Memorial Sloan-Kettering Cancer Center

Memorial Sloan-Kettering Cancer Center, Dept. of Epidemiology & Biostatistics Working Paper Series

Year 2005 *Paper* 4

Computing the Total Sample Size When Group Sizes Are Not Fixed

Mithat Gonen*

*Memorial Sloan-Kettering Cancer Center, gonenm@mskcc.org

Computing the Total Sample Size When Group Sizes Are Not Fixed

Mithat Gonen

Abstract

This article is concerned with computing the total sample size required for a two-sample comparison when the sizes of the two groups to be compared cannot be fixed in advance. This is frequently encountered when group membership depends on a variable which is observable only after the subject is enrolled to the study, such as a genetic or a biological marker. The most common way of circumventing this problem is assuming a fixed number for the prevalence of the condition that will determine the group membership and compute the required sample size conditionally. In this article this practice is formalized by placing a prior distribution on the prevalence which results in an analytically tractable formula for the unconditional sample size. In particular a sample size inflation factor, a number that can be multiplied with conditional sample size, is presented. An example is given from the planning of a clinical trial investigating the prognostic role of molecular markers in gastrointestinal stromal cancer.

Computing the Total Sample Size When Group Sizes Are Not Fixed

Mithat Gönen
Department of Epidemiology and Biostatistics
Memorial Sloan-Kettering Cancer Center
New York, NY 10021
gonenm@mskcc.org

Summary. This article is concerned with computing the total sample size required for a two-sample comparison when the sizes of the two groups to be compared cannot be fixed in advance. This is frequently encountered when group membership depends on a variable which is observable only after the subject is enrolled to the study, such as a genetic or a biological marker. The most common way of circumventing this problem is assuming a fixed number for the prevalence of the condition that will determine the group membership and compute the required sample size conditionally. In this article this practice is formalized by placing a prior distribution on the prevalence which results in an analytically tractable formula for the unconditional sample size. In particular a sample size inflation factor, a number that can be multiplied with conditional sample size, is presented. An example is given from the planning of a clinical trial investigating the prognostic role of molecular markers in gastrointestinal stromal cancer.

1 Introduction

Two-sample comparisons are fundamental in statistics and the sample size requirements to adequately power these tests are well-studied. In particular, one can find a closed-form approximation to the sample size for every conceivable two-sample comparison, including normal, binary, ordinal, survival and count data (see, for example, Cohen, 1988; Fleiss, 1981; Gail, 1974, Noether, 1987; Lachin, 1981; Campbell, Julious and Altman, 1995). All of these formulae presume that group sizes can be fixed in advance, a reasonable assumption for randomized studies. In some instances, however, group sizes cannot be fixed in advance because group membership is determined by variable which is observable only after the subject is recruited to the study. A genetic or biological marker that is thought to have a bearing on the outcome of the patient, but can only be observed after the patient is enrolled in the clinical trial, is a commonly encountered example. The term "marker" will be used throughout the paper to denote the variable that determines group membership.

This article addresses the case where the marker is dichotomous and the corresponding comaprison of interest is a two-sample test. The two categories will be called positive (+) and negative (–) generically. Marker status for a given patient usually requires tissue or blood samples to be collected for which the patient has to be consented and enrolled to the study. Therefore the numbers of marker positive and marker negative patients are beyond the control of the investigators. An example is given in Section 4 on a problem of this nature that arised in the context of a clinical trial.

The proportion of marker positive patients will be referred to as *prevalence*. The current practice in handling these situations is to guess the prevalence, either eliciting it from clinical investigators or from the numbers reported in the literature for similar patient populations, and treat it as fixed. This results in a conditional procedure where the computed sample size provides adequate power only if the presumed prevalence is correct. As a partial remedy, it is customary to offer power calculations for a variety of other values of prevalence, providing an idea for which values the study will have acceptable power.

This procedure is in common use, mostly due to its simplicity, but it is an underestimate of the required sample size since it ignores the variability in prevalence. Here an unconditional approach which uses a prior distribution on the prevalence and averaging over it with respect to the prior is presented. This method is analytically tractable and results in a simple adjustment that can be used to inflate the sample size obtained conditionally. One must note that, although the term prior is used, our method is not Bayesian in a general sense. In particular a prior distribution for the parameter of interest is not used.

