communication, statistician William Huber described the justification above of the formula $2/(n + 1)$ as a thought experiment. It does not carry an obvious theoretical imperative. We can alter this thought experiment and the resulting formula slightly to give it a theoretical basis.

Imagine that the series of equatorial temperatures is n observations long. Then the probability that the next observation will set a record is $2/(n+1)$. This formula is recognizable in the context of plotting positions for a Q-Q plot, if one uses the $k/(n+1)$ approach to plotting the k -th order statistic from a set of n observations. The formula $2/(n+1)$ equals the probability that a new observation drawn from a continuous uniform distribution is less than the minimum, or more than the maximum, of a set of n observations drawn from that same uniform distribution; see Wackerly, Mendenhall, and Scheaffer, chapter 6.

Other approaches to the plotting position problem can suggest other heuristic arguments, other probabilities, and other formulas for the target significance level. These new formulas may have the form $(2-2a)/(n+1-2a)$, where a lies in the interval $[0,1)$; the formula proposed in the previous paragraph corresponds to $a=0$. As a increases, the fraction above decreases; therefore, out of all significance levels defined by the fraction above, the significance level $2/(n+1)$ makes rejection of the null hypothesis the easiest.

The formula $k/(n+1)$ represents the mean of the beta distribution with parameters k and $n-k+1$, which is the distribution of the k -th order statistic from a set of n observations drawn independently from the uniform distribution on $(0,1)$; hence $k/(n+1)$ is the mean of the k -th order statistic from the uniform distribution; again see Wackerly, Mendenhall, and Scheaffer, chapters 4 and 6. We may prefer a distribution-free approach, in which case we should consider the median of the k-th order statistic, since medians are preserved by order-preserving transformations of random variables. Although there is no closed formula for the general case of the median of the beta distribution with parameters k and $n-k+1$, we are only concerned here with the location of the median values of the minimum ($k=1$) and maximum ($k=n$), for which closed formulas do exist: $1-2^{-1/n}$ for the case $k=1$, and $2^{-1/n}$ for the case $k=n$. Our fourth candidate for the significance level is $2\left(1-2^{-1/n}\right)$. This is depicted as a yellow curve in Figure 5, and represents a slightly stricter standard of significance than $2/(n+1)$, being smaller by a factor of approximately $\ln 2$. This is closest to 0.05 when n is 27.

These two newest candidates for the significance threshold, based on order statistics, are larger than the first two candidates, which were based on a fixed relative effect size. We observed that confidence intervals constructed using the first two significance thresholds would have constant width regardless of sample size, by design. Hence confidence intervals constructed using the larger significance thresholds based on order statistics will become narrower as sample size increases, as is the case with current standard practice. The difference is that confidence intervals will become narrower with increasing sample size more slowly under our proposal, compared with current practice.

Next, we compute critical values of z, t, the relative effect size or signal-to-noise ratio (SNR), and r. Assuming that a random variable follows a standard normal distribution, or a t distribution with $n-1$ degrees of freedom, or represents a sample correlation from a bivariate normal distribution with zero population correlation, Figures 7 and 8 depict the minimum values of the absolute values of z and t (Figure 7), and SNR and r (Figure 8) respectively that would be considered significantly different from zero, versus the sample size n, using the two-sided significance thresholds $2/(n+1)$ (labeled “mean” in the diagram) and $2\left(1-2^{-1/n}\right)$ (labeled “median” in the diagram). Traditional values based on two-sided tests with $\alpha=.05$ are shown as well. The growth rate for the newly proposed critical values of the absolute values of z and t is approximately on the order of

$O\left(\sqrt{\ln(n)}\right)$, a very slowly growing function of n; and we compute $r = t/\sqrt{n-2+t^2}$, employing t with $n-2$ degrees of freedom.

The critical value for the relative effect size or SNR very closely tracks that of r. It also represents the radius of the confidence interval for the population mean around the sample mean, as measured in multiples of the population (respectively sample) standard deviation. Thus the order of SNR and of r is approximately $O\left(\sqrt{\ln(n)/n}\right)$.

