# Tests of the Mean with Distributional Uncertainty: An Info-Gap Approach

**Yakov Ben-Haim**
Yitzhak Moda'i Chair in Technology and Economics
Technion — Israel Institute of Technology
Haifa 32000 Israel

## Abstract

Statistical tests of the mean are quite common. Sometimes the analyst cannot validate the assumptions underlying the test, such as normality, symmetry, independence of measurements, etc. This causes unknown deviation of the actual sampling distribution from the distribution assumed by the test, and thus unknown size and power of the test. This distributional uncertainty makes it difficult to reliably choose the decision threshold (critical value) and sample size. We present a method for evaluating the robustness of a test to an unknown degree of distributional uncertainty, based on info-gap decision theory. Analysis of robustness is useful in evaluating effective size and power, and for selecting the decision threshold and sample-size. We study binary simple-hypothesis tests of the mean and consider both type I and type II errors. We show quantitatively that robustness to distributional uncertainty improves, at fixed nominal level of significance, as the effective level of significance deteriorates. Likewise, robustness improves as the effective power of the test deteriorates. Furthermore, we show how to choose the decision threshold and sample size in light of distributional uncertainty. We illustrate our results by application to the $t$ test and to a test of false nulls in epidemiology.

**Keywords:** binary hypothesis tests, distributional uncertainty, info-gaps, robustness, tests of the mean, $t$ test, chronic wasting disease, false nulls.

## 1 Introduction

Statistical tests of the mean value of a population property are exceedingly common, and numerous tests are available. These tests depend on various assumptions about the data and the population. Sometimes normality is assumed, and almost invariably random sampling is posited: the measurements are made independently but with the same instrument and from the same population which is unaffected by the sample. However, in many situations the data generating process is not normal, or the sample is not random: the measurement instrument is not constant, or the sample is biased, or the measurements influence one another to some extent, or the statistical character of the population which is sampled is not constant. Determination of the level of significance and power of the test, and selection of the sample size, depend on the test which is used and its underlying assumptions. In some situations the analyst is unable to characterize the violation of test assumptions and is thus unable to adjust the test accordingly, and unable to reliably evaluate the level of significance and power or choose a sample size. We present a method for dealing with such situations.

Violation of the test assumptions can result in deviation of the actual sampling distribution from the distribution upon which the test is based. In situations where the violations are poorly known, the distributional deviations are similarly uncertain. We will refer to this as *distributional uncertainty.*

Considerable effort has been devoted to deriving methods which are robust to distributional uncertainty. Careful test design is a major antidote, though not always adequate. In some cases the distributional uncertainty can be characterized as a mixture of several (or many) distributions of known structure. Given adequate data, methods exist for estimating the parameters of the distributions and their weights in the mixture (Titterington *et al.* 1985). In other situations Monte Carlo methods are used to construct a sampling distribution based on prior knowledge of the distributional complexity (Robert, 2004). In these situations one can evaluate the robustness of a test as the extent of difference between simulated type I and type II error rates and the theoretical error rates in the absence of distributional uncertainty. Non-parametric methods exist which avoid

or weaken some assumptions about the sampled distribution. These tests do posit random sampling, or identity of two distinct distributions, or other assumptions (Johnson, 1995), and some are strictly valid only asymptotically. Numerical methods are available for evaluating the robustness of non-parametric statistics to specific violations, such as small-sample applications. However, non-parametric statistics can be very sensitive to a small number of outlying measurements. This focusses attention on the problem of long tails of the sampled distribution. The jacknife technique (Mooney and Duval, 1993), or trimmed means (DeGroot, 1986), attempt to rectify the effects of outliers. More generally, robustness can be evaluated as insensitivity to small deviations from the distributional assumptions (Huber, 1981), leading to $M$ estimates and other techniques.

The distributional uncertainty on which we focus here is more unstructured than that for which these methods are explicitly designed. We illustrate our concept of distributional uncertainty, and its origin in ecological assessment and epidemiology, in section 2. Briefly, however, we consider situations in which the sampling distribution is highly uncertain and may be skewed, heavy tailed, multi-modal or non-random in ways which are unknown to the analyst. Distributional uncertainty, in the sense which concerns us here, arises for example in the use of historical data from diverse and unknown sources, taken with a variety of protocols (or lack of protocols in any professional sense), sampled from different and varying populations whose identity is imperfectly known. In such situations one must account for enormous and highly unstructured variability of the sampling distribution.

This sort of distributional uncertainty cannot be handled by the analysis of compound hypotheses. Distributional uncertainty presents us with an unbounded infinity of possible distributions—hypotheses—so it would seem impossible to formulate a compound hypothesis, or to identify a mixture of distributions.

We study two sets of problems. First, in the face of severe distributional uncertainty, what level of significance and power can one reliably ascribe to a binary simple-hypothesis test of the mean? We develop a method for quantitatively evaluating the reduction in level of significance and power, as distributional uncertainty increases. This analysis supports judgments about the effective level of significance and power, as expressed by their robustness to distributional uncertainty. Second, we show how to choose the decision threshold (critical value) and sample size when facing distributional uncertainty.

Our analysis employs info-gap decision theory for evaluating the robustness to large and highly unstructured uncertainty in the sampling distribution. We illustrate our results with simple $t$ tests of the mean, but the methodology is applicable to a broad range of statistical tests.

We begin, in section 2, with an intuitive discussion of the origin and nature of distributional uncertainty. Section 3 formulates the binary hypothesis test which we study. Section 4 presents an info-gap model for distributional uncertainty. Section 5 formulates the info-gap robustness functions for type I and type II errors. A numerical example illustrating the decisions and judgments which the analyst must make is presented in section 6. A concluding discussion appears in section 8.

## 2 Origins of Distributional Uncertainty in Ecology and Epidemiology

Recall that by 'distributional uncertainty' we mean uncertainty in the form of the sampling distribution which results from unknown violations of assumptions underlying a statistical test. Distributional uncertainty is not uncommon in ecological assessment, arising from violations of test assumptions which the analyst is unable to characterize.

