Minimax Regret in Practice - Four Examples on Treatment Choice[1]

Patrick Eozenou, Javier Rivas and Karl H. Schlag
European University Institute[2]

June 20, 2006

[1]We would like to thank Esther Duflo, Al Roth and Dr. Manu Sabeti for sending us their data and David Yves Albouy for putting the data on his homepage. We would like to thank Decio Coviello and Avner Shaked.

[2]Economics Department, European University Institute. Via della Piazzuola 43, 50133 Florence (Italy). Contact author: schlag@iue.it.

Abstract

Minimax regret is an exact method for deriving a recommendation based on a small sample. It can incorporate costs in its measurement of opportunity loss (regret) in terms of not making the best choice. In this paper we present the methodolgy and implement it in four examples from different fields: medicine, development policy, experimental game theory and macro economics. We focus on the comparison between two treatments with unknown response. Recommendations based on the binomial average rule, the correlated binomial average rule and on the empirical success rule are derived. Point estimates of the treatment effect round off the picture.

Key words: correlated binomial average rule, empirical success rule, estimation, minimmax regret, treatment effect.

JEL classification numbers: C44, D81, C72, C90.

1 Introduction

Recently there has been a growing attention of regret and of minimax regret. This is a method for making decisions based on observations that does not rely on priors. The difference to classical hypothesis testing is that the value of a decision is taken into account. The statistician or decision maker does not only wish to learn which treatment is better but takes the value of implementing a treatment into account. This incorporation of values is captured by regret, the nonparametric or distribution-free component is contained in the minimax quantifier. One may say that minimax regret brings economics to statistics. This is because, superficially speaking, economics is about cost and benefit.

More specifically, consider someone who has to evaluate the results from testing a new treatment. The approach of hypothesis testing could be to test the null hypothesis that the new treatment is at most as good as the old versus the alternative hypothesis that the new treatment yields better results on average. The statistician is worried about wrongly rejecting the null hypothesis which means wrongly recommending the new treatment even if it is not better. So the statistician is very unhappy when accepting the new treatment if it likely that the old treatment is better, however small the margin may be (wrongly rejecting the null hypothesis). Minimax regret on the other hand can incorporate the cost of the new treatment. The statistician does not care which recommendation is made if the benefits, when taking cost into account, are very similar. Instead, focus is on not recommending the new treatment when the old was substantially better or vice versa, not keeping the old treatment when the new one is substantially better.

Regret goes back to Savage (1951) who was interpreting the ideas of Wald (1950) and has its foundations in the axioms of Milnor (1954). Recently there has been a revival of regret in statistics starting with Manski (2004). Manski (2004) considers a specific rule for evaluating treatments called the *empirical success rule* and uses this to investigate how one should deal with covariate information. Schlag (2006b) derives the minimax regret rule with and without covariate information, the rule is called the *binomial average rule*. The binomial average rule is a randomized rule and as such typically does not yield a unique recommendation. However it needs much less data to obtain the same performance as the empirical success rule.

We derive recommendations based on this methodology in four examples. The outset of this paper is to demonstrate the particular value of the binomial average rule. However, Karl Schlag recently also discovered an alternative randomized rule called the *correlated binomial average rule* which turns out to be superior in the applications. This alternative rule also attains minimax regret but typically involves less variance in the recommendation. We also demonstrate how to use these randomized rules to derive a deterministic recommendation at the cost of increasing maximal regret, a useful method if the recommendation is already close to being determinis-

tic. Everything is compared to the performance of the deterministic empirical success rule. Point estimates of the treatment effect round off the picture.

We have selected four examples from very different areas for this demonstration that have the common feature that sample sizes are small by nature of the application. We have a medical example in which computer equipment is tested as an alternative means for diagnosis. In the Indian school example two alternative policies are tested in order to improve learning of children. Our data on experimental game theory concerns a cross country comparison of generosity. Finally, in our fourth example we look at the design of institutions for achieving economic performance.[1]

The paper starts with an extensive informal presentation of treatment choice under minimax regret. We purposely refrain from as much jargon as possible to make this paper accessible to a wide audience from different disciplines. An important supplement of this paper is the matlab program we provide at the following adress <http://www.iue.it/Personal/Schlag/Welcome.html#research> to allow users to derive recommendations under the binomial average rule and the correlated binomial average rule on their own data sets.

The paper is organized as follows. We first present an extensive introduction to the methodology. Then we separately present the four examples. Finally we conclude.

2 Informal Introduction to Minimax Regret

In the following we give a brief summary of the method and underlying theory that is being applied in this paper to four specific examples. The approach itself is called *choice under minimax regret*. The presentation is kept as basic as possible to facilitate first time reading of this topic.

Consider a statistician who has to choose between two alternative options given a set of data sampled. For better illustration, consider a more specific scenario in which a medical statistician has to recommend one of two methods (or treatments) to doctors for examining their future patients. Let us call these method X and method Y. In a later example we compare manual location to computer-assisted navigation to determine where to apply a shock wave therapy against pain.

We organize this introduction along the following items:

1) sample and outcomes,
2) being a better method,
3) measuring outcomes and cost,
4) setting the outcome range,
5) hypothetical recommendation,
6) empirical success rule,

[1] These four data sets come from Sabeti-Aschraf (2005), Banerjee et al.(2005), Roth et al. (1991) and Acemoglu et al. (2001) respectively.

7) regret,
8) minimax regret,
9) binomial average rule - motivation,
10) transforming data,
11) binomial average rule in an example,
12) binomial average rule and minimax regret,
13) the empirical success rule.

