Power surfaces

Mary C. Phipps*

Abstract

The link between α-level tests and p-values is stated in terms of a *power surface* and is demonstrated visually using tests based on continuous and discrete test statistics. Cross-sections of the power surface in one plane result in the usual power curves, while cross-sections in another plane provide a useful empirical tool for evaluating tests. Examples are given using Monte-Carlo simulation and bootstrap techniques.

Keywords: inference, power, p-value, bootstrap, simulation.

1 Introduction

It has become common practice to quote a *p-value* as the result of a test of significance, rather than to report a conclusion of *significance* or *non-significance* at a specified level α. Often it is necessary to perform an approximate test based on the non-parametric bootstrap of Efron (1979), and in these cases a bootstrap p-value is quoted. These bootstrap p-values are sometimes used in place of a power curve to evaluate the performance of a test, by plotting the ordered bootstrap p-values against the order statistics of the uniform

*Address: School of Mathematics and Statistics, University of Sydney, Sydney, Australia. 2006

distribution. The interpretation of these plots (see for example Fisher et al (1994) and Bennett et al (1995)) exploits the fact that the p-value has a uniform distribution under the null hypothesis. It will be shown in §2 that these q-q plots of the bootstrap p-values can also be viewed as cross-sections of a simulated *power surface*, and so can also be interpreted in terms of power.

1.1 The power surface for tests returning p-values

One gain in testing hypotheses using a p-value is that the p-value provides more information than a conclusion for an α-level test. This is because the latter is simply the binary indicator $I(\text{p-value} \in [0, \alpha])$, where $I = 1$ for significance and $I = 0$ for non-significance.

There is a similar gain connected with the power of a test. For simplicity consider null and alternative hypotheses H_0 and H_1 concerned with the value of a parameter θ, and suppose that t is the observed value of a continuous test statistic T. For a specific value θ_1 specified by H_1, the probability of observing a value of T as bad or worse than t is $G_{\theta_1}(t)$, the *significance probability.* Let us write $p = G_{\theta_0}(t)$ for the corresponding p-value. Then $t = G_{\theta_0}^{-1}(p)$, and so $G_{\theta_1}(t) = G_{\theta_1}(G_{\theta_0}^{-1}(p))$. Evaluating this at $p = \alpha$ gives the *power*, $G_{\theta_1}(G_{\theta_0}^{-1}(\alpha))$ of the α-level test.

Taking this one step further, we obtain $G_\theta(G_{\theta_0}^{-1}(\alpha))$, the *power curve* of an α-level test as a function of θ. Finally, permitting α to vary as well (rather than changing the notation to p as above) gives a *power surface* for the test as a function of α and θ. This surface contains the essence of the p-value approach for hypothesis testing.

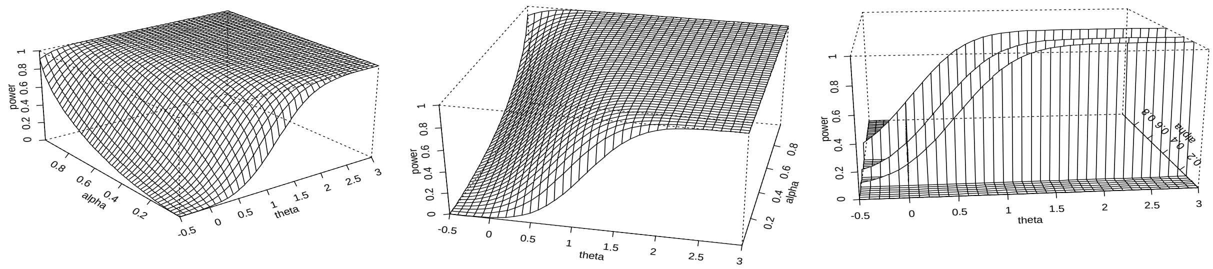
The idea extends to a discrete test statistic T, but the resulting power 'surface' is a set of disjoint power curves in this case.