The main antidote to violation of test-assumptions is of course careful test design. This typically requires good basic understanding of the processes which are studied. However, measurements are sometimes made for the very purpose of augmenting our (sometimes quite deficient) understanding of these processes. For instance, Boone and Krohn (1999) show that the accuracy of model-based predictions of occurrence of avian species is a function of the frequency of species occurrences; not surprisingly, rare species are more difficult to model accurately than common species. Similarly, Craft *et al.* (1999) study the rate of restoration of ecological attributes in artificially constructed marshes as compared to natural marshes, noting that there are no long-term comparative studies. If the factors which influence long-term restoration and growth are incompletely understood, it may be difficult to characterize the relevant statistical properties of the control and test sites and to verify that they are equivalent. Finally, it is sometimes necessary to use very small samples, such as when data are based on large-scale natural experiments (Carpenter, 1989). Tests based on phenomena which are rare and poorly understood, or newly identified and unstudied, are vulnerable to distributional uncertainty.

There are also other potential causes of distributional uncertainty. Franklin (1999) uses a range of obser-

vational data from many different sources over the past 150 years—of varying accuracy and reliability—to evaluate change in bird assemblages in northern Australia. Some of these sources were trained biologists, though professional protocols changed over the sampling period. Some observers were casual or untrained observers who may exert less effort, and thus miss the rare events, or who are enthusiastic in the search for rare occurrences and may systematically over-report extreme observations. While historical observational data are an important and valuable source, it is difficult to verify that test-assumptions are not violated.

McCarthy (1998) uses museum collections to evaluate trends in marsupials and monotremes, recognizing that variable collection efforts introduce uncertainties. Similarly, Burgman *et al.* (1995) recognize that "collection frequencies will reflect changing trends in museum and herbarium collections", which introduces uncertainties in evaluating extinction threats based on historical development of collections. Stewart-Oaten *et al.* (1992) study tests of changes of a mean population property, before and after an impact, where the impact cannot be replicated (e.g., construction of a power plant). They note that data from such measurements "do not necessarily satisfy" the assumptions of standard tests. They state that "there is no panacea" for violation of test assumptions, and if the assumptions "are seriously wrong, alternative analyses are needed. This will often require a long time series of data." These authors discuss many sources of violation of test assumptions, stressing the importance of unknown skewness of distributions or correlations among measurements.

Evidence for violation of test assumptions is not uncommon in epidemiological studies. Bausch *et al.* (2003) report non-normal distributions of large samples, and non-random selection of participants, with disproportionate participation of particular subpopulations, due perhaps to the fear of stigma.

In short, analysts not infrequently face considerable uncertainty about the actual sampling distribution of their data. There surely is a true sampling distribution from which the data were obtained, but this distribution is unknown, and unknowable on the basis of available information. On the other hand, there is undoubtedly a population property—such as a mean—which is reflected in some way in the data. It is the aim of the statistical test to discriminate something about this population property, and to assess the confidence of this discrimination. A conventional statistical approach would be to transform the pdf, or modify the test, for formulate a compound or mix-distribution hypothesis, to accommodate violations of specific assumptions. We cannot do that because we don't know what specific violations have occurred. That's precisely the distributional uncertainty which we are studying.

## 3 Binary-Hypothesis Test

We have a set of measurements $X = \{x_1, \ldots, x_n\}$ which do not necessarily constitute a random sample of any known distribution, as discussed in sections 1 and 2. These data reflect a population mean, $\mu$, but they suffer from an unknown degree of distributional uncertainty. We wish to use this data to decide between two simple hypotheses:

$$H_0: \quad \mu = T_0 \tag{1}$$
$$H_1: \quad \mu = T_1 \tag{2}$$

where each $T_i$ is a specified number, and $T_1 > T_0$.

Let $y$ be a statistic, for instance the $t$ statistic, and let $F_i(y)$ denote the cumulative distribution function (cdf) of $y$ under $H_i$. For any distribution $F(y)$, let $q_\alpha(F)$ denote the $(1-\alpha)$th quantile of $F(y)$:

$$q_\alpha(F) = \inf\{y: \ F(y) \geq 1-\alpha\} \tag{3}$$

We reject $H_0$ with significance $\alpha$ if:

$$y \geq q_\alpha(F_0) \tag{4}$$

The size, $\alpha$, is the probability of falsely rejecting the null hypothesis, $H_0$, and the power, $1-\beta$, is the probability of correctly rejecting $H_0$. $\beta$ is the probability of falsely rejecting $H_1$. If the cdf's are continuous at $q_\alpha(F)$ then the size $\alpha$, and power, $1-\beta$, are:

$$1-\alpha = F_0[q_\alpha(F_0)] \tag{5}$$
$$\beta = F_1[q_\alpha(F_0)] \tag{6}$$

## 4 Info-Gap Models for Distributional Uncertainty

Suppose that the data $X$ are not believed to be a random sample, or other assumptions underlying the test which is to be used are violated, but the nature of the violation is not known. In other words, suppose that the data are subject to distributional uncertainty. Let $y$ be the test statistic (perhaps the $t$ statistic, but not necessarily), and let $\widetilde{F}_i(y)$ denote the best (or perhaps only) guess of the distribution of the test statistic $y$, under hypothesis $H_i$. For instance, our best guess might be that $\widetilde{F}_0(y)$ is the $t$ distribution with $n-1$ degrees of freedom for the regular $t$ statistic $y = (\overline{x} - T_0)/(s/\sqrt{n})$ with sample mean and variance

$\overline{x}$ and $s^2$, while $\widetilde{F}_1(y) = \widetilde{F}_0(y-\delta)$ where $\delta = (T_1 - T_0)/(s/\sqrt{n})$.

$\widetilde{F}_i(y)$ is our best guess of the pdf of $y$ but we don't know how wrong this guess is, and we have no "worst case" estimate. This distributional uncertainty in $y$, under hypothesis $H_i$, is represented by an info-gap model, $\mathcal{U}_i(h)$, which is an unbounded family of cdf's centered on $\widetilde{F}_i(y)$. For example, the uniform-bound info-gap model for uncertainty in the cdf of $y$ is:

$$\mathcal{U}_i(h) = \Big\{ F(y): \quad F(y) \in \mathcal{P},\ |F(y) - \widetilde{F}_i(y)| \le h, \quad \forall\ y \Big\}, \quad h \ge 0 \tag{7}$$

where $\mathcal{P}$ is the set of all normalized non-negative cdf's. The info-gap model is an unbounded family of nested sets, $\mathcal{U}_i(h)$, of cdf's. In the absence of uncertainty, that is, when $h = 0$, the set is a singleton containing only the estimated cdf:

$$\mathcal{U}_i(0) = \{\widetilde{F}_i\} \tag{8}$$

The sets become more inclusive as the horizon of uncertainty increases:

$$h < h' \quad \text{implies} \quad \mathcal{U}_i(h) \subseteq \mathcal{U}_i(h') \tag{9}$$

The horizon of uncertainty, $h$, is unknown, so there is no known worst case or largest set of cdf's other than the set of all mathematically allowed cdf's (which occurs for $h \ge 1$). Eqs.(8) and (9) are the "contraction" and "nesting" axioms, respectively.