The main results appear in the next section and the choice of prior is discussed in Section 3. Section 4 contains a case study on the planning of a clinical trial investigating the prognostic role of molecular markers in gastrointestinal stromal cancer (GIST). and Section 5 contains the concluding remarks.

2 Methodology

The following establishes some necessary notation: $\dot{n}$ is the sample size required for a two-sample comparison with equal group sizes, $\tilde{n}$ is the sample size required for the same two-sample comparison with unequal group sizes when the groups sizes can be specified in advance (conditional sample size) and n is the sample size required for the same two-sample comparison with unequal group sizes when the groups sizes cannot be specified in advance (unconditional sample size). It will be assumed that, for the endpoint in question, a procedure is available for computing $\dot{n}$. It is well-known that

$$\tilde{n}|p = \frac{\dot{n}}{4p(1-p)} \tag{1}$$

where p denotes the prevalence of one of the groups with respect to the total sample size. The conditional approach calls for inserting the best guess, say $p = p_0$, in (1). In this article this approach is formalized by using an unconditional formula:

$$\begin{aligned} n &= E_p(\tilde{n}|p) \\ n &= \int_0^1 [4p(1-p)]^{-1} \, \dot{n} \, dG(p) \end{aligned} \tag{2}$$

To apply (2) one would need to specify $G(p)$. Note that assuming p known can be considered as a special case of (2) where $G(p)$ is a point mass at p_0 so this common practice is equivalent to specifying *some* $G(.)$.

A more realistic distribution for p is the standard Beta family with parameters α and β, denoted by $Be(\alpha, \beta)$. The density and cumulative density functions of $Be(\alpha, \beta)$ will be denoetd by $f(p; \alpha, \beta)$ and $F(p; \alpha, \beta)$, respectively.

Beta family is not only flexible enough to represent varying levels of prior information about p, but, because of the way $p(1-p)$ appears in the integral, leads to a simple analytical solution, provided $\alpha > 1$ and $\beta > 1$, as well:

$$n = \frac{(\alpha+\beta-1)(\alpha+\beta-2)}{4(\alpha-1)(\beta-1)}\dot{n}$$

The term that multiplies $\dot{n}$ will be called as the sample size inflation factor (SSIF) since it represents the amount one needs to inflate the sample calculation if groups are assumed equal sizes:

$$SSIF(\dot{n}) = \frac{(\alpha+\beta-1)(\alpha+\beta-2)}{4(\alpha-1)(\beta-1)} \tag{3}$$

Remark 1: Note that when $\alpha < 1$ or $\beta < 1$

$$\begin{aligned} \lim_{p\to 0} f(p;\alpha,\beta) &= 0 \\ \lim_{p\to 1} f(p;\alpha,\beta) &= 0 \end{aligned}$$

Therefore, from a practical standpoint, the requirement that both $\alpha > 1$ and $\beta > 1$ assures, by removing all the appreciable probability in the neighborhoods of the boundary points, that prevalence is sufficiently away from 0 or 1.

Remark 2: The mean of the beta distribution, $\alpha/(\alpha+\beta)$, can be taken to be equal to p_0. When α and β tend to infinity in such a way that $p_0 = \alpha/(\alpha+\beta)$ remains constant, the density approaches a point mass at p_0. SSIF, in this case, approaches $4\alpha\beta/(\alpha+\beta) = p_0(1-p_0)$. This establishes the conditional approach as a limiting case of the unconditional one.

Remark 3: It is sometimes of interest to compute the power for a given sample size. Since p appears in the upper limit of the integral in a power calculation, direct derivation for a power inflation factor is elusive. Nevertheless one can recompute the approximate power of a study using n in the power equation that is appropriate for the problem at hand.