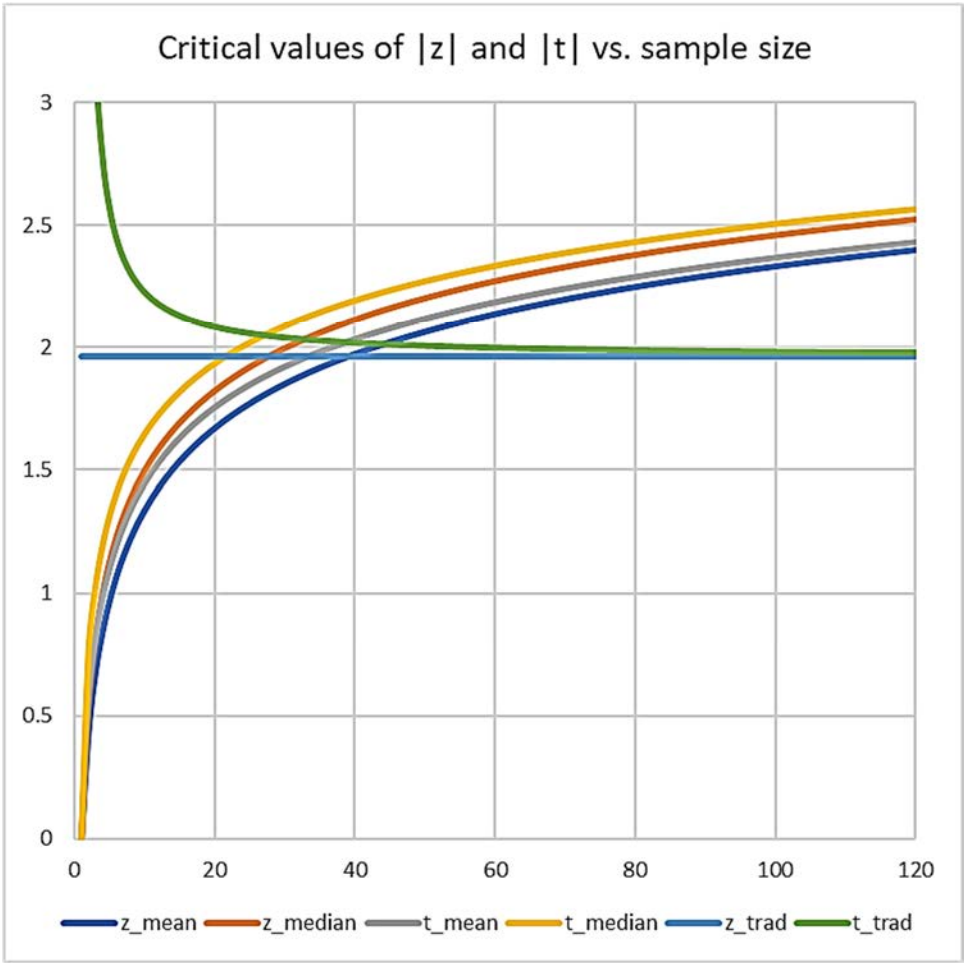

Figure 7. Critical values of |z| and |t| vs. sample size

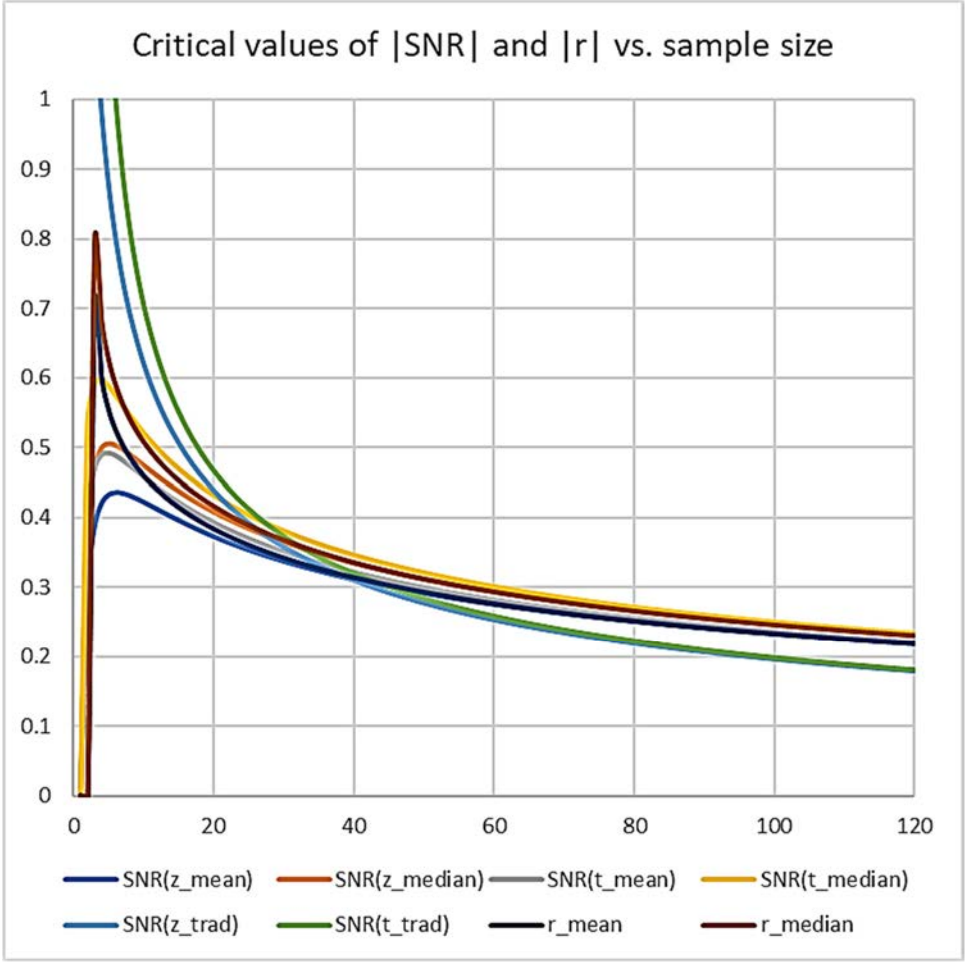

Figure 8. Critical values of |SNR| and |r| vs. sample size

For comparison with traditional practice, we offer in table 3 a brief selection of significance levels, and minimum significant positive levels for t, SNR, and r, associated to various sample sizes, in table 2, using the distribution-free approach, where $\alpha = 2\left(1 - 2^{-1/n}\right)$.