1. Sample and Outcomes

The data or *sample* available to the statistician results from testing patients who are randomly drawn from the pool of potential patients (the procedure is also called a randomized experiment). On each patient one of the two methods has been tested and the *outcome* of each test recorded where the assignment of methods to patients was not systematic.

We first consider the setting in which each method has been tested on the same number of patients, the sample is called *balanced*.

The outcome can be in terms of success or failure of the treatment or it can be a more differentiated. One may want to measure the level such as the grade on a school exam or the change in level such as change in pain due to treatment.

2. Being a Better Method

Each method does not achieve the same outcome on each patient. As the doctor is assumed not to know a priori how any given patient responds we choose to evaluate each method according to the average outcome it generates. So we call one method *better* than the other if its average outcome is higher. This average refers to the average outcome that it would achieve if we would apply the method to all patients in the world. In this paper it is assumed that there is a large, essentially infinite number of potential patients. So the statistician will never actually know which method is better. Hence, his recommendation will go hand in hand with a measurement error that here will be formulated in terms of "regret" (more on this later).

3. Measuring Outcomes and Cost

In order to measure average outcomes we have to associate outcomes with numbers (values) where better outcomes receive higher numbers. Moreover, these numbers have to be such that higher averages are preferred without any concern for difference in the variation. The specific assignment of numbers results from how the statistician chooses to compare different outcomes.

Often outcomes are already presented in the appropriate format given by the underlying question. For instance, the statistician may be interested in the average school grade. A minor adjustment has to be made if lower average outcomes are preferred (such as less pain after the treatment). In this case one has to simply invert the scale. For instance if the original scale measures the degree x of pain on a scale

from 0 to 100, so x is the outcome of the treatment, then associate outcome x with the number $100 - x$, thus associating higher numbers with better outcomes (less pain).

In other cases we have to construct the numbers associated to outcomes from scratch. When there are only two outcomes such as success and failure then we can associate success with value 1 and failure with value 0. Typically however there are other events too like partial success or success with some side effects.

Method specific costs have to be incorporated into these numbers. Different methods may be associated to different costs. The simplest way to incorporate costs is to calculate a unit cost per patient, measured on the same scale as outcomes are, and subtract this cost from the numerical value of the outcome. So if method X is more costly than method Y then subtract a *unit cost* c from the numerical outcome of the method X.[2]

In the following we will assume that outcomes are directly specified in terms of such numbers.

4. Setting and Normalizing the Outcome Range

In order to apply the method of this paper the statistician must provide a range of outcomes that contains any outcome that can possibly occur. Let a be the lowest possible outcome and b the highest possible outcome, $a, b \in \mathbb{R}$. The statistician may not derive this range by looking at the data gathered or making some wild guess. Instead this range has to be a consequence of the setting itself. For the education example grades belong by definition to a fixed scale ranging from 0 to 100 so we set $a = 0$ and $b = 100$. In the medical example of this paper pain is measured on a scale $[0, 100]$ so if we wish to measure the change in pain then the range is given by $[-100, 100]$. If we also subtract a unit cost c from one of the two methods then the range becomes $[-c, 100]$ or $[-100 - c, 100]$ respectively.

Once the outcome range $[a, b]$ has been set it simplifies the presentation to normalize outcomes so that we only deal with ranges given by $[0, 1]$. In order not to change any of the further results it is important that this normalization is linear (actually, affine). Hence, normalize outcome y by transforming it into $\frac{y-a}{b-a}$.

5. Hypothetical Recommendation

As we mentioned at the outset, the statistician has to recommend which method to use, X or Y. Let EX be the true average (formally, expected) outcome of method X if method X would be applied infinitely often to some random patient (without the same patient being called twice) and let EY respectively be the true average outcome of Y. So if the statistician would know that $EX > EY$ then he would recommend

[2] A standard procedure in decision theory for constructing such numerical values when there is a best and a worst outcome is as follows. Assign value 1 to the best and 0 to the worst outcome. For any other outcome assign the value x such that you do not care if this outcome occurs or instead if you receive outcome 1 with probability x and outcome 0 with probability $1 - x$.

method X to the entire population. However the statistician does not know which method is better. He may have some hunch, however the method of this paper does not allow for incorporation of such a hunch. Instead it considers a statistician who has genuine uncertainty about which of the two methods is better.

6. Empirical Success Rule

A natural recommendation would be to recommend the method that achieved a higher average outcome in this data sample. This recommendation rule will be called the *empirical success rule*.[3] Let $\bar{X}$ and $\bar{Y}$ denote the respective average outcomes of each method observed in the sample. So method X is recommended if $\bar{X} > \bar{Y}$. If there is a tie then each method is recommended equally likely. This paper is about demonstrating the advantages of an alternative rule called the *binomial average rule*. Before we present it we have to describe how we compare rules. This is where regret comes in.

7. Regret

Regret is a measure of lost opportunity. It compares what you get with what would have been best if you knew the truth. If you choose method X but method Y is better so $EY > EX$ then the difference $EY - EX$ is called the regret. More generally, if $z = 1$ indicates that method X is recommended and $z = 0$ that method Y is recommended then regret r is defined by $r = \max\{EX, EY\} - zEX - (1 - z)EY$. In particular, regret is equal to 0 if the statistician has actually recommended the truly best method. Sometimes we refer to regret as the error of the recommendation. So if both methods are equally good then regret r is equal to 0.