1.2 Graphical presentation of power surfaces

Figs 1(a) and 1(b) show portion of a typical power surface (based on a continuous test statistic T with θ defined on an interval) from two different angles. Fig 1(c) shows a 'surface' for which the only achievable p-values (axis labelled *alpha*) are 0.1, 0.2, 0.4 and 1.0 (being based on a discrete test statistic T with θ defined on an interval).

Figure 1

(a) (b) (c)



The power surface, $P(\theta, \alpha) = G_\theta(G_{\theta_0}^{-1}(\alpha))$, is a three dimensional variant of a power curve. The vertical axis is labelled *power* since, in the plane $alpha = \alpha$, the significance probability *is* the power of the α-level test. The third axis is labelled *theta*, representing the alternative hypothesis.

Cross-sections of the power surface in the plane $alpha = \alpha$ yield the usual power curves on an α-level test, while cross-sections in the plane $theta = \theta_1$ plots $G_{\theta_1}(t)$ against $G_{\theta_0}(t)$ as t varies over the domain of T, and is called a *significance probability curve.*

1.3 A particular example

To illustrate, consider the uniformly most powerful (UMP) test of $H_0 : \theta = 0$ against the alternative $H_1 : \theta > 0$, based on a random sample $\mathbf{X}$ (with realisation $\mathbf{x}$) from the normal variable, $\mathcal{N}(\theta, 1)$. The test statistic is $T = \bar{X}$, for which $G_\theta(t) = 1 - \Phi(\sqrt{n}(t - \theta))$,[1] and

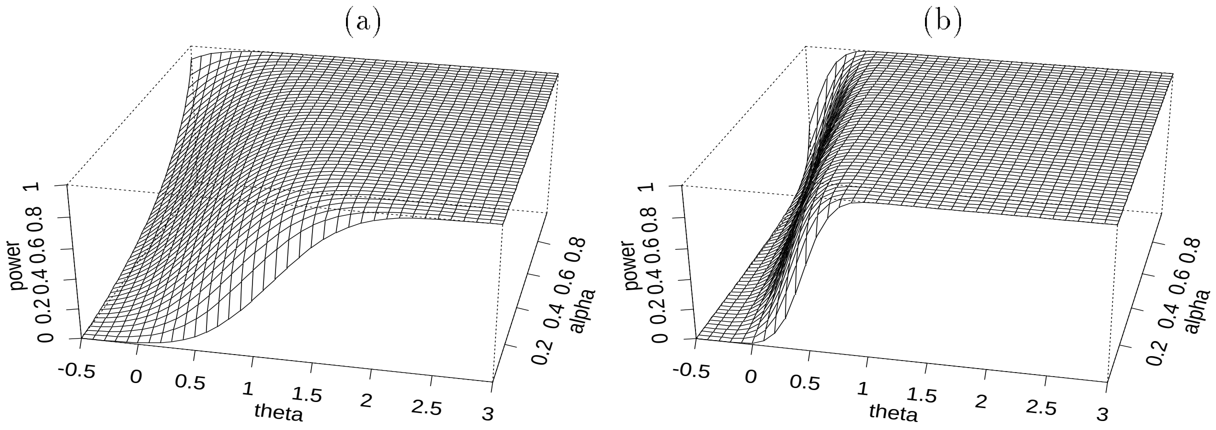
[1] $\Phi(t)$ is the distribution function of the standard normal variable

the best critical region (BCR) for the α-level test is $\{\mathbf{x} : \bar{x} \geq t_\alpha\}$, where $t_\alpha = G_0^{-1}(\alpha) = \frac{1}{\sqrt{n}}\Phi^{-1}(1-\alpha)$. The power surface is

$$P(\theta, \alpha) = G_\theta(G_0^{-1}(\alpha)) = 1 - \Phi(\sqrt{n}(\frac{1}{\sqrt{n}}\Phi^{-1}(1-\alpha) - \theta))$$