The uniform-bound info-gap model of eq.(7) entails enormous uncertainty in the cdf's. For sufficiently large $h$, the set $\mathcal{U}_i(h)$ contains densities which are highly asymmetric, multi-modal, with heavy or light tails, and with bumps, dimples, or "atoms" (infinite probability density at a single value of $y$) arbitrarily far from the mean resulting in arbitrarily large moments. Most importantly, the uncertainty in the cdf's which is represented by an info-gap model such as eq.(7) is different from estimation error or deviation from an asymptotic form. The info-gap model represents distributional uncertainty arising from unknown and possibly serious violation of fundamental assumptions underlying the hypothesis test. We do not motivate the structure of the info-gap model from consideration of estimation analytics or convergence (as in the Berry-Esseen inequality, Feller, 1971). Rather, the family of sets in eq.(7) reflects distributional uncertainty.

Other forms of info-gap model can be used if further information is available to constrain the relevant cdf's (Ben-Haim, 2006). For instance, one might have information indicating that the error of the estimated cdf is localized, e.g. on the tails, so the inequality in eq.(7) is modified in the envelope-bound info-gap model:

$$\mathcal{U}_i(h) = \Big\{ F(y): \ F(y) \in \mathcal{P},\ |F(y) - \widetilde{F}_i(y)| \le h\psi(y), \quad \forall\ y \Big\}, \quad h \ge 0 \tag{10}$$

where $\psi(y)$ is a known function. A related class of info-gap models is treated by Fox *et al.* (2007).

Alternatively one might make the judgment that probability atoms do not occur, but that the distribution may have bumps or dimples, or the tails may be heavy or light in unknown ways. An info-gap model which represents this is the fractional-error model applied to the probability density function (pdf) rather than to the cdf:

$$\mathcal{U}_i(h) = \Big\{ f(y): \ f(y) \in \mathcal{D},\ |f(y) - \widetilde{f}_i(y)| \le h f_i^\star, \quad \forall\ y \Big\}, \quad h \ge 0 \tag{11}$$

where $\mathcal{D}$ is the set of all normalized non-negative pdf's and $f_i^\star$ is a normalization constant with units of probability density. For instance, one might choose $f_i^\star = \max_y \widetilde{f}_i(y)$, which is the value of the pdf at its mode.

A much more restrictive info-gap model than eq.(11) is:

$$\mathcal{U}_i(h) = \Big\{ f(y): \ f(y) \in \mathcal{D},\ |f(y) - \widetilde{f}_i(y)| \le h \widetilde{f}_i(y), \quad \forall\ y \Big\}, \quad h \ge 0 \tag{12}$$

To understand the difference between the uncertainty models in eqs.(11) and (12), consider the case where $y$ varies from $-\infty$ to $+\infty$ and the estimated pdf, $\widetilde{f}_i(y)$, has tails which diminish asymptotically to zero. The uncertainty set $\mathcal{U}_i(h)$ in eq.(11) allows bumps as large as $hf_i^\star$ arbitrarily far out on the tail. This is not the case for the set $\mathcal{U}_i(h)$ in eq.(12) for which a bump cannot be larger than $h\widetilde{f}_i(y)$ which will become very small for large $y$. The info-gap model of eq.(11) allows much more deviant tails than the info-gap model of eq.(12).

## 5 Robustnesses for Type I and Type II Errors: Formulation

Consider a test of size $\alpha^\star$, namely, a test which rejects $H_0$ when:

$$y \ge q_{\alpha^\star}(\widetilde{F}_0) \tag{13}$$

$\alpha^\star$ is the "nominal" size of the test since it is based on the best-estimate of the cdf under $H_0$, $\widetilde{F}_0$.

We now define the robustness of this test with respect to distributional uncertainty in $\widetilde{F}_0$, for falsely rejecting $H_0$. The robustness is the maximum horizon of uncertainty, $h$, up to which the test at nominal size $\alpha^\star$ falsely rejects $H_0$ with probability no greater than $\alpha$:

$$\widehat{h}_0(\alpha^\star, \alpha) = \max\left\{h : \left(\min_{F\in\mathcal{U}_0(h)} F[q_{\alpha^\star}(\widetilde{F}_0)]\right) \geq 1-\alpha\right\} \tag{14}$$

We use the quantile $q_{\alpha^\star}(\widetilde{F}_0)$ because the test is implemented with the quantile of the best-guess distribution under $H_0$, $\widetilde{F}_0(y)$, and is of nominal size $\alpha^\star$, while the actual size (probability of falsely rejecting $H_0$) is then determined by the unknown true distribution under $H_0$, $F(y)$, which is info-gap-uncertain.

$\widehat{h}_0(\alpha^\star, \alpha)$ is related to type I error (falsely rejecting $H_0$). Specifically, $\widehat{h}_0(\alpha^\star, \alpha)$ is the greatest horizon of uncertainty up to which the probability of type I error is no greater than $\alpha$. The following expression for $\widehat{h}_0(\alpha^\star, \alpha)$, for the info-gap model in eq.(7), is derived in appendix A:

$$\widehat{h}_0(\alpha^\star, \alpha) = \alpha - \alpha^\star \tag{15}$$

or zero if this is negative. We refer to $\alpha$ as the effective size, while $\alpha^\star$ is the nominal size. Section 6 explains how the analyst evaluates and chooses the effective size.

Note that, for any choice of $\alpha^\star$, the robustness curve for type-I error, $\widehat{h}_0(\alpha^\star, \alpha)$ vs. $\alpha$, is entirely independent of the form of the test statistic. The implementation of the test, eq.(13), does depend on the type of test, through the value of the quantile $q_{\alpha^\star}(\widetilde{F}_0)$.