3 Choosing α and β

Elicitation of α and β is critical for applications. In some cases, especially for markers that are frequently encountered in the literature, there will be some prior reports which provide prevalence estimates for the markers. Table 1, further explained in the next section, is an example of a summary of previous reports for a particular marker in gastrointestinal stromal cancer (GIST). Gnanadesikan, Pinkham and Hughes (1967) showed that, to fit a beta distrubution to such data, maximum likelihood equations can be expressed in terms of digamma functions and solved iteratively. Specifically, if there are s comparable studies in the literature which report prevalences x_j based on a sample size of w_j, $j = 1, \ldots, s$, then maximum likelihood estimates $\hat{\alpha}$ and $\hat{\beta}$ will satisfy

$$\frac{1}{s}\sum_{j=1}^{s} w_j \log(x_j) = \Psi(\hat{\alpha}) - \Psi(\hat{\alpha}+\hat{\beta}) \tag{4}$$

$$\frac{1}{s}\sum_{j=1}^{s} w_j \log(1-x_j) = \Psi(\hat{\beta}) - \Psi(\hat{\alpha}+\hat{\beta}). \tag{5}$$

where $\Psi(s) = \Gamma'(s)/\Gamma(s)$ is the digamma function.

If there is no reliable literature survey but the investigators have strong convictions about a range for the possible values of p, the prevalence, denoted by (p_L, p_U), such that $P(p \in (p_L, p_U)) = 1 - q$, where $1 - q$ can be thought of the "confidence" associated with this choice. In this case

$$F(p_U;\alpha,\beta) - F(p_L;\alpha,\beta) = 1 - q \tag{6}$$

by definition and there are several (α,β) pairs that satisfy (6). Therefore a further restriction on the parameters space is needed. One possibility is setting $p_0 = \alpha/(\alpha+\beta)$. Another possibility is imposing symmetry in terms of the tail probabilities:

$$F(p_L;\alpha,\beta) = \frac{q}{2} \tag{7}$$

Strictly speaking (7) can be satisfied only when $\alpha = \beta$, that is when the beta distribution is symmetric around its mean. But it may be possible to find a solution that yields approximately equal tail probabilities. Either approach involves solving a non-linear system. Since F is not available in closed-form a solution can be obtained only numerically using an iterative method. An automated iteration is rarely necessary and a manual trial-and-error approach, with the help of a statistical software that evaluates F, is sufficient.

4 Case Study: KIT mutations in GIST

Nearly all gastrointestinal stromal tumors express the receptor tyrosine kinase KIT that has a role in cell proliferation and survival. Most GISTs have a gain of function mutation in the KIT proto-oncogene resulting in constitutive KIT activation in the absence of its natural ligand. STI571, commerically known as Gleevec, is a selective molecular inhibitor of KIT and has shown remarkable activity in patients with metastatic GIST. In June 2002, a multicenter, randomized, placebo-controlled trial (American College of Surgeons

Oncology Group Study Z9001) was opened to test the benefit of adjuvant STI571 in patients following the resection of their GIST.

The investigators propose to study the tissue specimens collected from the patients enrolled in this multicenter trial. In particular they hypothesize that the molecular characteristics of primary GIST predict the clinical outcome after adjuvant STI571 therapy. This is formulated in a sequence of two-sample hypothesis tests which compare the patients with KIT mutation with those who do not which is unknown at the time a grant proposal is written for this moelcular study.

Antonescu et. al. (2003) summarize the available evidence on KIT mutations in GIST patients. A total of seven studies are published and their relevant results are summarized in Table 1. For the moment we ignore the tissue type and maximize the likelihood to find $\alpha = 3.95$ and $\beta = 2.71$ (Figure 1). This results in a sample size inflation factor of 1.31. One of the goals of the study is to see larger tumors (> 5cm) is associated with KIT mutations. Using the method of Fleiss (1981) one finds a total sample size of 118 to detect a difference of 20% (70% to 90%) in the prevalence of KIT mutations in small and large tumor categories. This calculation assumes equal group sizes. Therefore if the study is planned for 155 patients, it will, on average, have 80% power to test this hypothesis.