and it is clear that $P(\theta, \alpha)$ is increasing in θ holding α fixed provided $\theta > 0$. It is also clear that $P(\theta, \alpha)$ is increasing in α holding θ fixed. This property of UMP tests is noted by Hogg and Craig (1978), who make the remark that "the word *uniformly* is associated with θ; $\cdots$ " adding that "there is a certain uniformity property, also associated with α, that is not always noted in statistics texts". This property is demonstrated for this particular UMP test in Figs 2(a) and 2(b) which show the power surfaces when the sample sizes are $n = 4$ and $n = 25$ respectively. Other properties are summarised below.

Figure 2



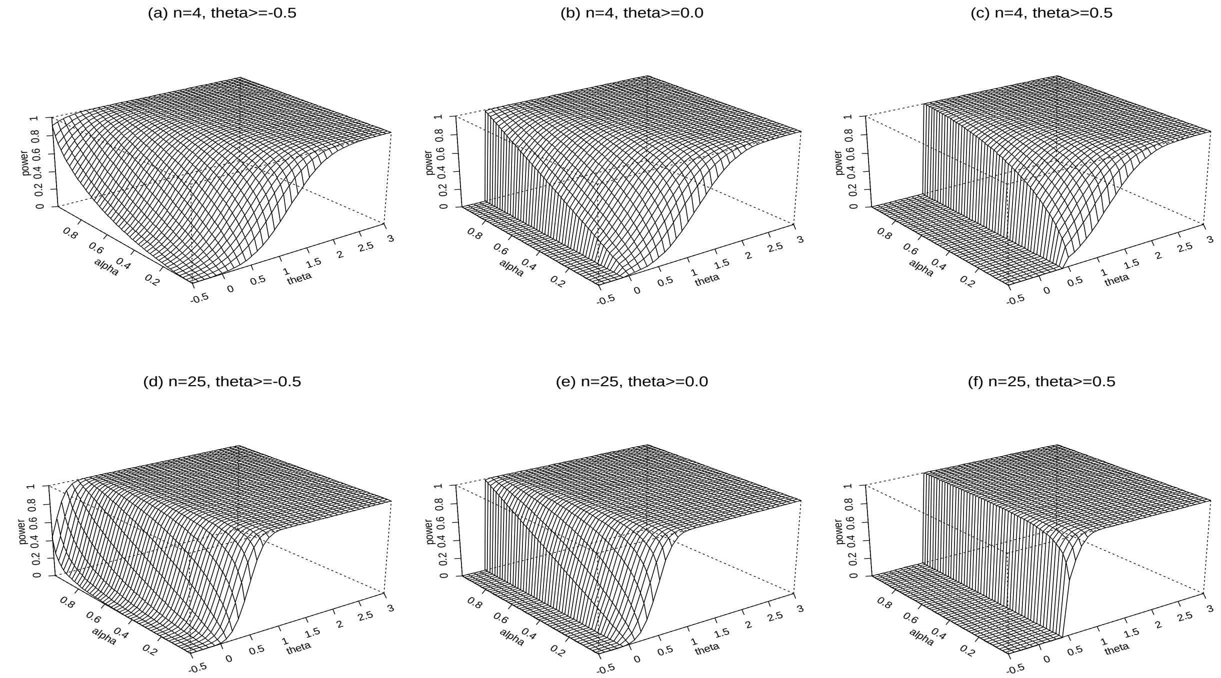
- The surfaces are increasing functions of both θ and α, for all alternatives in H_1,

- the surface of Fig 2(a) lies entirely below the surface of Fig 2(b) for $\theta > 0$ but entirely above for $\theta < 0$. This is because the UMP test based on $n = 25$ observations is more powerful than that based on $n = 4$ observations for alternatives under H_1.

- The usual *power curve* at a specified value of α is simply the cross-section of the surface in the plane at $alpha = \alpha$.