We now define a different robustness, related to type II error (falsely accepting $H_0$). $\widehat{h}_1(\alpha^\star, \beta)$ is the greatest horizon of uncertainty up to which the probability of falsely accepting $H_0$, with a test of nominal size $\alpha^\star$, is no greater than $\beta$:

$$\widehat{h}_1(\alpha^\star, \beta) = \max\left\{h : \left(\max_{F\in\mathcal{U}_1(h)} F[q_{\alpha^\star}(\widetilde{F}_0)]\right) \leq \beta\right\} \tag{16}$$

Let $1-\beta^\star$ be the nominal power:

$$1-\beta^\star = 1 - \widetilde{F}_1[q_{\alpha^\star}(\widetilde{F}_0)] \tag{17}$$

The following expression for $\widehat{h}_1(\alpha^\star, \beta)$, for the info-gap model in eq.(7), is derived in appendix B:

$$\widehat{h}_1(\alpha^\star, \beta) = 1 - \beta^\star - (1-\beta) \tag{18}$$

or zero if this is negative. We refer to $1-\beta$ as the effective power, while $1-\beta^\star$ is the nominal power. Section 6 explains how the analyst evaluates and chooses the effective power.

Note that, for any choice of $\alpha^\star$, the robustness curve for type-II error, $\widehat{h}_1(\alpha^\star, \beta)$ vs. $\beta$, depends on the form of the test, unlike for the type-I robustness. This is because the value of $\beta^\star$ depends on $\alpha^\star$ through the cdf's of the test statistic, $\widetilde{F}_0$ and $\widetilde{F}_1$.

## 6 Decisions and Judgments

The analyst must make two decisions and two judgments. The analyst must *decide* on the nominal test size $\alpha^\star$ and the sample size $n$. Together these decisions determine the decision threshold $q_{\alpha^\star}(\widetilde{F}_0)$ in eq.(13). Also, the analyst must *judge* what are reliable and acceptable values of the effective size $\alpha$ and effective power $1-\beta$ by considering the robustness functions $\widehat{h}_0(\alpha^\star, \alpha)$ and $\widehat{h}_1(\alpha^\star, \beta)$. (Recall that $\alpha$ is the probability of falsely rejecting $H_0$, while $1-\beta$ is the probability of correctly rejecting $H_0$.)

We will illustrate these decisions and judgments with an example employing the $t$ test. The test statistic, $y$, is $(\overline{x} - T_0)(s/\sqrt{n})$ where $\overline{x}$ is the sample mean, $s^2$ is the sample variance, and $n$ is the sample size. The estimated distribution under $H_0$, $\widetilde{F}_0(y)$, is the cdf of the $t$ statistic with $n-1$ degrees of freedom. The estimated distribution under $H_1$ is $\widetilde{F}_1(y) = \widetilde{F}_0(y-\delta)$ where $\delta = (T_1 - T_0)/(s/\sqrt{n})$. The true distributions under $H_0$ and $H_1$ are unknown and the uncertainty in each cdf is represented by the info-gap model in eq.(7).

| $\alpha^\star = 0.01$ | | $\alpha^\star = 0.05$ | |
|---|---|---|---|
| $n$ | $1-\beta^\star$ | $n$ | $1-\beta^\star$ |
| 5 | 0.1027 | 3 | 0.1784 |
| 7 | 0.3185 | 4 | 0.3736 |
| 9 | 0.5400 | 5 | 0.5390 |
| 12 | 0.7644 | 7 | 0.7457 |
| 31 | 0.9980 | 31 | 0.9997 |

Table 1: Size and power in the absence of distributional uncertainty.

The need for these judgments disappears in the absence of distributional uncertainty, since $\alpha^\star$ is the actual size and the actual power, $1-\beta^\star$, is entirely determined by $\alpha^\star$ and $n$. Values of $\alpha^\star$ and $1-\beta^\star$ are shown in table 1 for several sample sizes. Power, $1-\beta^\star$, improves (gets larger) as level of significance $\alpha^\star$ gets worse (gets larger) at fixed sample size $n$. Likewise, $1-\beta^\star$ improves as $n$ increases at fixed $\alpha^\star$.

However, the presence of distributional uncertainty makes it necessary to form judgments on effective size $\alpha$ and power $1-\beta$. These judgments are based on the robustness functions, plots of which appear in figs. 1 and 2: $\widehat{h}_0(\alpha^\star, \alpha)$ vs. $\alpha$ (positive slope) and $\widehat{h}_1(\alpha^\star, \beta)$

vs. $1-\beta$ (negative slope).

Consider first the robustness curve for type-I error, $\widehat{h}_0(\alpha^\star, \alpha)$. The horizontal intercept of $\widehat{h}_0(\alpha^\star, \alpha)$ is the nominal size, $\alpha^\star$, because $\widehat{h}_0(\alpha^\star, \alpha^\star) = 0$. This means that a test designed for size $\alpha^\star$ has no robustness to distributional uncertainty if one requires that the effective size actually equal $\alpha^\star$. The positive slope of $\widehat{h}_0(\alpha^\star, \alpha)$ vs. $\alpha$ means that positive robustness is obtained only for effective size, $\alpha$, greater (worse) than the nominal size $\alpha^\star$. Stated differently, the positive slope of $\widehat{h}_0(\alpha^\star, \alpha)$ expresses a trade-off: the robustness against distributional uncertainty improves as the effective level of significance, $\alpha$, get worse: robustness is exchanged for significance.

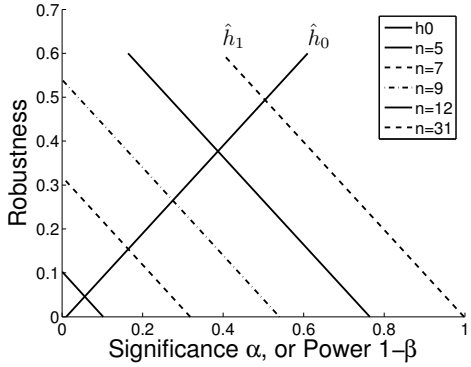


Figure 1: Robustness curves for the $t$ test, $\widehat{h}_0(\alpha^\star, \alpha)$ for falsely rejecting $H_0$, and $\widehat{h}_1(\alpha^\star, \alpha)$ for falsely rejecting $H_1$. Nominal size is $\alpha^\star = 0.01$. $\widehat{h}_1(\alpha^\star, \alpha)$ calculated at 5 different sample sizes: $n = 5$, 7, 9, 12 and 31. $\delta = 1$.