Table 3. Significance levels and minimum significant positive levels for t, SNR, and r

n	α	t $df = n - 1$	$SNR =$ $\bar{x}/s = t/\sqrt{n}$	r $df = n - 2$
2	0.5858	0.7612	0.5383	–
6	0.2182	1.4079	0.5748	0.5895
24	0.0569	2.0044	0.4092	0.3938
120	0.0115	2.5663	0.2343	0.2299
720	0.0019	3.1131	0.1160	0.1154
5040	0.0003	3.6403	0.0513	0.0512

Recall the standard minimum sample size formulas for detecting differences in population means or population proportions respectively: $n = \left(z_{\alpha/2}\sigma/SE\right)^2$ and $n = \left(z_{\alpha/2}/(2SE)\right)^2$, where SE stands for sampling error. (The second formula is a conservative sample size formula, making no prior assumptions on the likely population proportion.) Traditionally, the significance level α has been set arbitrarily, but now we have made the argument that α should itself be a function of n. From this follows an algebraic consequence: that with a sample of size n, there is a minimum sampling error that can be prescribed, in the case of proportions; and in the case of means, there is a minimum ratio of sampling error to standard deviation.

In other words, given a sample of size n, there is a positive lower bound on the fineness of resolution with which we may distinguish between two populations. This bound decreases as a function of n, and is essentially the relative effect size or SNR in table 2. For the case of estimating population proportions, however, we no longer set α arbitrarily; rather, it is a function of n, which in turn will be computed based on the largest acceptable sampling error.

It is common in polling practice to cite a "margin of error" of $\pm 3\% = \pm .03$. This is achieved when the sample size n is at least 3492, when we solve $n = \left(t_{(\alpha/2),(n-1)}/(2 \times .03)\right)^2$, using $\alpha = 2\left(1 - 2^{-1/n}\right)$; but then $\alpha = .0004$, so the confidence level is 99.96%. In contrast, if we use the traditional sample size of approximately 1000, we maintain a high confidence level of 99.86% ($\alpha = .0014$), but now the margin of error is about $\pm 5\%$. All inferences from such a poll are predicated on the assumption that the sample is indeed representative of the population, of course.

With the figures above in hand, we can offer a tentative heuristic argument to support using these apparently arbitrary formulas for significance thresholds. If a null hypothesis is assumed to hold in some experiment, as we gather data, our attention will be drawn to the observation that appears to conflict with the null hypothesis most strongly; because if the most extreme observation does not shake our belief in the null hypothesis, the other observations will not do so either. Our curves represent the thresholds where one observation bursts beyond the bounds that would be expected from the behavior of the other observations.

These tentative heuristic arguments do not preclude further discussion and recommendations for alternative significance threshold functions, either stricter or more lenient. However, any such significance threshold function should lie between a constant significance function and one derived from a constant relative effect size. In particular, it is arguable that the significance level vary, presumably decreasing, with the number of predictors. We address this issue next.

Multiple tests will affect the formula for significance level. Suppose we wish to investigate whether Y is approximately a linear function of $X_1, \dots, X_k$ based on n observations. Assume that variables have been transformed as in the first section of this paper to have support on $(-\infty, \infty)$ and to be approximately normal, and assume the conditions of multiple linear regression are met. Suppose we have settled on some formula for a target significance level as a function of sample size and call it $\alpha(n)$.

Consider a forward selection process. We would be tempted to declare a significant correlation between Y and one of the X's if the p-value of the correlation were less than $\alpha(n)$. There is some risk of a type I error (false positive) in the case of one correlation;

but now we are permitting k different variables to compete for the honor of being in significant correlation with Y, so we are increasing our family-wise type I error risk.

This implies that the first variable, if any, to be declared in significant correlation with Y must meet a more stringent standard. If the multiple tests are known to be independent, we lower the significance threshold to $1-\left(1-\alpha(n)\right)^{1/k}$, the Šidák correction; if the multiple tests are not known to be independent, we use the even lower significance threshold $\alpha(n)/k$, the Bonferroni adjustment. When $\alpha(n)$ is sufficiently close to zero, these formulas converge, as shown in Figure 9, which shows the case where k is 2. The blue curve shows the Šidák correction, and the red line is the Bonferroni correction.