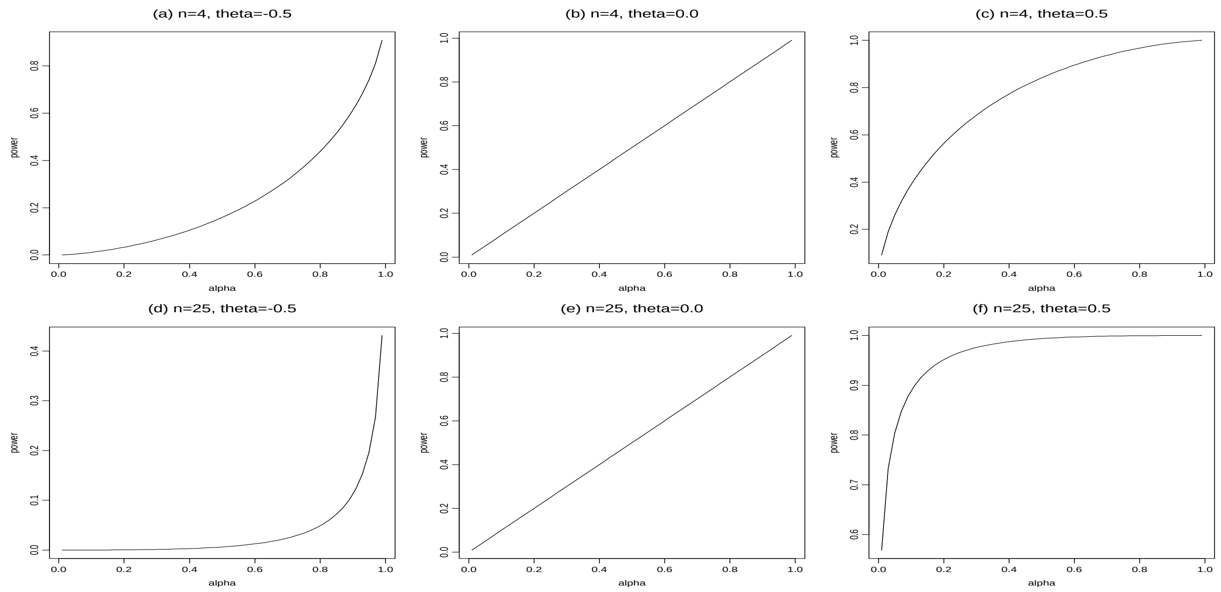
- Taking cross-sections of the power surface in a plane with θ fixed, gives another comparison of the tests which is not restricted to fixed α. Cross-sections of the power surface of Fig 2(a) in the planes $\theta = -0.5$, 0 and 0.5 are shown in Figs 3(a), 3(b) and 3(c) respectively, and repeated for Fig 2(b) in Figs 3(d), 3(e) and 3(f). These curves plot $G_{\theta_1}(t)$ against $G_{\theta_0}(t)$ for fixed θ_1 as t varies, where $\theta_0 = 0$.

Figure 3



These *significance probability curves* are displayed again in two dimensions in Figs 4(a) to 4(f), but note that the surface is viewed from behind.

Figure 4



It is clear that good tests are those for which the significance probability $G_{\theta_1}(t)$ (axis labelled *power*) is large compared with $G_{\theta_0}(t)$ (axis labelled *alpha*). Comparisons for this particular example are easily made using Fig 4.

- Cross-sections of power surfaces like those in Fig 1(c) in the plane $\theta = \theta_1$, result in points rather than a curve, but as in the discussion above, the position of these points relative to the diagonal indicates whether or not the test is good.

- Even when the alternatives are not simply changes in the value of a scalar parameter, cross-sections by *theta* have the same interpretation as above.