One can visually confirm (see Figure 2) the fit of the prior by overlaying the model probabilties (smooth curve) with the actual probabilities (step function). While the two distributions have substantial differences in the middle, the overall fit is reasoanble. Another way to check if the chosen prior is sensible is to see if the implied tail probabilities are sensible. In this case

$$\begin{aligned} F(0.2, 2, 1.35) &= 0.015 \\ F(0.9, 2, 1.35) &= 0.025 \end{aligned}$$

and the investigators found these extreme probabilities to be reasonable. If this were not the case, a calibration of the prior would have been necessary.

Fitting a single distribution to prior data does not reveal possible inconsistencies in the literature. In this case, one can observe from Table 1 that even large studies disagreed with each other. It is possible to visualize this by fitting a mixture prior to the data, i.e. each study k would be assigned a prior with $\alpha_k - 1$ being equal to the number of patients with KIT mutations and $\beta_k - 1$ being equal to the number of patients without KIT mutations, for $k = 1, \ldots, 7$. This choice of prior parameters can be motivated by Bayesian practice, see Gelman et. al. (2000) for example. If one plots the mixture of these betas (using equal mixture probabilities since sample sizes are already factored in through the choice of α_k and β_k), as done in Figure 3, a disturbing feature is revealed: the priors for the big studies do not overlap for the most part, suggesting a hidden covariate (perhaps more than one) which explains what mode the particular study belongs to in this multi-modal distribution.

It turns out that one such covariate is not hidden: one might notice that the studies which had paraffin-embedded tissues (marked with P in the table) reported lower prevalences than those studies which has fresh frozen tissues (marked with F in the table). This was not a surprise to the study pathologist. Processing fresh frozen tissue requires an on-site tumor bank, a substantial investment, and, as a result, many small institutions or community hospitals still prefer paraffin embedding despite the fact that it involves a sequence of operations which may inadvertently alter the genetic makeup of the tissue and result in false negatives. Z9001 will encompass more than one hundred institutions and will inevitably produce both paraffin and fresh frozen tissue. For this reason the recommendation for this particular example is to use the SSIF of 1.31. In single-instituion studies, or in cases where it is known that all samples will come from similar source of tissue, it would be prudent to fit a prior only to the relevant reports. Figure 4 gives examples of a prior fit only to the studies reporting fresh frozen tissue (solid curve) and only to the ones reporting paraffin-embedded tissue (dotted curve).

5 Discussion

This article presents an approach to computing the sample size for two-sample comparisons when the experimenters do not have control over group sizes. The proposed method is an improvement over the common practice of fixing the group sizes and computing the sample size conditionally since it explicitly models the variability in group sizes. It has the advantage of being directly applicable to any two-sample problem for which the total size can be computed with known group sizes. In addition it can be expressed as a simple factor that multiplies the conditional sample size.

Modeling the uncertainty about group sizes is achieved through a parametric prior model, reminscent of Bayesian analysis, and the choice of prior parameters are critical for the appropriate application of the method. The example from the GIST clinical trial demonstrates different aspects of choosing these prior parameters. This is an application-specific issue and a close interaction between the clinical investigators and the statisticians is essential for this purpose.

This approach essentially computes the sample size which will, on average, deliver the desired power. Proponents of the conditional approach may object to this, since, the actual power of the study will depend only on the observed group sizes and not the ones that could have been observed but were not. In fact it can be shown that group size is an ancillary statistic and, when observed, conditioning on it will always result in more efficient inferences. On the other hand, the idea of averaging over the sample space is fundamental to the frequentist school of inference and the notion of power has a distinct frequentist flavor. Therefore the idea of an "average" sample size is consistent with the practice of Neyman-Pearson style hypothesis testing. From a practical standpoint using a single best guess leads to underestimating the sample size and using an extreme point can be excessively conservative.

Similarly, one can argue that, if there is substantial evidence that the marker prevalence is multi-modal, averaging over the entire mixture may be misleading since only the component in which the current study will be is relevant. Although this can be countered using the same principle from the previous paragraph, one needs to be careful about which previous studies are relevant for the current one. One possible source of multi-modality is a systematic difference, such as the pathological methods used to determine marker positivity. In this case only the methods that are the same as, or very similar to, the study at hand should be considered. This is critically different from a unimodal prior postulates that all previous studies are essentially exchangable as far as the prevalence is concerned.