8. Minimax Regret

Of course the statistician will never be able to calculate regret as he will never know the true average outcomes. Hence the statistician resorts to theory to find a recommendation for which it is known that regret is never too large. How is this done? The idea is that a worst case analysis is undergone. For each rule describing how the statistician makes a recommendation based on the data one derives the maximal possible regret without making any assumption on how effective each method is. Since the statistician only knows that each method will yield an outcome in $[0, 1]$ the maximum is taken over all possible methods with this property. Maximum regret is then taken as a measure of how good the recommendation rule is. The next step is to search for a rule that achieves the lowest such maximal regret. Theory shows that the binomial average rule will have this property, we say that it attains *minimax regret*.

Up to here this seems very mysterious. How should one protect oneself against the worst case without knowing anything? The statistician does know something as he has data to base his recommendation on. Let us assume the statistician observes

[3] Results relating to the empirical success rule are based on Manski (2004).

in the sample that method X was always very extremely successful and achieved outcome 1 each time while method Y only yielded failures in the form of outcome 0. Two different things he can conclude. It could be that method X is better than method Y. It can also be that method Y is better than method X, only the data was not representative. It just turned out that the few patients that do not react well to the method Y (that is better on average) were tested. However, the larger the data sample the more likely it is to have tested the methods on a representative sample and hence on average it seems like recommending the method that yielded higher outcomes in the sample is the right thing to do.

Consider now an alternative sample in which $\bar{X} = 0.7$ and $\bar{Y} = 0.67$ given bounds $[0, 1]$. Of course method X achieved a higher average in the sample than method Y but the difference is only very small, it seems like both are equally good. As regret is zero if both methods are equally good regardless of what is recommended the statistician is not worried about this case. Instead one has to think about whether X could be much better or Y could be much better. Since we have no strong evidence in either way, we can protect ourself by recommending each method approximately equally likely.

To make this last point clearer consider briefly the recommendation without any sample. The maximal regret of recommending method X is equal to 1, a regret arising when $EX = 0$ and $EY = 1$. On the other hand, the maximal regret of recommending each method equally likely is equal to $1/2$, this bound on regret is achieved when $EX = 0$ and $EY = 1$. With probability $1/2$ the worse method X is recommended which yields regret of 1 while with probability $1/2$ the better method Y is recommended which yields regret 0. On average (or in expectation), we obtain regret equal to $\frac{1}{2} * 1 + \frac{1}{2} * 0 = \frac{1}{2}$.

So on the one hand it is natural to recommend the method that achieved higher outcomes in the sample. On the other hand we want to take the differences in observations into account and recommend each method approximately equally likely when there is no strong evidence in favor of one of the two methods. This intuition is the first step to understanding that the empirical success rule might not be as good as it originally sounded. It disregards the magnitude of the observed differences between the two methods. Of course if the sample is very large then the empirical success rule will most likely pick up the better method due to the law of large numbers. However we have no indication of how large actually the sample has to be for this to be true. The strength of the methodology of this paper is that it already yields good results for very small samples.

9. Binomial Average Rule (BAR) - Motivation

The binomial average rule builds on the intuition that when outcomes are binary, so any outcome is either a success or a failure, then recommending the method that

was more successful seems the right thing to do.[4] Of course this also means that method X is recommended if $\bar{X} = 0.4$ and $\bar{Y} = 0.35$. The difference is very small. When we mentioned the problem of small differences we did not know where these differences came from. Were outcomes lower or were there fewer successes? How should one compare method X that always achieved outcome 0.4 to method Y that achieved a success in 35% of the tests and a failure in 65%? However, as we are now briefly considering the case where there are only two outcomes we are comparing one method that yielded a success 40% of the time to one that yielded a success 35% of the time. When there are only two outcomes then it is quite natural to recommend the method that yielded more successes as recommended by the empirical success rule. Theory underlines this intuition that actually the empirical success rule is the unique way to minimize maximum regret when outcomes are binary.

So is there a way to eliminate the issue of comparing many different outcomes? The idea underlying the binomial average rule is that one first transforms each outcome to create a new data set containing only the extreme values 0 and 1. After this transformation it is as if the statistician is facing binary outcomes only and hence can recommend the more successful method. It is important to keep in mind that more successful refers to outcome after this transformation, not in terms of the original data set.

10. Transforming Data

So how is this transformation done? What should we do if outcome 1/3 has been observed for one patient who received method Y. The idea is to replace this observation 1/3 by either 1 or 0 and to do this probabilistic. Replace it by 1 with probability 1/3 and by 0 with probability $1 - 1/3 = 2/3$. The procedure is to do this for each outcome in the sample, always using the outcome observed as the probability that it is replaced by 1. Notice that this transformation treats similar outcomes similarly, it respects levels as the level determines the probability of being replaced by 1.

At this point we have to step back a bit and discuss the role of randomization. A naive approach would be to transform all data in this way and then to recommend based on the transformed data. While this approach is not incorrect it does ignore the fact that the data transformation involved randomization. The data could have been transformed differently as the transformation was probabilitistic. To see this assume that all outcomes in the sample are in the interior of the range $(0, 1)$. Thus it is possible yet unlikely that the transformation results in method X having only $0s$ while method Y only has $1s$ in which case Y is recommended. The opposite is similarly possible, yielding a recommendation of method X. To include this multiplicity of recommendations explicitly the binomial average rule is defined by taking the expected recommendation based on this transformation procedure. The

[4]Results relating to the binomial average rule are due to Schlag (2006b).

computer program available from the authors online[5] simply does this transformation many times and takes the average recommendation. Thus the recommendation will typically be random, e.g. recommend method X with probability 0.8 and recommend method Y with probability 0.2.