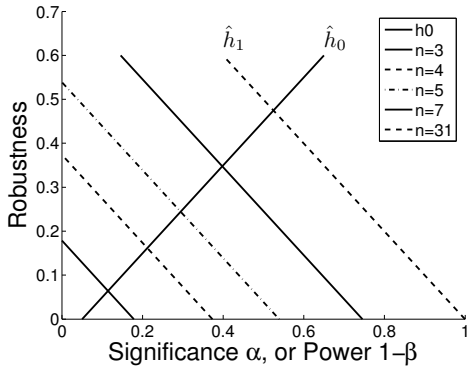


Figure 2: Robustness curves for the $t$ test, $\widehat{h}_0(\alpha^\star, \alpha)$ for falsely rejecting $H_0$, and $\widehat{h}_1(\alpha^\star, \alpha)$ for falsely rejecting $H_1$. Nominal size is $\alpha^\star = 0.05$. $\widehat{h}_1(\alpha^\star, \alpha)$ calculated at 5 different sample sizes: $n = 3$, 4, 5, 7 and 31. $\delta = 1$.

We now can see how one makes judgments of reliable effective size, $\alpha$. A test designed for size $\alpha^\star = 0.01$, as in fig. 1, has no robustness for size 0.01. However, consider an effective size $\alpha = 0.05$ and refer to eq.(15). The test designed for $\alpha^\star = 0.01$ will falsely reject $H_0$ with probability no greater than 0.05 if the actual cdf, $F(y)$, differs from the estimated cdf, $\widetilde{F}_0(y)$, by no more than 0.04 in cumulative probability. For instance, type I error will have probability no larger than 0.05 if the tails of the true distribution are too heavy or too light by no more than 4% of the total probability weight. The distributional uncertainty may arise from the presence of an outlying sub-population. The probability of type I error will not exceed 0.05 provided the sub-population is no larger than 4% of the total, regardless of how it is distributed. Similarly, at effective size $\alpha = 0.1$, a test designed for size $\alpha^\star = 0.01$ is robust to distributional uncertainty up to 0.09 in cumulative probability.

Now consider the robustness curves for type-II error, $\widehat{h}_1(\alpha^\star, \beta)$, eq.(18). The horizontal intercept of $\widehat{h}_1(\alpha^\star, \beta)$ is the nominal power, $1-\beta^\star$, because $\widehat{h}_1(\alpha^\star, \beta^\star) = 0$. This means that a test designed for size $\alpha^\star$ has no robustness to distributional uncertainty if one requires that the effective power actually equal $1-\beta^\star$. The negative slope of $\widehat{h}_1(\alpha^\star, \beta)$ vs. $1-\beta$ means that positive robustness is obtained only for effective power, $1-\beta$, lower (worse) than the nominal power $1-\beta^\star$. Stated differently, the negative slope of $\widehat{h}_1(\alpha^\star, \beta)$ expresses a trade-off: the robustness against distributional uncertainty improves as the effective power, $1-\beta$, get worse: robustness is exchanged for power.

We can now see how one makes judgments of reliable effective power, $1-\beta$. A test designed for size $\alpha^\star = 0.01$ with sample size $n = 9$ (dot-dash in fig. 1), has no robustness for power 0.54 (the horizontal intercept and nominal power). However, consider an effective power $1-\beta = 0.44$ and refer to eq.(18). This test will falsely accept $H_0$ with probability of 0.44 if the actual cdf differs from the estimated cdf by no more than 0.1. At effective size $1-\beta = 0.44$, this test is robust to distributional uncertainty up to 0.1 in cumulative probability. For instance, if the tails err by as much as 10% of the total probability, or if a sub-population with unknown distribution has no more than 10% weight, then the probability of type II error will be no more than 0.44. Similarly, at effective size $1-\beta = 0.34$, this test is robust to distributional uncertainty up to 0.2 in cumulative probability.

Finally, let us consider the choice of the sample size. Only the type-II robustness is influenced by the sample size, as we see from eqs.(15) and (18) and from figs. 1 and 2. The nominal and effective power both

increase with increasing sample size, and are also substantially influenced by the nominal size $\alpha^\star$ as we see by comparing the two figures. The analyst decides on the sample size in light of the effective power and robustness which are needed. We illustrate the decisions and judgments with the aid of fig. 3, which is expanded from fig. 1.

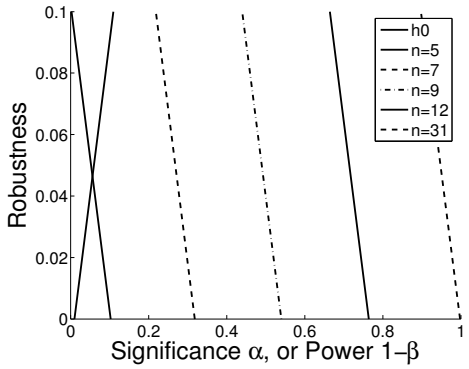


Figure 3: Expanded from fig. 1.

In fig. 3 we are contemplating the choice of nominal size $\alpha^\star = 0.01$. Consider the judgment that effective size $\alpha = 0.05$ is adequate and reliable because the robustness is $\widehat{h}_0(0.01, 0.05) = 0.04$, eq.(15). This judgment considers the robustness and the effective size together since they are linked through the trade-off between them. For instance, the judgment is that the tails are unlikely to err by more than 4%, and the 5% risk of type I error is acceptable. Now apply this robustness to type II error by requiring $\widehat{h}_1(\alpha^\star, \beta) = 0.04$. From fig. 3 we find effective powers of 0.50, 0.72 and 0.96 for sample sizes 9, 12 and 31. Judging that power of 0.50 is too small, we require a sample larger than $n = 9$. If power of 0.72 is adequate then we adopt a sample of size 12. Choosing a sample of size 31 would result in power of 0.96.

Let us continue our consideration of the judgment in the previous paragraph that effective size $\alpha = 0.05$ is adequate and reliable. Judgment is subjective, and this is a two-fold judgment since size and robustness are linked through the trade-off between them. Size, $\alpha$, is subjectively judged in terms of the risk of type I error. Robustness in this case can also be subjectively judged in terms of probability. For instance one might make the judgment that the distribution is distorted by an outlying sub-population whose weight is no more than a few percent of the main population. This robustness judgment can be cast in terms of risk: by accepting a robustness of 0.04 we are accepting the risk that the parent population is contaminated by an outlying population whose weight is no more than 4%.