Finally, a viable alternative to the proposed method is sample size re-estimation (Gould, 2001). One can, after a small group of patients is enrolled, estimate the prevalence and use it to re-estimate the sample size. In larger trials this could be feasible and may result in a more powerful design but it is unlikely to be useful in small trials. Another issue which limits its applicability even in large trials is that, most multi-center clinical trials involving pathological analysis designate a central laboratory which usually processes the samples in large batches for reasons of accuracy and cost. It is in fact very common that all samples are analyzed at the end of the trial. Under such a protocol sample size re-estimation will not be feasible.

In summary, the proposed method is a reasonable way to take into account the variability in group sizes at the design stage. It is simple to apply, although choosing the prior parameters should be done with care and in close collaboration with clinical invetsigators. This serves to reinforce the notion that the role of the statistician in designing a study is not only providing sample size and power for a variety of scenarios but actively participating in building the scenarios and deciding which of them are more relevant.

References

Antonescu CR, Sommer G, Sarran L, Tschernyavsky SJ, Riedel E, Woodruff JM, Robson M, Maki R, Brennan MF, Ladanyi M, DeMatteo RP, Besmer P. (2003) Association of KIT exon 9 mutations with nongastric primary site and aggressive behavior: KIT mutation analysis and clinical correlates of 120 gastrointestinal stromal tumors. *Clinical Cancer Research*, **15:**3329–37.

Campbell MJ, Julious SA, Altman DG. (1995) Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *British Mediacal Journal*, **311:**1145–1148.

Cohen, J. (1988) *Statistical power analysis for the behavioral sciences.* Lawrence Erlbaum Associates (Hillsdale, NJ).

Fleiss JL. (1981) *Statistical methods for rates and proportions.* John Wiley and Sons (New York; Chichester).

Gail M. (1974) Power computations for designing comparative Poisson trials (Corr: 83V39 p1137). *Biometrics* **30:**231–237

Gould AL. (2001) Sample size re-estimation: recent developments and practical considerations. *Statistics in Medicine* **20:**2625–43.

Gnanadesikan R, Pinkham RS, Hughes LP. (1967) Maximum likelihood estimation of the parameters of the beta distribution using from the smalles order statistics.. *Technometrics* **9:**607–620.

Lachin JM. (1981) Introduction to sample size determination and power analysis for clinical trials. *Controlled Clincal Trials* **2:**93–114.

Noether GE. (1987) Sample size determination for some common nonparametric tests. *J Am Stat Assoc* **82:** 645–647.

Table 1: Published studies on KIT mutations in GIST. Tissue indicates the type of tissue used for molecular analysis: P for paraffin-embedded tissue and F for fresh frozen tissue. NR stands for not reported.

Study	1	2	3	4	5	6	7
Tissue	NR	P	P	P	P	F	F
Sample size	6	35	46	200	124	45	120
KIT mutation	5	13	9	111	71	40	94