A common objection is that a randomized recommendation (as typically made by the binomial average rule) is not very easy to implement. However this is not necessarily the case. One simply can recommend method X to 80% of the doctors and 20% to the remaining doctors.[6] For instance one need not convince all to use say the new method Y. However one is not allowed to let the patients choose as then the methods would not be assigned at random. For those seeking an alternative we describe below a way to create a deterministic (i.e. non randomized) recommendation at the cost of increasing the error.

It is important to point out that randomness is inherent to any recommendation based on data. Data itself has been gathered at random. Inviting a new set of patients and testing the two methods again will typically result in different outcomes.

11. Binomial Average Rule in an Example

Let us show how the binomial average rule is implemented in a simple example. Assume that two observations are available for each method. Say method X yielded 0 and 1 while method Y yielded 1/3 and 1/3. Since the data of X is already binary we need not transform it. The following table shows the outcome of the transformation of the data of Y together with the probability that this occurs and the consequent recommendation.

Table 1: Binomial Average Rule

Transformed data for Y	*Probability*	*Recommendation*
0 , 0	4/9	X
0 , 1	2/9	X and Y equally likely
1 , 0	2/9	X and Y equally likely
1 , 1	1/9	Y

Consequently, the method X is recommended with probability $\frac{4}{9} + \frac{2}{9}\frac{1}{2} + \frac{2}{9}\frac{1}{2} + 0 = \frac{2}{3}$ while method Y is recommended with probability $1 - \frac{2}{3} = \frac{1}{3}$.

12. Binomial Average Rule and Minimax Regret

It turns out the binomial average rule is the best one can do for a given sample size in the sense that it attains minimax regret. In other words, it guarantees the lowest value of maximal regret possible. If one prefers a different rule such as the empirical success rule because it is simpler and deterministic then one should take into account

[5] At http://www.iue.it/Personal/Schlag/Welcome.html#research

[6] To recommend methods to doctors or hospitals instead of letting each doctor randomize can be a way to ensure that the allocation of methods is truly random.

how much worse it performs. The binomial average rule provides a benchmark on how low maximal regret can be pushed for a given number of observations.

We present the lowest value of maximal regret, achievable by the binomial average rule. This value will be independent of the specific data gathered and only depend on the size of the sample. Of course the specific data will influence the recommendation but not the error in terms of maximal regret of this recommendation. While there is an exact expression for the value of minimax regret, the following simple formula is correct up to three decimals. Let n be the number of outcomes observed for each method. Then regret under the binomial average rule is bounded above by

$$\frac{0.17 * (b-a)}{\sqrt{2n-0.2}} \ . \tag{1}$$

There is no rule that can ensure a strictly lower value of maximal regret. So if $n = 6$ observations of each method are available then maximal regret is bounded above by $0.05 * (b-a)$, i.e. by 5% relative to the range of the outcomes.

13. The Empirical Success Rule

Unfortunately the maximal regret of the empirical success rule is not known. Simulations show that at least $n = 10$ observations of each method are needed to ensure maximal regret below 5%. Notice that simulations are not very trustworthy for deriving an upper bound on maximal regret given the large number of possible outcome distributions. An analytic upper bound for general sample size on regret under the empirical success rule is given by

$$\frac{b-a}{\sqrt{2n}} e^{-\frac{1}{2}} \tag{2}$$

without any indication of how good this bound is for a given number of observations.

Accordingly, one can only ensure a maximal regret of 5% under the empirical success rule by obtaining $n = 74$ observations of each method. If we compare the two bounds in (1) and (2) we find that more than 12 times the number of observations are needed (provided $n \geq 2$) to be sure that the empirical success rule achieves the same error or lower than the binomial average rule.

Gathering more data reduces the error. Notice however that this reduction is very slow, $n = 145$ observations of each method are needed at least to push maximal regret below 1% of the range when using the binomial average rule (the bound in (2) yields $n = 1840$ for the empirical success rule). As the error is reduced with more data gathered, the more likely the statistician will recommend the better method under either of the two rules. Using the language of statistics, both the binomial average rule and the empirical success rule are uniformly consistent estimators of the better method.

2.1 Comparison to Hypothesis Testing

At this point it is important to step back and briefly compare the minimax regret approach to the more standard methodology of hypothesis testing. We only present a flavor of the argument.

Should the statistician be worried about making a *bad* choice or about making the *wrong* choice? Under minimax regret the statistician is worried about making a bad choice where bad refers to the fact that there is a large difference between the average outcome of the better method and the average outcome of the chosen method. A bad choice means the statistician would have a large regret if he knew the truth. So bad does not refer to the size of the average outcome but to the comparison between own recommendation and the best possible recommendation if true outcome distributions were known.

On the other hand, in classical statistical hypothesis testing the statistician is worried about making the wrong choice, about saying that method X is better than Y while in reality method X is worse than method Y. To make the difference clearer, assume that $EX = 0.6$ and $EY = 0.57$ given range $[0, 1]$. Now assume that we repeatedly gather a sample of $2n$ observations and let the statistician make a recommendation. Assume that in 80% of the data samples the statistician recommends method Y. Then the statistician has made the wrong choice in 20% of the cases while the regret of his choice is very small as it is equal to $0.2 * 0.03 = 0.006$. So the classic statistician who uses hypothesis testing is very unhappy as 20% is typically not acceptable as a mistake while the statistician using minimax regret is quite happy, achieving an expected regret of only 0.6%.