It may be convenient and familiar for some analysts to judge robustness in this example in terms of probability and risk as just described, However, this is not necessary. Info-gap models of uncertainty are inherently non-probabilistic, and value judgments about robustness can be formed non-probabilistically. Judgments of acceptable risk are based on experience and context. In the same way, analysts can acquire subjective feel for fractional error, or other non-probabilistic quantities, which leads to judgments of acceptable robustness. The concept of analogical inference has been employed to form non-probabilistic value judgments of robustness (Ben-Haim, 2006, chap. 4).

Let us now return to our discussion of choosing the sample size, three paragraphs before, and remove a simplification which we made: applying the same robustness to both type I and type II errors. Having accepted robustness of 0.04 for type I error, $\widehat{h}_0(0.01, 0.05) = 0.04$, we then evaluated the sample size in terms of the same robustness for type II error, $\widehat{h}_1(\alpha^\star, \beta) = 0.04$. This is justified if one faces the same severity of distributional uncertainty for both hypotheses. However, one might well image situations in which the distributional uncertainty is different for the two hypotheses. For instance, one hypothesis may represent a "healthy" state which is more thoroughly studied than the "unhealthy" state represented by the other hypothesis. In such a situation one makes separate judgments of robustness and its trade-off partner (either size or power) for each hypothesis. The judgment of effective size, $\alpha$, is linked to a judgment of $\widehat{h}_0$-robustness. Then one chooses the sample size to yield what is judged to be acceptable type-II robustness, $\widehat{h}_1$, at acceptable power.

## 7 Example: Chronic Wasting Disease

**Verbal description.** Chronic wasting disease (CWD) in deer can be detected by inoculating a particular strain of mice with an extract from the antler velvet of the infected deer. The prion protein (PrP) which is characteristic of this disease is expressed in the mice after a time $t$ which is randomly distributed. This distribution is highly uncertain, and it has been observed that PrP expression with antler velvet from diseased deer frequently does not occur even anomalously long after the mean time (Angers *et al.* 2009). The expression of PrP is much more reliable if the injections are made from the brains of the deer. However, brains may not be available. For instance, antler velvet is used in various traditional Asian medicines which may be the only source for testing.

Suppose that we have inoculated $n$ mice and after incubation times $t_1, \ldots, t_n$, no expression of the PrP is observed in any of the mice. How confident are we

that CWD is not present in the deer?

**System model.** Let $p(t)$ denote the probability density function (pdf) of the incubation time, with cumulative distribution function (cdf) $P(t)$. We assume that the incubation times are statistically independent, so the probability of a false null—true presence with no observed expression of the PrP—is:

$$P_{\rm fn}(t_1,\,\ldots,\,t_n) = \prod_{i=1}^{n}[1-P(t_i)] \tag{19}$$

**Uncertainty model.** Let $\widetilde{p}(t)$ and $\widetilde{P}(t)$ denote the estimated pdf and cdf. Let $t_{\rm s}$ denote a point on the upper tail beyond which the estimated pdf is quite uncertain. For instance we might choose $t_{\rm s}$ to be 2 standard deviations from the mean. We will define an info-gap model in which there are functions whose upper tail, beyond $t_{\rm s}$, decays as $1/t^2$, much slower than the decay of exponential or normal distributions.

Let $\mathcal{P}$ denote the set of non-negative normalized pdf's. The info-gap model, for $h \ge 0$, is:

$$\mathcal{U}(h) = \left\{ p:\ p\in\mathcal{P}, p(t) \le \widetilde{p}(t) + \frac{t_{\rm s}h}{t^2}\forall t \ge t_{\rm s} \right\} \tag{20}$$

The first condition assures that the functions are mathematically legitimate pdf's. The second condition allows the upper tail, beyond $t_{\rm s}$, to exceed the exponential by as much as $t_{\rm s}h/t^2$, conditional on the rest of the distribution being able to adjust to assure non-negativity and normalization.

Note that $\int_{t_{\rm s}}^{\infty} t_{\rm s}h/t^2\,{\rm d}t = h$. Thus the horizon of uncertainty, $h$, represents the fraction of the entire statistical weight which is uncertain. For instance, if the uncertainty of the pdf is thought of as an uncertain mixture of populations, then $h$ is the fraction of the non-$\widetilde{p}$ population.

**Performance requirement.** The probability of a false null must be less than a critical value:

$$P_{\rm fn}(t_1,\,\ldots,\,t_n) \le P_{\rm fnc} \tag{21}$$

**Robustness function.** The robustness is defined as:

$$\widehat{h}(n, P_{\rm fnc}) = \max\left\{ h:\ \left( \max_{p\in\mathcal{U}(h)} P_{\rm fn} \right) \le P_{\rm fnc} \right\} \tag{22}$$

We will evaluate the inverse of $\widehat{h}(n, P_{\rm fnc})$.

Let us denote the inner maximum in eq.(22) by $m(h)$, which is the inverse of $\widehat{h}(n, P_{\rm fnc})$. We will assume that all the observed times, $t_1,\,\ldots,\,t_n$, exceed $t_{\rm s}$, so they fall in the domain of the uncertain tail. In this case, $m(h)$ is evaluated with the upper envelope at horizon of uncertainty $h$, provided that this distribution can be normalized. For each individual observation:

$$\begin{aligned} \max_{p\in\mathcal{U}(h)}[1-P(t_i)] &= \min\left[1,\ \int_{t_i}^{\infty}\left(\widetilde{p}(t) + \frac{t_{\rm s}h}{t^2}\right)\,{\rm d}t\right] \\ &= \min\left[1,\ 1-\widetilde{P}(t_i) + \frac{t_{\rm s}h}{t_i}\right] \end{aligned} \tag{23}$$

Since the $n$ observations are independent we find the inner maximum in eq.(22) to be:

$$m(h) = \prod_{i=1}^{n} \min\left[1,\ 1-\widetilde{P}(t_i) + \frac{t_{\rm s}h}{t_i}\right] \tag{24}$$

Plotting $m(h)$ vs $h$ is equivalent to $P_{\rm fnc}$ vs $\widehat{h}(n, P_{\rm fnc})$.