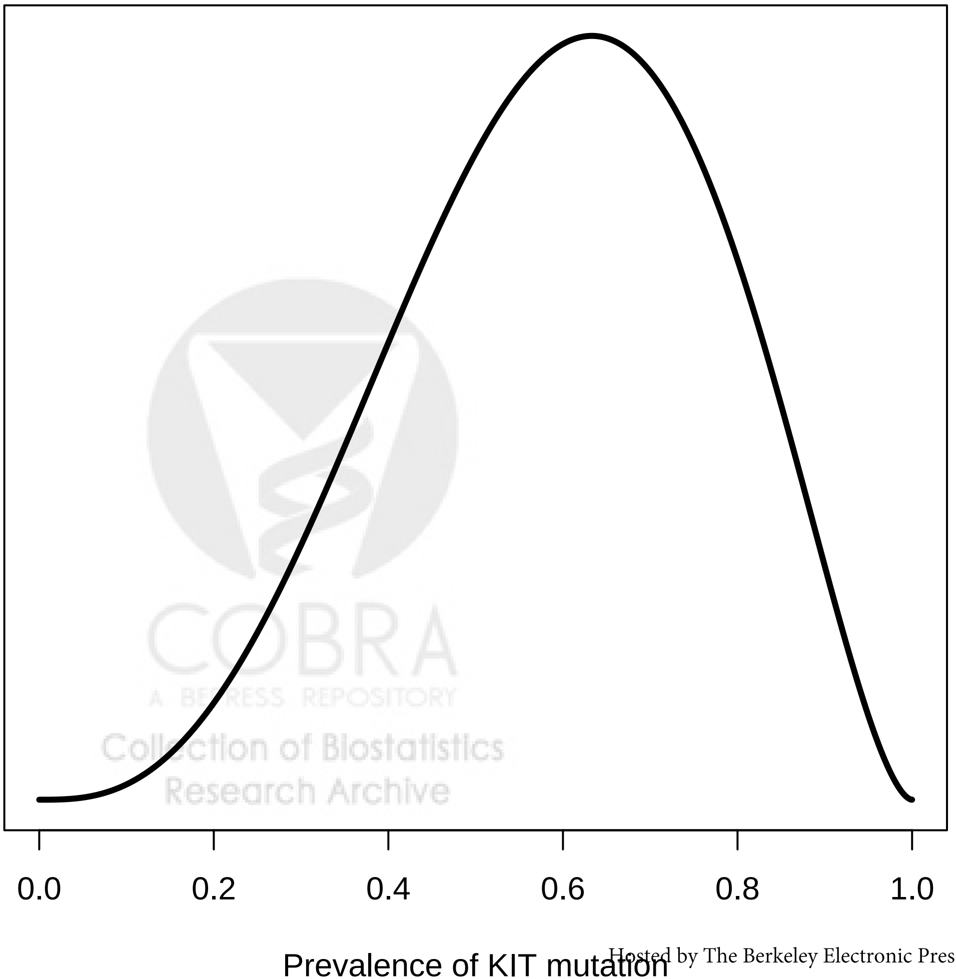
Figures

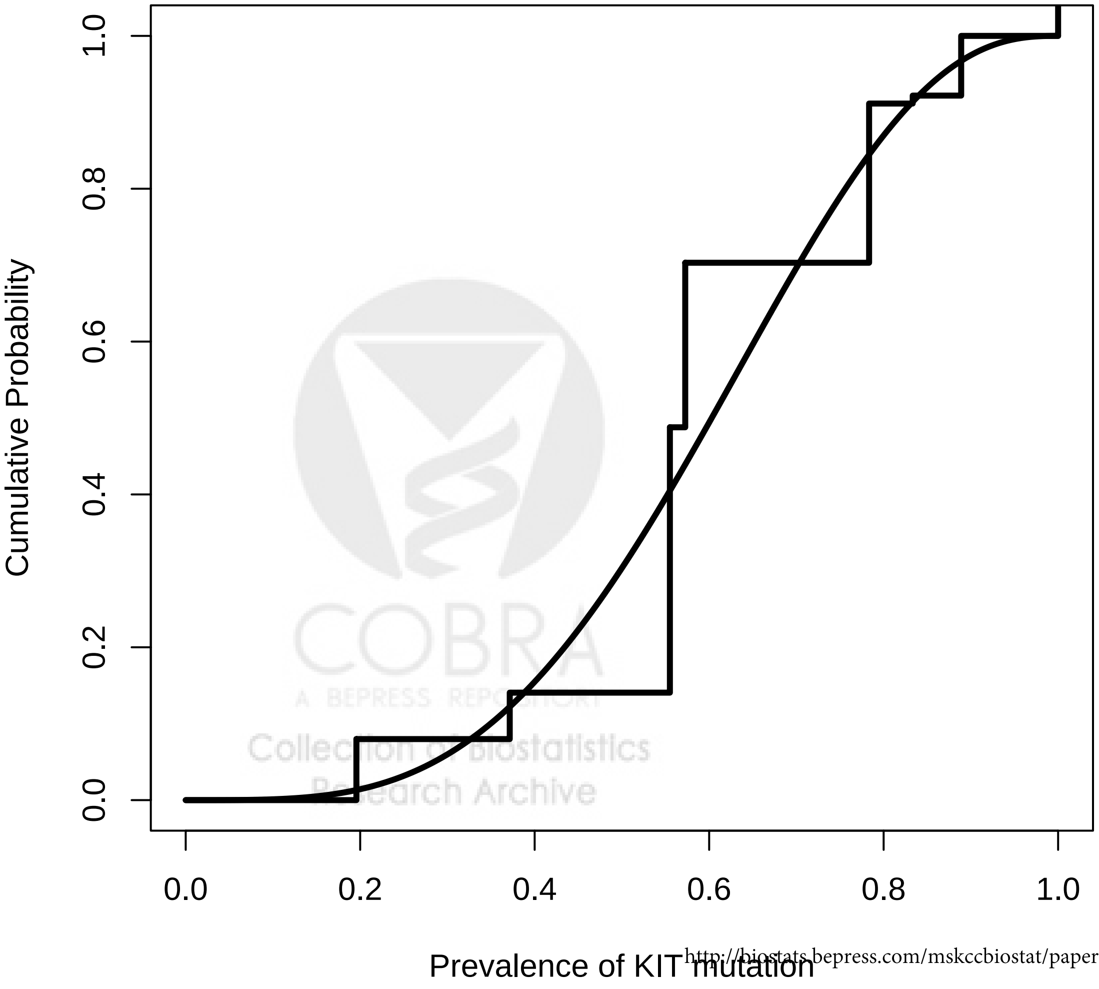
Figure 1: Beta prior with parameters $\alpha = 3.95$ and $\beta = 2.71$.