We believe this example nicely shows how adding value to a choice and not only being worried about finding the best choice is very natural. Being worried with a bad choice means that the statistician only has to try to learn which method is best when the methods are quite different which means that these are situations in which it is easier to learn. Consequently, the minimax regret approach allows to achieve low levels of maximal regret already in small samples sizes (e.g. 6 observations of each method ensures maximum regret below 5%).

Notice that an additional difference between the two approaches is how to behave when there is no evidence that one method is better than the other. Hypothesis testing will recommend the status quo given by the null hypothesis while minimax regret behavior will recommend each method approximately equally likely.

2.2 Extensions

Given that the basics have been explained above we comment on the following modifications and extensions:

- what to do if we do not have the same number of observations of each method,
- an alternative rule that does better at putting more weight on one recommendation,
- an alternative to the empirical success rule for achieving a deterministic recommendation,
- what to do if the outcomes come in pairs where each method has been tested on the same source (e.g. patient), and
- estimating the treatment effect of each method and how the two methods X and Y differ.

2.2.1 Unbalanced Samples

So far we assumed that an equal number of data was observed for each of the two methods X and Y. The sample was *balanced*. However, often data is not balanced. In this case there are two options.

As the binomial average rule needs a balanced sample, one option is to simply drop randomly some of the data points of the method that was observed more often. Here it is important that the observations are dropped at random independent of the particular outcomes. So if method X was tested n_X times and Y tested n_Y times with $n_Y > n_X$ then $n_Y - n_X$ outcomes are dropped from method Y and then the binomial average rule is applied to yield maximal regret of

$$\frac{0.17 * (b-a)}{\sqrt{2\min\{n_X, n_Y\} - 0.2}}.$$

The computer program provided by the authors drops different data points in each loop of the simulation. This procedure does not attain minimax regret, below we investigate the induced increase in maximal regret.

The alternative is to derive the best rule for unbalanced samples in terms of minimizing maximum regret. The explicit rule is quite intricate and would unnecessarily complicate our presentation. When the sample is not too unbalanced, like in our examples, then the best rule is not that much better than our method described above of dropping observations. To see this we provide an analytic upper bound on how much regret can be further reduced by using the best rule for the unbalanced case without actually knowing the rule. Assume method X was tested n_X times and Y tested n_Y times with $n_Y > n_X$. Then the best method for the unbalanced sample can never be better than the best method when we add more observations of method X to obtain a balanced sample with n_Y observations of each method. Hence the value of minimax regret in the unbalanced sample is bounded below by

$$\frac{0.17 * (b-a)}{\sqrt{2n_Y - 0.2}}.$$

If

$$\frac{0.17 * (b-a)}{\sqrt{2n_X - 0.2}} - \frac{0.17 * (b-a)}{\sqrt{2n_Y - 0.2}}$$

is small then our more conservative approach of dropping observations to create a balanced sample is not that bad. In all our examples the samples are nearly balanced, the resulting errors of dropping observations negligible, e.g. if $n_X = 26$ and $n_Y = 30$ then the increase in maximal regret due to dropping four observations is bounded above by $0.0017 * (b-a)$.

The empirical success rule can also be applied to unbalanced samples. So here there is no need to drop data. The upper bound on regret can be adjusted to accommodate for the unbalancedness and is now given by

$$\frac{1}{2} e^{-\frac{1}{2}} \left(\frac{1}{n_X} + \frac{1}{n_Y} \right)^{\frac{1}{2}}.$$

It is useful to observe that this bound is identical to the one we gave above when $n_X = n_Y$.

2.2.2 The Correlated Binomial Average Rule (CBAR)

The binomial average rule was introduced as a rule that is able to attain minimax regret. However it is not the *unique* rule. In the following we present a variant we call the *correlated binomial average rule* that also attains minimax regret.[7]

The binomial average rule creates a lot of variance in the recommendation due to the fact that it transforms each data point independently into the most extreme outcomes. This transformation was done in order to create a rule that is more sensitive to levels. However, the disadvantage for practitioners is that this rule tends to recommend each method with some probability. In the following we present a rule that creates less variance when data is transformed which tends to reduce the degree of randomization in the recommendation itself. E.g. in the education example below the new program is recommended under the binomial average rule in 73% of the schools. This number is increased to 90% under the correlated binomial average rule.

For the analytic proof that the binomial average rule attains minimax regret it is very important that the transformation does not change the expected value of each observation. However, the binomial average also assumes that this transformation is done independently for each data point. There is no need for this. It is natural not to add correlation between different observations of the same method as the entire strength of the data lies in the independence of the different data points underlying

[7] This is the first appearance of this rule in the literature.

the same method. However, nothing per se should speak against correlating the transformation across data from different methods. This is what the correlated binomial average rule does.

The correlated binomial average rule is implemented as follows. First the data is randomly matched in pairs, one observation for each method. Now consider one such pair (y_1, y_2). Pick randomly a number in $[0, 1]$ where each number is chosen equally likely. Say z has been chosen. Then transform any outcome y_i in this pair into 0 if $y_i < z$ and transform it into 1 if $y_i \geq z$, $i = 1, 2$. Apply this method independently to each pair and then proceed as under the binomial average rule (and hence as in the empirical success rule as now all observations are binary).