Eq.(24) can be simplified when the observations, $t_i$, are large, so that $1-\widetilde{P}(t_i)$ is nearly zero. For $h \le 1$:

$$m(h) \approx \frac{t_{\rm s}^n h^n}{\prod_{i=1}^{n} t_i} \tag{25}$$

Equating this to $P_{\rm fnc}$ and solving for $h$ yields an approximate expression for the robustness which is valid when the observations are large:

$$\widehat{h}(n, P_{\rm fnc}) \approx \frac{1}{t_{\rm s}}\left( P_{\rm fnc}\prod_{i=1}^{n} t_i \right)^{1/n} \tag{26}$$

Denoting the geometric mean of the $n$ observations by $\overline{t}_{\rm gm}$, this becomes:

$$\widehat{h}(n, P_{\rm fnc}) \approx \frac{\overline{t}_{\rm gm}}{t_{\rm s}} P_{\rm fnc}^{1/n} \tag{27}$$

The geometric mean observation, $\overline{t}_{\rm gm}$, will change as the sample grows, but the dominant effect of sample size is in the term $P_{\rm fnc}^{1/n}$ which grows rapidly as $n$ increases when $n$ and $P_{\rm fnc}$ are small. Furthermore, when $P_{\rm fnc}$ is very small, the slope of $\widehat{h}$ vs $P_{\rm fnc}$ increases as $n$ increases. This means that, when $P_{\rm fnc}$ is small, the cost of robustness, in units of increased $P_{\rm fnc}$, is small when $n$ is large.

**Example.** Fig. 4 shows robustness curves, based on eq.(24), for 5 sample sizes with the following data $t_i$ = 500, 530, 510, 520, 505 days. The bottom curve ($n$ = 1) uses only the first datum; the next curve uses the first 2 data; etc. The estimated distribution is normal with mean and standard deviation of 450 and 20 days. $t_{\rm s} = 490$.

The positive slopes of the curves express the trade-off between robustness, $\widehat{h}$, and critical probability of false null, $P_{\rm fnc}$. Large robustness is obtained only by accepting large $P_{\rm fnc}$. The robustness is zero at the estimated value of $P_{\rm fnc}$.

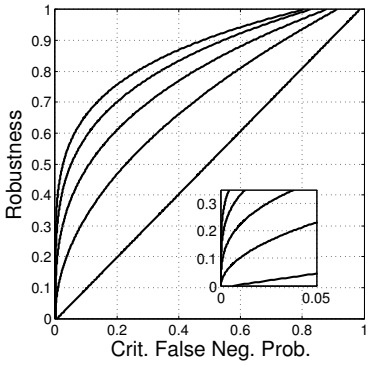


Figure 4: $\widehat{h}(n, P_{\text{fnc}})$ vs $P_{\text{fnc}}$, $n = 1$ to 5 (bottom to top).

The robustness increases substantially as the sample size increases from $n = 1$ to 2. The marginal increase in robustness decreases with increasing $n$. From the insert in the figure we see that the slope of the robustness curve increases dramatically as the sample size increases. A high slope means that the robustness can be increased without significantly increasing the critical probability of false null, $P_{\text{fnc}}$.

## 8 Methodological Conclusion

This paper concentrates on binary simple-hypothesis statistical tests, subject to distributional uncertainty, by which we mean uncertainty in the sampling distribution resulting from unknown violations of the test assumptions. We have focussed on two decisions and two judgments which the analyst must make. How can one *decide upon* the decision threshold and the sample size, and how does one *judge* the effective size and power of a test? We have developed a generic approach to these questions based on info-gap decision theory, and illustrated the method with the $t$ test and with a test for false nulls. The method can be applied to other tests as well.

Consider a test which is designed to have nominal level of significance $\alpha^\star$. The robustness of this test with respect to distributional uncertainty, for falsely rejecting $H_0$ in eq.(1), is denoted $\widehat{h}_0(\alpha^\star, \alpha)$ and defined in eq.(14). $\widehat{h}_0(\alpha^\star, \alpha)$ is the greatest horizon of distributional uncertainty up to which the test, with nominal size $\alpha^\star$, falsely rejects $H_0$ with probability no greater than $\alpha$. That is, $\widehat{h}_0(\alpha^\star, \alpha)$ is the greatest horizon of uncertainty up to which the probability of type I error (false rejection of $H_0$) is no greater than $\alpha$, when using a test with nominal size $\alpha^\star$.

$\widehat{h}_0(\alpha^\star, \alpha)$ is necessarily zero when $\alpha = \alpha^\star$, implying that the test has no robustness to distributional uncertainty at its nominal size, $\alpha^\star$. The robustness is positive for $\alpha > \alpha^\star$, and the robustness increases as $\alpha$ gets larger. This expresses the trade-off between robustness to distributional uncertainty on the one hand, and effective level of significance on the other hand, as illustrated by eq.(15) and the lines of positive slope in figs. 1–3.

The robustness function $\widehat{h}_0(\alpha^\star, \alpha)$ is the basic tool for choosing the decision threshold, $q_{\alpha^\star}(\widetilde{F}_0)$ in eq.(13), and for evaluating the effective size, $\alpha$, of the test. If $\widehat{h}_0(\alpha^\star, \alpha)$ is large then one has confidence that the probability of falsely rejecting $H_0$ is no greater than $\alpha$. What constitutes a 'large' robustness, and 'how large is large enough' are delicate value judgments, somewhat like the choice of level of significance. We discussed this in section 6, though there is no absolute answer.

We have also considered the robustness to distributional uncertainty in evaluating the effective power. For any test designed for size $\alpha^\star$, the robustness to distributional uncertainty, for falsely accepting $H_0$ (type II error), is denoted $\widehat{h}_1(\alpha^\star, \beta)$, defined in eq.(16). The power, $1 - \beta$, is the probability of correctly rejecting $H_0$. $\widehat{h}_1(\alpha^\star, \beta)$ is the greatest horizon of distributional uncertainty up to which the test, with nominal size $\alpha^\star$, will falsely accept $H_0$ with probability no greater than $\beta$. The robustness is zero when $\beta$ is the value obtained, at size $\alpha^\star$, in the absence of distributional uncertainty. That is, there is no robustness for the nominal power. The robustness increases as the power decreases, as illustrated by eq.(18) and the lines of negative slope in figs. 1–3.

The robustness functions $\widehat{h}_1(\alpha^\star, \beta)$ and $\widehat{h}_0(\alpha^\star, \alpha)$ are the basic tools for choosing the sample size and for evaluating the effective power of a test. If $\widehat{h}_1(\alpha^\star, \beta)$ is large then one has confidence that the probability of correctly rejecting $H_0$ is no less than $1 - \beta$ with the chosen sample size. Once again, judgments of adequate power and large robustness are subjective.