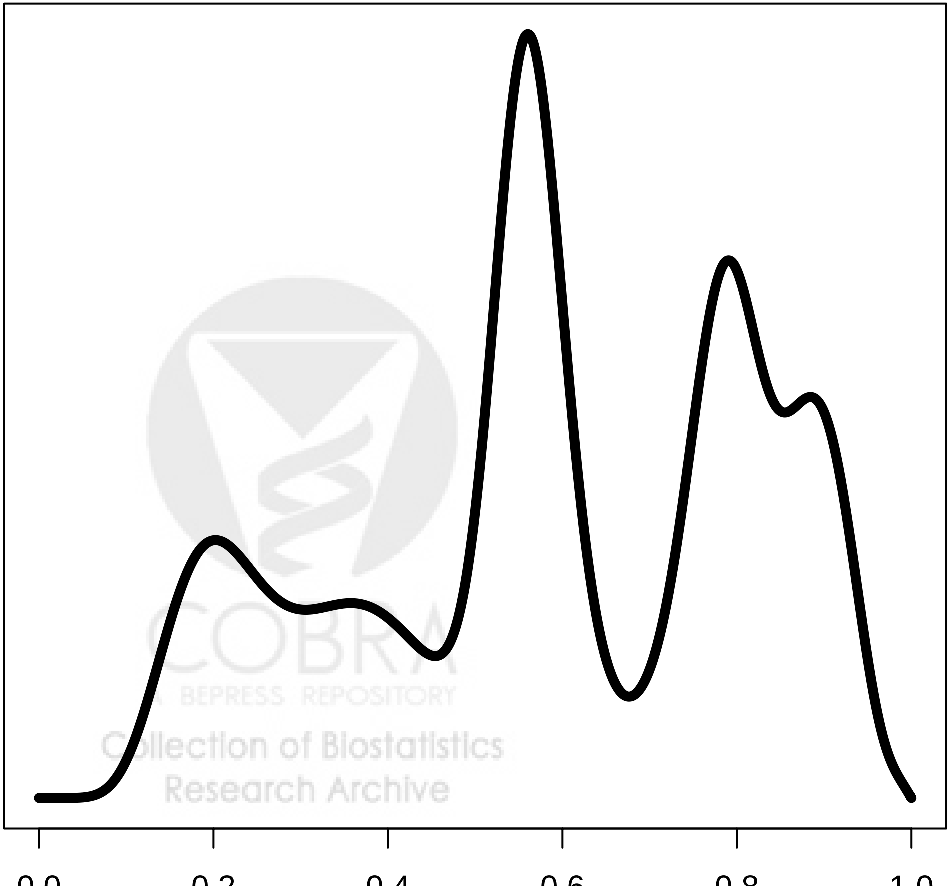
Figure 2: Empricial probabilities from Table 1 and the fitted cumulative prior ($\alpha = 3.95$ and $\beta = 2.71$).