For illustration how this works assume that there is only a single observation of each method, assume X yielded 1/4 while Y yielded 1/3. We illustrate the transformation for the correlated method as well as for the original method in the table below.[8]

Table 2: Correlated Binomial Average Rule

Transformed Data for X and Y	*Probability under Correlated Binomial Average Rule*	*Probability under Binomial Average Rule*	*Recommendation*
0 , 0	2/3	1/2	X and Y equally likely
0 , 1	1/12	1/4	Y
1 , 0	0	1/6	X
1 , 1	1/4	1/12	X and Y equally likely

It is easily argued that the correlated binomial average rule also attains minimax regret. In practical examples we find that it shifts the recommendation away from the mid point, thus yielding a recommendation with lower variance.

Little more is known about how the correlated binomial average rule compares to the binomial average rule. We make two observations.

1. If the outcomes underlying the two methods take only the maximum and the minimum value in the range then the two rules yield the same recommendation as the underlying transformations do not change the data.

2. It is well known for minimax regret rules that any probabilistic combination of two such rules will also attain minimax regret. The reasoning is as follows. Since each rule attains the same value of maximal regret on its own, the maximal regret obtained by mixing between the two rules cannot be strictly above this maximal value. Consequently, by combining the recommendations of the

[8] Notice that whenever there are only two data points, as in this example, both rules make the same average recommendation. Here X is recommended with probability 11/24.

binomial average rule and of the correlated binomial average rule we get a confidence interval for minimax regret recommendation. However it is still an open question whether this interval is tight in the sense that any combination outside will not be a minimax regret recommendation.

2.2.3 Deterministic Recommendation

We acknowledge the fact that sometimes it is simply not feasible to have a probabilistic recommendation. One solution is to use the empirical success rule at the cost of the 12 fold larger sample size needed to guarantee the same error. An alternative approach is to adjust the recommendation of the correlated binomial average rule at the cost of increasing maximal regret. Assume that the recommendation is 95% on method X and 5% on method Y, associated to maximal regret of 2% in terms of the range of the outcomes. Then one can instead recommend method X to all but this means that maximal regret is increased by 5% to 7%.

The idea is as follows. If method X is recommended to all then in 5% of the cases, namely when the minimax regret solution is to recommend method Y, then you do not do what you are supposed to do. The maximal regret due to recommending something different is equal to the range of the interval of outcomes. Hence, in 5% of the cases the increase is maximaly $(b-a)$.

Formally, let p_X be the probability of recommending method X given by a rule that attains minimax regret r^* (in percent of range). Then

$$\begin{aligned} \max\{EX, EY\} - EX &= \max\{EX, EY\} - p_X EX - p_Y EY + p_Y (EY - EX) \\ &\leq r^* (b-a) + p_Y (b-a). \end{aligned}$$

2.2.4 Paired Experiments

Sometimes the data is gathered by testing each method on the same patient, possibly at different times. When data comes in pairs where each observation contains two outcomes, one for each method, we speak of a *paired experiment*. This situation arises naturally when comparing the outcome before and after some treatment, for instance a treatment aimed at reducing pain. So method X could correspond to the state of the patient before the treatment and Y to the state after the treatment. Method X is better if the treatment is not effective (on average). Method Y is better if the treatment is effective. Notice that there are two observations for the same patient, hence observations are not independent. The specific characteristics of the patient will influence both outcomes simultaneously. Our approach so far assumed independence of all observations. So what should we do?

An interesting theoretical insight for paired experiments is that we can eliminate one observation for each patient without increasing maximal regret. We only have to

make sure that the sample is balanced after this elimination. We are throwing away data it is inconsequential for maximal regret.

One only has to be careful when computing the value of minimax regret in paired examples. If there are 60 observations resulting from 30 patients then after elimination we obtain 15 independent observations of each method which yields a maximal regret of $\frac{0.17}{\sqrt{2*15-0.2}} \approx 3.1\%$. If the number of patients is odd, say 25, then a balanced sample of 12 independent observations can only be achieved. The computer program provided at ?? has an option for dealing with paired experiments.

3 Point Estimation

We remind the reader that the minimax regret approach described above is aimed at choosing a method that achieves the highest true average outcome. However, this approach does not provide insight either on how large the average outcome of the method chosen will be nor on the difference between the average outcomes of the two methods. In the following we show how one can estimate these two values.

Any such estimate is of course prone to some error and the estimate will depend on how the error is measured. The classic approach is to punish wrong estimates by the square of the difference between the estimate and the true value (quadratic loss), taking the average (or expectation) over all possible data sequences that can be generated. Notice that quadratic loss punishes estimates that are very wrong. With this measure of error one then proceeds as above, choosing for each estimate the maximum error and then finding the estimate that minimizes this maximum error. An estimate that results from this procedure is said to attain *minimax risk*.[9]

A common approach in statistics is to confine attention to estimates that are unbiased. Accordingly, one searches for the estimate that attains minimax risk among all unbiased estimates. An estimate is called *unbiased* if on average it equals the true value, otherwise it is called *biased*. Unbiasedness would be the obvious feature of choice if one would make an estimate on infinitely many different samples. As the statistician is facing a single sample, adding unbiasedness will typically increase the error as he has less estimates to choose from. This contradicts his aim to minimize the maximum error. Hence it is important to quantify the loss due to considering only unbiased estimates.

3.1 Average Outcome

Consider a sample of n independent observations of method X. Then the sample average $\bar{X}$ attains minimax risk among all unbiased estimates of EX.[10] If one instead

[9] Results on biased estimates in this section are based on (Hodges and Lehmann, 1950).