2 Comparing tests empirically

It is still possible to judge tests when the sampling distribution of T is not known, by producing cross-sections of the power surface empirically. Simulating a power curve (cross-section for a fixed α) requires simulation over a range of values of θ. However, if interest is in the performance of the test at a specific value of θ, it may be more informative to produce a simulated significance probability curve (cross-section for that fixed θ). This can be done even when a power curve may be inappropriate, for example when the alternative is not simply a change in the value of a scalar parameter. These cross-sections may be simulated using either Monte-Carlo or bootstrap techniques, and the method of construction depends on which of these techniques is used.

2.1 Monte-Carlo techniques

Monte-Carlo simulation techniques can be used when the sample $\mathbf{X}$ is a random sample of a variable X with *known density* for every θ. Suppose for illustration that the data $\mathbf{x}$ are to be used to test $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ and that the BCR is of the form

$\{\mathbf{x} : t \geq c\}$ where $t = T(\mathbf{x})$ is an observed value of the test statistic $T = T(\mathbf{X})$. Writing $P(T \geq t) = G_\theta(t)$, the significance probability curve at $\theta = \theta_1$ is $G_{\theta_1}(t)$ against $G_{\theta_0}(t)$, as t varies, or equivalently as α varies.

If G_θ is known for all θ, it is not necessary to use empirical methods, but if it is not known, the curve is simulated by generating N random samples from the known density of X with $\theta = \theta_1$. For each sample calculate the sample values $t_1, \ldots, t_N$ of T. Repeat for $\theta = \theta_0$, calculating the sample values $t_1^0, \ldots t_N^0$.

- If T is discrete, estimate $G_{\theta_1}(t)$ by $\hat{G}_{\theta_1}(t) = \frac{\sum_{j=1}^{N} I(t_j \geq t)}{N}$, and $G_{\theta_0}(t)$ by $\hat{G}_{\theta_0}(t) = \frac{\sum_{j=1}^{N} I(t_j^0 \geq t)}{N}$ for achievable values of t, plotting the points $(\hat{G}_{\theta_0}(t), \hat{G}_{\theta_1}(t))$.

- If T is continuous, $G_{\theta_0}(T)$ is uniformly distributed since

$$P_{\theta_0}(G_{\theta_0}(T) \leq \alpha) = G_{\theta_0} G_{\theta_0}^{-1}(\alpha) = \alpha.$$

  Estimate $G_{\theta_1}(t)$ as above at $t = t_1^0, \ldots t_N^0$, and produce a 'uniform' q-q plot as usual.

  Of course, if the *null* sampling distribution G_{θ_0} is *known*, simulation under H_0 is not necessary. Instead, estimate $G_{\theta_1}(t)$ at $t = G_{\theta_0}^{-1}(\alpha_i)$, where $\alpha_1, \ldots, \alpha_N$ are evenly spaced over the interval (0,1). It is usual to choose either $\frac{i}{N+1}$ or $\frac{i-0.5}{N}$ for α_i.

2.2 Bootstrap techniques

Bootstrap techniques can be used to produce a bootstrap p-value for a sample of size n when the parent variable X has *unknown density*. Briefly, the sample yields an observed value t of the test statistic T, and the sample itself is used to provide an empirical distribution from which B bootstrap samples each of size n are obtained. These provide B bootstrap replicates of the test statistic T under the *null hypothesis*, and the bootstrap p-value is estimated as a proportion in the usual way.

2.2.1 One method of evaluating tests with bootstrap p-values

If a test is to be evaluated in the light of a particular alternative, N *simulated* bootstrap p-values can be found by generating N samples each of size n under the alternative of interest. Suppose that the ordered bootstrap p-values are $p^*_{(1)}, p^*_{(2)}, \ldots, p^*_{(N)}$ One approach in use is to produce a *uniform* q-q plot of the p-values. In this way, the distribution of p-values is compared with the distribution expected if the samples had been generated instead under the *null hypothesis*. In the latter case, the p-values would be uniformly distributed.