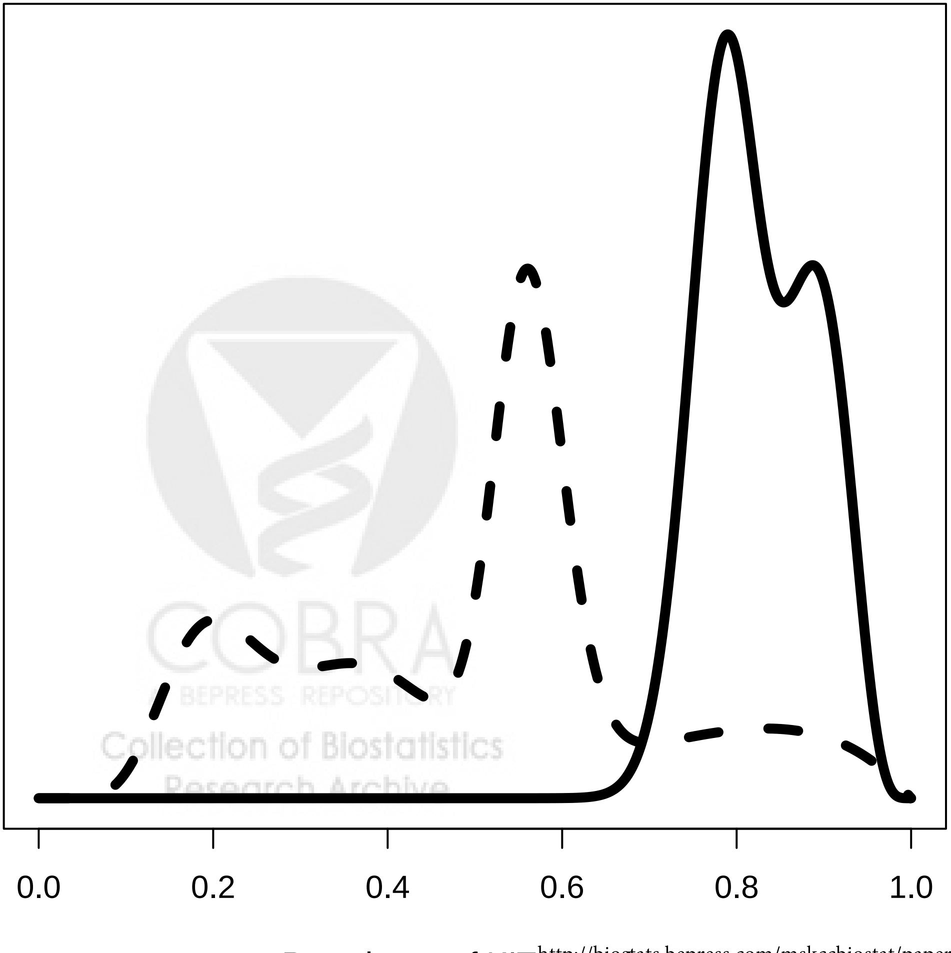
Figure 3: Mixture prior for all the seven studies in Table 1

Figure 4: Mixture priors for studies using fresh frozen tissues (solid) and studies using paraffin tissue (dotted).


0.0
0.2
0.4
0.6
0.8
1.0
Prevalence of KIT mutation


Cumulative Probability
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Prevalence of KIT mutation


0.0
0.2
0.4
0.6
0.8
1.0
Prevalence of KIT mutation


0.0
0.2
0.4
0.6
0.8
1.0
Prevalence of KIT mutation