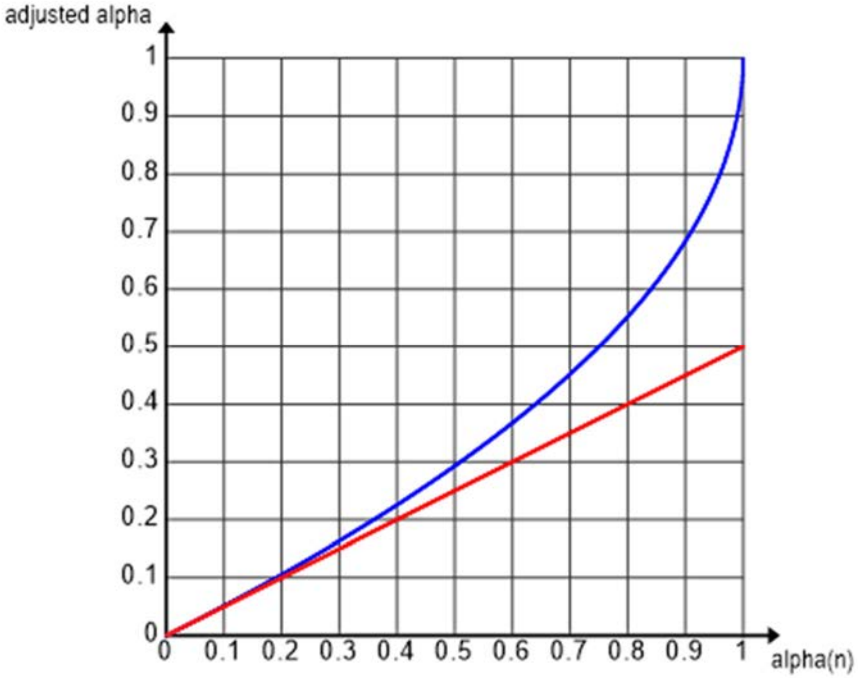

Figure 9. Adjusting the significance level for two tests

If there is a proposed predictor variable whose correlation with Y has a p-value less than $1-\left(1-\alpha(n)\right)^{1/k}$, or $\alpha(n)/k$ if the tests are not known to be independent, choose the most highly correlated variable as the first to enter the model. Once this relation has been established, there remain $k-1$ variables competing to enter the model as the second most significant variable. The threshold for significance is now $1-\left(1-\alpha(n)\right)^{1/(k-1)}$, or $\alpha(n)/(k-1)$ if the tests are not known to be independent; these are slightly larger. The most significant predictor variable, if any, meeting this threshold is added to the model. The significance threshold for the third variable will be $1-\left(1-\alpha(n)\right)^{1/(k-2)}$ or $\alpha(n)/(k-2)$ as the case may be, and so on.

The process described above is the Holm-Bonferroni method; see Holm. In this process, k represents the number of parameters that are estimated in a model. In a linear model including an intercept, k must represent the number of proposed predictors plus the intercept. It could represent the number of terms in an ARIMA model, the number of principal components in a dataset, the optimal number of clusters for a dataset, and so on.

This method of variable selection demands parsimony; the more variables one throws into one's dataset, the more likely it is that we will find a spurious correlation that is noise rather than signal. This requires us to find ways to reduce the dimension of the space of potential explanatory variables.

This goal can be accomplished by using principal components analysis on the proposed predictor variables $X_1, \ldots, X_k$, after those variables have been treated as described in the first section for support and normality issues. The resulting dataset should be a better approximation of a multivariate normal dataset. Once the principal components have been identified, one could then run principal components regression: that is, a multiple regression of Y against the principal components identified in the analysis of the proposed predictor variables $X_1, \ldots, X_k$. This should also help protect against potential multicollinearity in the model.

Although there has been some discussion of the idea of linking the reporting of p-values to the reporting of sample sizes, our specific proposal that the significance threshold should decrease with increasing sample size in a particular fashion appears to be new. The techniques discussed in this section, including setting the threshold significance level as a decreasing function of sample size based on the median of the minimum and the median of the maximum, together with principal components regression based on the Holm-Bonferroni method and the Bonferroni correction, should go a long way towards prevention of type I (false positive) errors, helping us to avoid being fooled by randomness, one theme of Taleb's Incerto works; also see Banerjee's new work on p-values.

Before leaving this subject, we consider an alternate approach to determining a threshold for significance that depends on sample size. This alternate approach makes use of the distribution of the sample coefficient of determination r^2. If the population coefficient of determination ρ^2 is zero, then $r^2 \sim \beta((k-1)/2, (n-k)/2)$, where k is the number of predictors in the model (including the constant term), and n remains the sample size; see for example the note of Papadopoulos. The mean of this distribution is $(k-1)/(n-1)$; see Wackerly, Mendenhall, and Scheaffer. Then we can declare the observed coefficient of determination to be significant if it exceeds this mean. That is, we declare significance if $r^2 > (k-1)/(n-1)$. This formula is easy to use and to explain.

In our previous threshold approach, we determined that the minimum significant value of $|r|$ was on the order of $\sqrt{\ln(n)/n}$, which implies that the minimum significant value of r^2 is on the order of $\ln(n)/n$. Given a fixed number of predictors k, the threshold $(k-1)/(n-1)$ makes it easier to find significance. In the interests of controlling false positives, we will not recommend using this alternate approach.

4. Continuous Decision Rules

If an outcome of interest has a continuous quantitative description, then an ordinal variable would be more informative than a nominal variable; a discrete quantitative variable would be more informative than an ordinal variable; and a continuous quantitative variable is the most informative of all. Decision making is traditionally presented in dichotomous, either/or, yes/no terms. Indeed, the consumers of statistics often require dichotomous judgments: is the transaction fraudulent; is the drug safe; and so on. We explore continuous interpretations of decision making.

The traditional framework of null hypothesis significance testing (NHST) is to contrast a null hypothesis (typically an equation or set of equations regarding population parameters) against an alternative hypothesis, the negation of the null hypothesis; set a significance level α between 0 and 1; collect data; compute a test statistic based on the data; compute the p-value of the test statistic under the null hypothesis; and then either reject the null hypothesis if the p-value is less than α, or fail to reject the null hypothesis otherwise.

Suppose, however, that before taking that last step, we express some uncertainty about our decision, wondering what might happen were the experiment to be repeated. We might attempt to compute a probability that the null hypothesis should be rejected. Traditionally, that probability was known to be one if the p-value is less than α; otherwise the probability was known to be zero.