This approach is particularly useful if the null and alternative hypotheses are not simple, as for example in Fisher et al (1995), who use a variant of Silverman's (1981) procedure for investigating multimodality to test for unimodality against the general alternative of a mixture of distributions. Taking a specific mixture of distributions, N samples of size n are generated, and for each a bootstrap p-value $\hat{G}_{\theta_0}(t)$ is estimated, not as is the case in Monte-Carlo simulation, a significance probability $\hat{G}_{\theta_1}(t)$. The procedure is repeated for different alternatives. To judge the test, Fisher et al (1994) produce a 'uniform' q-q plot by plotting the points $(x_i, y_i) = (\frac{i-0.5}{N}, p^*_{(i)})$ $i = 1, \ldots, N$. (Note that $\frac{i}{N+1}$ could also have been used in place of $\frac{i-0.5}{N}$.) They call this plot the *P-P plot of significance probabilites* and compare it with the line $y = x$ as follows:

> " A test has size α when the curve is above the 45° line $y = x$ on $[0,\alpha]$, for each f (density) in H_0. A conservative test is above this line in that region. An "ideal test" would give a curve close to the 45° line for f at the boundary between H_0 and H_1, and be as far below the 45° line as possible for f in the interior of H_1. A test which is near the 45° line when f is in the interior of H_1 lacks power."

2.2.2 The power surface and bootstrap p-values

There is another interpretation of this 'uniform' q-q plot related to the power surface. Recall that when using Monte-Carlo techniques to simulate the cross-section of a power surface in the plane $theta = \theta_1$, the estimated significance probabilities $\hat{G}_{\theta_1}(t)$ were plotted against $G_{\theta_0}(t)$. In effect, this was a 'uniform' q-q plot of significance probabilities, for continuous T, since, under the null hypothesis $G_{\theta_0}(T)$ is uniformly distributed.

With bootstrap techniques, even though the samples are generated under a specific alternative hypothesis $\theta = \theta_1$, it is the p-values $G_{\theta_0}(t)$ which are estimated *not* $G_{\theta_1}(t)$. As above, assume the ordered sample bootstrap p-values are $p^*_{(1)}, p^*_{(2)}, \ldots, p^*_{(N)}$. For the specific alternative $\theta = \theta_1$, $G_{\theta_1}(T)$ is uniformly distributed (when T is continuous) since

$$P_{\theta_1}(G_{\theta_1}(T) \le t) = G_{\theta_1} G^{-1}_{\theta_1}(t) = t.$$

Thus, the expected ith order statistic of $G_{\theta_1}(T)$ is $\frac{i}{N+1}$. The corresponding ith sample order statistic of $\hat{G}_{\theta_0}(T)$ is $p^*_{(i)}$.

To simulate the cross-section of the power surface in the plane $theta = \theta_1$, plot the points $(x_i, y_i) = (p^*_{(i)}, \frac{i}{N+1})$ for $i = 1, \ldots, N$. This is obviously a 'uniform' q-q plot of bootstrap p-values with the axes reversed. (For discrete T, the adjustment is similar to that used for Monte Carlo simulation.)

The standard interpretation of this plot compares the bootstrap p-values with those expected for samples generated under the *null hypothesis*. The plot can now be viewed also as the cross-section of a power surface in the plane specified by the *alternative hypothesis*.

References

1. Bennett, M.R., Robinson, J., Phipps, M.C., Karunanithi, S., Lin, Y.Q. & Cottee, L. (1995) Quantal components of spontaneous excitatory junction potentials at

vusualised varicosities. *Journal of the Autonomic Nervous System* (in press).

2. Efron B. (1979) Bootstrap methods: Another look at the jackknife. *Annals of Statistics,* 7: 1-26.

3. Fisher, N.I., Mammen, E. and Marron, I. (1994) Testing for multimodality. *Computational Statistics and Data Analysis,* 18: 499-512.

4. Hogg, R.V. and Craig, A.T. Introduction to Mathematical Statistics 4th Edition (1978). Prentice Hall, New Jersey.

5. Silverman, B.W. (1981) Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society. B Series,* 43: 97-99.