[10] This is proven in Schlag (2006a).

does not restrict attention to unbiased estimators then it is natural that the range of the variable will also play a role. If the sample average $\bar{X}$ is close to one of the borders then given that very bad estimates are punished severely under quadratic loss one is tempted to shift the estimate towards the median of the range. Of course this bias should decline as the sample size increases. This is actually what happens for the estimate that attains minimax risk, the estimate of EX being

$$\frac{1}{1+\sqrt{n}}\left(\sqrt{n}\bar{X}+\frac{1}{2}(a+b)\right). \tag{3a}$$

Note how the bias to the mid point $\frac{1}{2}(a+b)$ decreases in the sample size, for $n \geq 16$ it is always less than 10% of the range, for $n \geq 81$ it is always less than 5% of the range. So there is a non negligible bias for sample sizes that are not too large.

3.2 Average Difference between Outcomes

Consider a balanced sample of $2n$ independent observations of method X and method Y, so n of each. We are now interested in estimating how different the two methods are in terms of true average outcomes. So we are now interested in estimating $EX - EY$. While $\bar{X} - \bar{Y}$ remains an unbiased estimator of $EX - EY$ its properties in terms of whether it is best are not known. However we can present the biased estimate of $EX - EY$ that attains minimax risk, it is given by

$$\frac{\sqrt{2n}}{1+\sqrt{2n}}\left(\bar{X}-\bar{Y}\right). \tag{4}$$

Notice that the difference between the two biased estimators of EX and EY is not equal to the biased estimator of the difference $EX - EY$. Again the bias of the estimate is non negligible for small sample sizes, e.g. the sample difference is scaled down by more than 10% when $n \leq 40$.

3.3 Choice and Estimate

Can it happen that method Y is chosen with positive probability but method X is estimated to be better? Yes. The binomial average rule will typically be random and hence each method will be recommended with positive probability. On the other hand, the estimate for the difference is deterministic. Is this a conflict? No, as the objectives are different, estimating which method is best versus estimating the magnitude of the difference between the two methods. In fact, there is a more consistent way of solving both objectives by selecting a randomized estimate of $EX - EY$.[11]

[11] See Schlag (2006a).

4 Medical Data

We consider the following data collected from a medical randomized study (Sabeti-Ashraf et al. 2005) aimed at comparing two diagnosis methods. Fifty patients participated in a study trial and were randomly divided into two balanced groups. All patients were suffering from shoulder tendinitis.[12] Before and after the treatment sessions, all patients were clinically tested with the visual analog scale (VAS) to measure their pain level. Both groups were exposed to the same low-energy shock wave treatment 3 times in weekly intervals during 12 weeks. An important aspect of the shock wave treatment process is the precision of the location of the shock wave focus point. In group 1 (*feedback group*), the patient and the therapist determined the location point by palpation (manual location) while in group 2 (*navigation group*), a radiographically guided 3-D, computer-assisted navigation system was used.

We consider a decision maker concerned about reducing the level of pain reported by the patient where pain is measured by VAS. We address two questions: 1) Does shock wave therapy reduce pain? 2) Is there a difference between the two methods for locating the shock wave focus point? VAS 1 and VAS 2 measure the pain before and after the treatment. The following table give the descriptive statistics of our sample.

Table 3: Descriptive Statistics

Outcome	*Mean*	*St.Dev*
Feedback Group (25 patients)		
VAS 1	68	15
VAS 2	33	20
VAS 2 - VAS 1	-35	23
Navigation Group (25 patients)		
VAS 1	66	22
VAS 2	18	21
VAS 2 - VAS 1	-48	28

VAS 1 measures the pain before the shock wave treatment.
VAS 2 measures the pain after treatment.
Both Measures range from 0 (no pain) to 100.

We examine the first question. We use an inverted scale in order to make higher outcomes better for the patient. The outcome of shock therapy method is recorded by $100 - VAS\ 2$. No treatment is associated to the alternative method, the outcome is recorded by $100 - VAS\ 1$. The statistician has to recommend one of these two methods, treatment or no treatment.

[12] More specifically, patients all suffered from calcific tendinitis of the suprasinatus tendon.

This is a paired experiment since every patient reports both VAS 1 and VAS 2. While outcomes are independent across patients they are not so across methods. Therefore, as discussed above, we need to randomize over our sample in such a way that for each patient we use only one of the two values VAS 1 or VAS 2, putting equally many patients in each category.

We separately investigate each group. We do not provide an estimate using the empirical success rule as the associated upper bound on regret is equal to 12%.

Table 4: Effect of Shock Wave Therapy

Decision Rules	*N*	*Minimax regret in VAS scale*	*Treatment probability (bin. average)*	*Treatment probability (corr. bin. average)*
Probability of choosing Shock Wave Therapy				
Feedback group	25	3.3	0.97	0.99
Navigation group	25	3.3	≈ 1	≈ 1

No recommendation can be made that ensures regret less than 3.3 on the VAS scale. The lowest value of maximal regret can be sustained by recommending shock wave therapy by using palpation on 99% of the patients or by using the computer assisted diagnosis on practically 100% of the patients (a more precise estimate is 99.8%).

As the probability for the recommendation to treat using palpation is so close to the border, we can instead recommend the treatment to all, increasing maximal regret by $(1 - 0.99) * 100 = 1$ on the VAS scale .

Thus, we find that in either group the decision maker can recommend shock therapy to all patients and achieve a maximal regret of at most 4.0 on the VAS scale.[13] Shock therapy works.