Instead, we propose that the probability of rejecting the null hypothesis, $P(p)$, should be a non-increasing function of the p-value p, such that $P(0) = 1$ and $P(1) = 0$. Then the classic method of NHST is illustrated in Figure 10, for $\alpha = 0.05$.

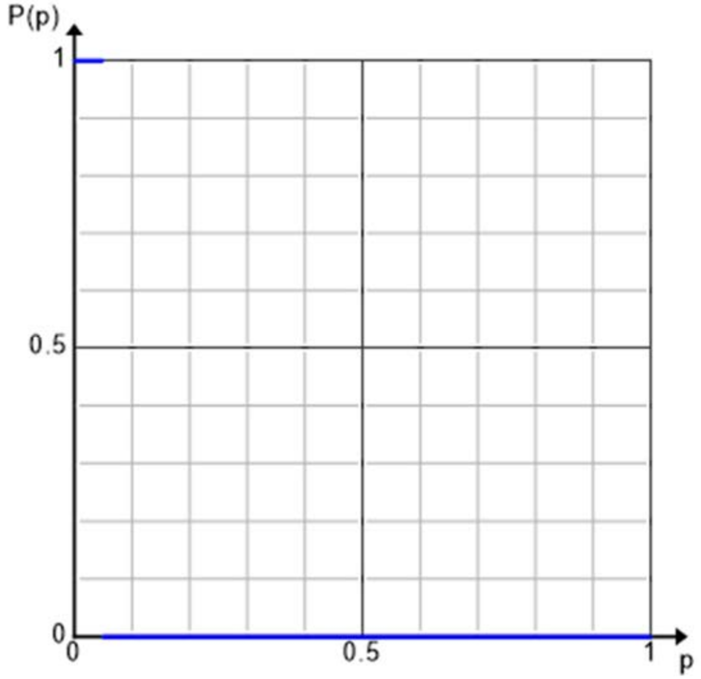

Figure 10. Classic NHST with $\alpha = 0.05$

Since both P and p are supported on the unit interval, we invoke the method of an earlier section of this paper. Consider the class of functions $P(p)$ such that the logit of $P(p)$ is a linear function of the logit of p. That is, $\text{logit}(P) = \beta_0 + \beta_1 \text{logit}(p)$, or $\ln(P/(1-P)) = \beta_0 + \beta_1 \ln(p/(1-p))$. Exponentiating both sides and solving for P yields:

$$P = \frac{e^{\beta_0}\left(\frac{p}{1-p}\right)^{\beta_1}}{1 + e^{\beta_0}\left(\frac{p}{1-p}\right)^{\beta_1}}$$

This function has domain and range equal to $(0,1)$, and is decreasing if and only if β_1 is negative. As p approaches 0 or 1 respectively, P approaches 1 or 0 respectively if β_1 is negative.

For convenience, we set β_0 to equal the logit of α. Then we can rewrite our formula for P:

$$P = \frac{\alpha\left(\frac{p}{1-p}\right)^{\beta_1}}{1-\alpha+\alpha\left(\frac{p}{1-p}\right)^{\beta_1}}$$

As β_1 approaches zero, the function $P(p)$ approaches the constant function at α. That is, if we ignore the p-value, we reject the null hypothesis with probability α; and that is consistent with our traditional understanding. Also, for any value of β_1, $P(1/2)$ is α; so, if the data seem equally consistent or inconsistent with the null hypothesis, reject the null hypothesis with probability α ; and that too is consistent with our traditional understanding.

When $\beta_1 = -1$, we observe two interesting properties: this is the only negative value of β_1 for which the graph of $P(p)$ lacks an inflection point on $(0,1)$; and, more interesting to statisticians, this is the only value of β_1 for which $P(\alpha) = 1/2$. This last observation can be read as stating that when the p-value is less than α, we lean more towards rejecting the null hypothesis, but when the p-value is more than α, we lean more towards not rejecting the null hypothesis. The narrative of the last sentence is the most consistent with mainstream practice in NHST. We depict the function $P(p)$ for $\beta_1 = -1$ and $\alpha = 0.05$ in Figure 11.