We have concentrated on tests of the mean with binary simple hypotheses, both because such tests are exceedingly common in practice, and because the main aim was to demonstrate the methodology of info-gap theory for evaluating effective size and power and for selecting the decision threshold and sample-size. The methodology developed in this paper can be extended to other test structures, and to tests of quantities other than the mean. Furthermore, the close relation between hypothesis tests and confidence intervals enables the application of the methodology to evaluating and selecting confidence intervals.

**Acknowledgements**

The author is indebted to Mark A. Burgman, Ayala Cohen, David Fox, Malka Gorfine, Mick McCarthy, Andrew Robinson and Miriam Zacksenhouse for valuable comments. Funding for this research was

provided by the U.S. Department of Agriculture under USDA/ERS/PREISM Cooperative Agreement No.58-7000-8-0095.

## 9 References


Angers, Rachel C., *et al.*, 2009, Chronic Wasting Disease Prions in Elk Antler Velvet, *Emerging Infectious Diseases,* Vol. 15, No. 5, pp.696–703.

Bausch, Daniel G. *et al.* 2003, Risk Factors for Marburg Hemorrhagic Fever, Democratic Republic of the Congo, *Emerging Infectious Diseases,* Vol. 9, No. 12, pp.1531–1537.

Ben-Haim, Yakov (2006). *Info-Gap Decision Theory: Decisions Under Severe Uncertainty,* 2nd edition, London: Academic Press.

Boone, Randall B. and William B. Krohn (1999). Modeling the occurrence of bird species: are the errors predictable? *Ecological Applications* **9**, 835–848.

Burgman, Mark A., Roger C. Grimson and Scott Ferson (1995). Inferring threat from scientific collections, *Conservation Biology* **9**, 923–928.

Carpenter, Stephen R. (1989). Replication and treatment strength in whole-lake experiments (1989). *Ecology* **70**, 453–463.

Craft, Christopher, Judy Reader, John N. Sacco and Stephen W. Broome (1999). Twenty-five years of ecosystem development of constructed *Spartina alterniflora* (Loisel) marshes, *Ecological Applications* **9**, 1405–1419.

DeGroot, Morris H. (1986). *Probability and Statistics,* 2nd ed., Reading, MA: Addison-Wesley.

Feller, William (1971). *An Introduction to Probability Theory and Its Applications,* vol. 2, 2nd ed. New York: Wiley.

Fox, David R., Yakov Ben-Haim, Keith R. Hayes, Michael McCarthy, Brendan Wintle, Piers Dunstan (2007). An info-gap approach to power and sample size calculations, *Environmetrics* **18**, 189–203.

Franklin, Donald C. (1999). Evidence of disarray amongst granivorous bird assemblages in the savannas of northern Australia, a region of sparse human settlement, *Biological Conservation* **90**, 53–68.

Huber, Peter J. (1981). *Robust Statistics,* New York: Wiley.

Johnson, Douglas H. (1995). Statistical sirens: The allure of nonparametrics, *Ecology* **76**, 1998–2000.

McCarthy, Michael A. (1998). Identifying declining and threatened species with museum data, *Biological Conservation* **83**, 9–17.

Mooney, Christopher Z. and Robert D. Duval (1993). *Bootstrapping: A Nonparametric Approach to Statistic Inference,* London: Sage Publications.

Robert, Christian P. (2004). *Monte Carlo Statistical Methods,* 2nd ed., New York: Springer.

Stewart-Oaten, Allan, James R. Bence, and Craig W. Osenberg (1992). Assessing effects of unreplicated perturbations: No simple solutions, *Ecology* **73**, 1396–1404.

Titterington, D.M., A.F.M. Smith, U.E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions,* Chichester: Wiley.


## A Evaluating the Robustness $\widehat{h}_0(\alpha^\star, \alpha)$ for Falsely Rejecting $H_0$

In this appendix we derive $\widehat{h}_0(\alpha^\star, \alpha)$ based on the info-gap model in eq.(7).

First $V(x) = 0$ if $x < 0$, $V(x) = x$ if $0 \le x \le 1$, $V(x) = 1$ if $x > 1$.

Let $m_0(h)$ denote the inner minimum in the definition of the robustness in eq.(14). The robustness, $\widehat{h}_0(\alpha^\star, \alpha)$, is the greatest horizon of uncertainty, $h$, at which $m_0(h) \ge 1 - \alpha$. $m_0(h)$ decreases with increasing $h$ because the sets $\mathcal{U}_0(h)$ of the info-gap model become more inclusive as $h$ increases (the nesting axiom). Hence the robustness is the greatest non-negative value of $h$ for which $m_0(h) = 1 - \alpha$. If there is no such value of $h$, then the robustness is zero.

The inner minimum in eq.(14) is obtained when $F(y)$ is as small as possible at $q_{\alpha^\star}(\widetilde{F}_0)$, subject to membership in $\mathcal{U}_0(h)$. From the info-gap model in eq.(7) we find:

$$m_0(h) = V\left(\widetilde{F}_0[q_{\alpha^\star}(\widetilde{F}_0)] - h\right) = V(1 - \alpha^\star - h) \quad (28)$$

where we recall that $\widetilde{F}_0[q_{\alpha^\star}(\widetilde{F}_0)] = 1 - \alpha^\star$. The greatest value of $h$ at which $m_0(h) = 1 - \alpha$ is the robustness, eq.(15).

## B Evaluating the Robustness $\widehat{h}_1(\alpha^\star, \beta)$ for Correctly Rejecting $H_0$

In this appendix we derive $\widehat{h}_1(\alpha^\star, \beta)$ based on the info-gap model in eq.(7).

Let $m_1(h)$ denote the inner maximum in the definition of the robustness in eq.(16). The nesting axiom implies that $m_1(h)$ increases monotonically as $h$ increases. Consequently the robustness, $\widehat{h}_1(\alpha^\star, \beta)$, is the greatest horizon of uncertainty, $h$, at which $m_1(h) = \beta$.

From the info-gap model in eq.(7), and using the step function $V(x)$ defined earlier, we find:

$$m_1(h) = V\left(\widetilde{F}_1[q_{\alpha^\star}(\widetilde{F}_0)] + h\right) \quad (29)$$

Equating this to $\beta$ and solving for $h$ we find the robustness in eq.(18) with the aid of the expression for the nominal power in eq.(17).