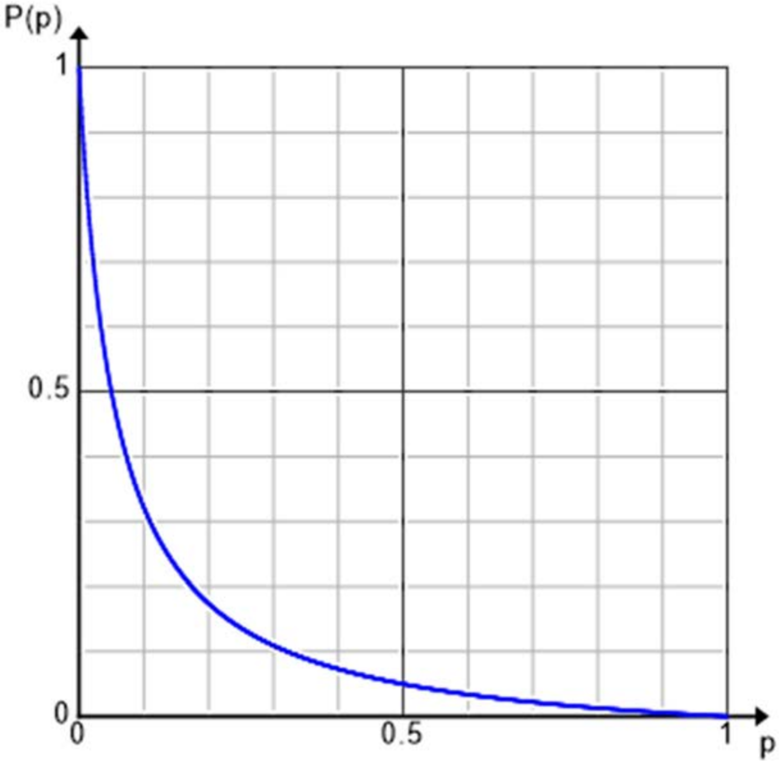

Figure 11. A continuous decision rule with $\alpha = 0.05$

For convenience, we supply the simplified version of the formula for $P(p)$, and for its complement, in the case $\beta_1 = -1$:

$$P = \frac{\alpha\,(1-p)}{(1-\alpha)\,p + \alpha\,(1-p)}$$

$$1 - P = \frac{(1-\alpha)\,p}{(1-\alpha)\,p + \alpha\,(1-p)}$$

This discussion begs the question: what does “the probability of rejecting the null hypothesis” mean? It could refer to the relative frequency of rejecting the null hypothesis in repeated, independent experiments producing similar data. More broadly, however, we can associate $1 - P$ with our degree of belief in the null hypothesis, and P with our degree of belief in the alternative hypothesis. This is not the same as “the probability that the null hypothesis is true, respectively false”. But it does capture the weight of evidence in support of, or respectively against, the null hypothesis, and thus the plausibility of the null and alternative hypotheses.

We might go further and define P to measure the “noteworthiness” of the data, as opposed to the significance, which is the p-value. In this fashion, we acknowledge that a p-value requires sample size for context, because for a fixed effect size, the p-value is already decreased by a large sample size; a p-value of 0.051 computed on a sample of size 10 is more interesting than a p-value of 0.049 computed on a sample of size 1000.

In this section, we have not addressed the issue of how α should be chosen. It was addressed, however, in the preceding section, as a function of sample size. Once an experiment is performed and a p-value computed, then the formulas above for $1 - P$ and for P respectively inform us as to the degree of plausibility we may find in the null hypothesis and in the alternative hypothesis respectively.

We illustrate these formulas in figure 12.

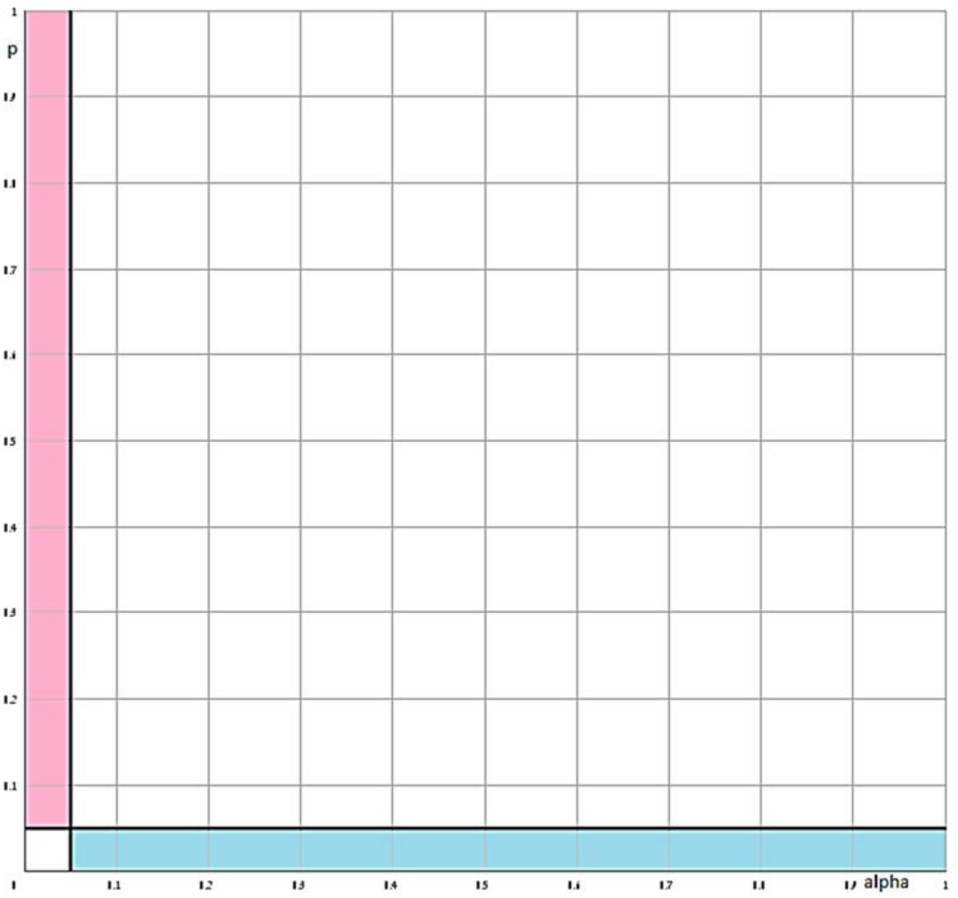

Figure 12. The p-by-α square

Figure 12 depicts the unit square, where the horizontal axis measures α and the vertical axis measures p, both varying between 0 and 1. The heavy vertical line within the square indicates the chosen value for α, and the heavy horizontal line within the square indicates the observed value for p. The area of the pink rectangle is the numerator for P; the area of the light blue rectangle is the numerator for $1 - P$; the denominator for both fractions is the sum of the areas of the two colored rectangles. Thus, the preponderance of evidence is in favor of the alternative hypothesis when the area of the pink rectangle exceeds the area of the light blue rectangle ($P > 1/2$); otherwise the preponderance of evidence is in favor of the null hypothesis.

We can even provide a quasi-Bayesian heuristic explanation of the formulas for P and $1 - P$. Bayes's rule would yield the given formulas, the first for $P(H_1|data)$ and the second for $P(H_0|data)$, provided the following interpretations were true:

- $p = P(data|H_0)$. This much, at least, is true by definition.
- $1 - \alpha = P(H_0)$ and $\alpha = P(H_1)$. This assertion cannot be accepted, since each of the two hypotheses is either true or false. Furthermore, it compromises the meaning of α, which is $P(reject\ H_0|H_0)$.
- $1 - p = P(data|H_1)$. This is the most difficult assumption, since it implies that the data can be observed only under one of the two hypotheses: and furthermore, this set of data must always be observed.

The form of the two formulas clearly invites some sort of Bayesian explanation, and we invite the reader to suggest a more acceptable explanation than that proposed above.

There has been much discussion recently about the banishment of the very concept of statistical significance itself, as well as dichotomous decision rules. The identification of the P and $1 - P$ functions might open the door to a potential compromise between traditional dichotomous rules and the alternative gaining popularity in the statistical community, which is: failing to give any guidance at all to decision makers who must make dichotomous choices. Various polychotomous rules are proposed below.

We propose a trichotomous decision rule as follows:

- if $P > 2/3$, we declare that there is substantial evidence in favor of the alternative hypothesis, and we find the alternative hypothesis plausible;
- if $1/3 < P \leq 2/3$, we declare that the evidence is ambiguous;
- if $P \leq 1/3$, we declare that there is substantial evidence against the alternative hypothesis, and we find the null hypothesis plausible.

This rule could easily be adjusted to replace the fractions above with $1/2 \pm c$, for other values of c between 0 and $1/2$.

A tetrachotomous decision rule would work as follows:

- if $P > 3/4$, we declare that there is strong evidence in favor of the alternative hypothesis, and we find the alternative hypothesis very plausible;
- if $1/2 < P \leq 3/4$, we declare that there is weak evidence in favor of the alternative hypothesis, and we find the alternative hypothesis somewhat plausible;
- if $1/4 < P \leq 1/2$, we declare that there is weak evidence against the alternative hypothesis, and we find the null hypothesis somewhat plausible;
- if $P \leq 1/4$, we declare that there is strong evidence against the alternative hypothesis, and we find the null hypothesis very plausible.

A pentachotomous decision rule would work as follows:

- if $P > 4/5$, we declare that there is strong evidence in favor of the alternative hypothesis, and we find the alternative hypothesis very plausible;
- if $3/5 < P \leq 4/5$, we declare that there is weak evidence in favor of the alternative hypothesis, and we find the alternative hypothesis somewhat plausible;
- if $2/5 < P \leq 3/5$, we declare that the evidence is ambiguous;
- if $1/5 < P \leq 2/5$, we declare that there is weak evidence against the alternative hypothesis, and we find the null hypothesis somewhat plausible;
- if $P \leq 1/5$, we declare that there is strong evidence against the alternative hypothesis, and we find the null hypothesis very plausible.

5. Summary and Recommendations

It has been our goal in this paper to sketch the outline of a program for the analysis and interpretation of data.

Given a set of observations of random variables, we begin by establishing the maximum theoretical support of all variables, with the assistance of a subject matter expert if required. If these are interval variables, we use the functions in the first section to transform support to $(-\infty, \infty)$. If these variables represent the integers within an interval, we use the transformation of the largest open interval containing those integers and only those integers, to the interval $(-\infty, \infty)$. Once the variables are supported on $(-\infty, \infty)$, we subtract the median, divide by the standard deviation, and use the functions of the first section to remove skewness and excess kurtosis. At this point, the first four moments of the transformed variables present themselves as being close to those of a standard normal random variable.

With the variables approximating normality, classical parametric methods are more easily enabled, giving us more power, and providing the optimal opportunity to distinguish signal from noise. We test for the existence of a non-trivial linear relation among specified variables in the dataset, by comparing the computed p-value to a significance threshold that varies with the sample size. We recommend the use of $\alpha = 2\left(1 - 2^{-1/n}\right)$, where n is the sample size. We note the simple lower and upper bounds $(2\ln 2)/(n+1) < 2\left(1 - 2^{-1/n}\right) < (2\ln 2)/n$, bounds that grow asymptotically close as n grows without bound.

If there are multiple tests occurring, as in the forward selection construction of a multiple regression model with k candidate predictors, we encourage the use of the Holm-Bonferroni method with the Bonferroni correction: test the most significant predictor against the significance level $2\left(1 - 2^{-1/n}\right)/k$, then the second most significant predictor against the significance level $2(1 - 2^{-1/n})/(k-1)$, and so on, until a non-significant result is found. If necessary with many predictors, regress against principal components of the predictors rather than the original data.

We refer readers to Macnaughton's survey on decision rules, where he argues that the p-value itself is best to detect effects, as well as the related paper by García-Pérez. In this paper, we have compromised between two approaches. The first is the traditional approach of comparing the p-value against a constant significance level α, no matter how small the effect size, as shown by the green line in figure 6. This runs the risk of too many false positive results. The opposite extreme consists of reporting only an effect size, no matter how small the p-value, as shown by the orange and gray curves in figure 6. This runs the risk of too many false negative results. Our compromise appears as the yellow curve in figure 6.

Once we have computed our p-value and our target significance level α, we can express our degree of belief in the null hypothesis as:

$$\frac{(1-\alpha)\,p}{(1-\alpha)\,p + \alpha\,(1-p)}$$

and our degree of belief in the alternative hypothesis, the noteworthiness, as:

$$\frac{\alpha\,(1-p)}{(1-\alpha)\,p + \alpha\,(1-p)}$$

We recommend changing statistics education to address, within the very first introductory course, the support of random variables and important probability distributions in general, as well the support of various types of quantities likely to be encountered in statistical practice. The formulas provided for support transformations, for the significance threshold as a function of sample size, and for the degree of belief in the null and alternative hypotheses as functions of the significance level and the p-value, are very easy for introductory statistics students to compute as well, and should be introduced as soon as possible.

Acknowledgments

The author wishes to express thanks for the tremendous patience of his wife, Yan, as the author demonstrated the interpretation of moments of probability distributions through statistical yoga.

References

Amrhein, Valentin and Sander Greenland, “Remove, rather than redefine, statistical significance,” *Nature Human Behaviour* (2018). DOI: 10.1038/s41562-017-0224-0, accessed October 7, 2018.

Banerjee, Subrato. “The p-value problem,” preprint.

Benjamin, D. J., J. O. Berger, … and V. E. Johnson, (2017), "Redefine Statistical Significance," *Nature Human Behaviour*. doi.org/10.1038/s41562-017-0189-z, accessed October 7, 2018.

Bohannon, John. I Fooled Millions Into Thinking Chocolate Helps Weight Loss. Here's How. https://io9.gizmodo.com/i-fooled-millions-into-thinking-chocolate-helps-weight-1707251800, accessed March 12, 2018.

García-Pérez, Michael A., “Thou shalt not bear false witness against null hypothesis significance testing,” *Educational and Psychological Measurement*, Vol. 77(4) (2017), 631-662.

Hahn, Gerald J. and Samuel S. Shapiro. *Statistical Models in Engineering*. John Wiley & Sons, Inc., 1967.

Holm, Sture, A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, Vol. 6(2) (1979), pp. 65-70.

Huber, William A. Private communication.

Ioannidis, John P. A., Why Most Published Research Findings Are False. *PLoS Med*, Vol. 2(8) (2005): e124; https://doi.org/10.1371/journal.pmed.0020124, accessed February 19, 2018.

Kleinbaum, David G., Lawrence L. Kupper, Azhar Nizam, and Keith E. Muller. *Applied Regression Analysis and Other Multivariable Methods*. Thomson Brooks/Cole, fourth edition, 2008.

Kutner, Michael H., Christopher J. Nachtsheim, John Neter, and William Li. *Applied Linear Statistical Models*. McGraw Hill, fifth edition, 2005.

Lakens, D. et al. (2018), "Justify Your Alpha," *Nature Human Behavior*, 2, 168–171. doi.org/10.1038/s41562-018-0311-x, accessed October 7, 2018.

Macnaughton, D. B., "The *p*-value is Best to Detect Effects" (2018), https://matstat.com/macnaughton2018d.pdf, accessed November 13, 2018.

Papadopoulos, Alecos. https://stats.stackexchange.com/questions/130069/what-is-the-distribution-of-r2-in-linear-regression-under-the-null-hypothesis, published 2017, accessed July 20, 2020.

Statistics of Income – 2014 Individual Income Tax Returns. Internal Revenue Service, 2015. https://www.irs.gov/pub/irs-soi/14in11si.xls, accessed February 25, 2018.

Taleb, Nassim Nicholas. *Fooled by Randomness*. Random House, second edition, 2005.

Taleb, Nassim Nicholas. *The Black Swan*. Random House, second edition, 2010.

Taleb, Nassim Nicholas. *Antifragile*. Random House, 2012.

van der Laan, Mark. Why We Need a Statistical Revolution. http://senseaboutscienceusa.org/super-learning-and-the-revolution-in-knowledge/, published 2015, accessed February 19, 2018.

Wackerly, Dennis D., William Mendenhall III, and Richard L. Scheaffer. *Mathematical Statistics with Applications*. Thomson Brooks/Cole, 2008.

Wasserstein, Ronald L. and Nicole A. Lazar, The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, Vol. 70, Iss. 2.2016; https://doi.org/10.1080/00031305.2016.1154108, accessed January 12, 2018.

Wilshire Associates, Wilshire 5000 Total Market Full Cap Index [WILL5000INDFC], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/WILL5000INDFC, accessed September 30, 2016.