Notice that the point estimate in the change in pain in the feedback group is equal to $\frac{1}{1+\sqrt{25}}\left(\sqrt{25}\,(-35) + \frac{1}{2} * 100\right) \approx -21$ while in the navigation group it is equal to -32.

Next we turn to the second question and examine how the shock wave focus point should be located. Now one method is associated to the feedback and the other to navigation. We do this in two scenarios. First, outcomes are measured by the level of not having pain after the treatment $100 - VAS\ 2$. Second, outcomes are measured in terms of change in pain $VAS\ 1 - VAS\ 2$. The results for the recommendations are reported below:

[13] $\left(\frac{0.17}{\sqrt{0.8+25}} + 0.0062\right) * 100 \approx 3.97$.

Table 5: Choice of the Location Method

Decision rules	*N*	*Minimax regret in VAS scale*	*Probability of navigation (bin. average)*	*Probability of navigation (corr. bin. average)*	*ESR**
Navigation versus Feedback					
(100 - VAS 2)	50	2.4	0.92	0.97	1(8.6)
VAS 1 -VAS 2	50	4.8	0.7	0.82	1(17.2)

**For the empirical success rule the upper bound on the maximal regret is given in brackets.*

The value of minimax regret given 50 observations equals $\frac{0.17}{\sqrt{50-0.2}} \approx 0.024$ in percentage of the range. For the analysis of pain after the treatment this means approximately 2.4 points on the VAS scale as the range is equal to $[0, 100]$. Regarding change in pain we obtain maximal regret equal to $200 * 0.024 = 4.8$ on the same scale as the outcome range now equals $[-100, 100]$. The larger range naturally makes the recommendation less extreme, now only recommending that 82% of the patients should use computer assisted navigation.

We can recommend the computer assisted navigation to all patients at a maximal regret of $2.4 + 0.03 * 100 = 5.4$ on the VAS scale.

We calculate the point estimates of the difference between the success of the navigation and of the feedback group. The navigation group is estimated to perform better than the feedback group, in terms of levels by $\frac{\sqrt{50}}{1+\sqrt{50}}(33 - 18) \approx 13$ points while in terms of change in pain by 11 points on the VAS scale.

We summarize:

1. We find that both location procedures are effective in reducing pain. Maximal regret of this recommendation is 4 points on the VAS scale. This is despite the fact that this statement is only based on 25 observations as we do not pool the data. Pain reduction was substantial, the point estimate being 21 points (feedback) and 32 points (navigation) on the VAS scale.

2. For the comparison between methods we have double the sample size but the difference was not so pronounced.

   When only looking at the absolute pain levels after the treatment we find large success of the computer assisted navigation. It can be recommended to all at 4.8 maximal regret on the VAS scale.

   When interested in change of well being and hence in change in pain, then the increased outcome range makes the recommendation less pronounced. Now only 82% should use computer assisted navigation at a maximal regret of 4.8 points on VAS scale.

5 Remedying Education Data

Two randomized experiments designed to assess means to improve education were conducted in major Indian cities (Baroda and Bombay) between 2001 and 2003 (Banerjee et al., 2005). First, a remedial education program (BAL) was implemented. A tutor was assigned to teach basic literacy and numeracy skills to children lagging behind in public schools.[14] Second, a computer-assisted learning (CAL) program was run. Two hours of shared computer time per week were provided to 4th grade students (two students per computer).

During the experiment, which lasted two years, all children from 3rd and 4th grade were given a test at the beginning of the year (pre-test) to assess their language and math skills. Then, some schools were randomly assigned to receive the programs, and children were again exposed to a test at the end of the year (post-test). While in the BAL treatment only a minority of students within each treated schools received a tutor, all students were exposed to the computer assisted learning program in the CAL treatment.[15] Finally, while assistance was provided both for language and math skills in the BAL program, the CAL program was instead focused on developing mathematical skills only.[16] The annual cost per student for the tutorial was 2.25$ while for the computer assisted learning it was 15$.

We illustrate how one can draw conclusions based on this randomized experiment. We select Baroda 4th grade students during the second year of the experiment (2002-2003). We choose this subset of students because they constitute the only cohort for which both programs were implemented, and hence the only cohort that allows us to compare the two programs. Our random sample is composed of 111 schools and it is structured as follows:

Table 6: Indian Data Structure

Number of Schools	*CAL (Computer)*	*No CAL*
BAL (Tutorial)	28 (group 1)	26 (group 2)
No BAL	27 (group 3)	30 (group 4)

We present the descriptive statistics together with the point estimate of the treatment effect of each group 1-3 compared to the untreated group 4. We do this separately for post-test grades and for the difference between post-test and pre-test grades.

[14] The program was labelled "*Balsakhi*", meaning "child friend".

[15] For the BAL treatment, only those students lagging behind were selected according to their pre-test score. They were then removed from classroom during two hours per week to receive tutorial assistance.

[16] Two hours of shared computer time per week (two students per computer) were provided under the supervision of an instructor. Students were exposed to a variety of computer games designed to emphasize basic comptencies in the mathematic curriculum.

Table 7: Post-Test Grades (range $[0, 100]$)

	Mean Score	*St.Dev*	*Point Estimate of Effect*
Group 1 (BAL+CAL)	58	9	10
Group 2 (BAL)	55	11	7.2
Group 3 (CAL)	49	7.8	2.5
Group 4	47	10	–

Table 8: Difference between Post- and Pre-Test Grades (range $[-100, 100]$)

	Mean Difference	*St.Dev*	Point Estimate of Effect
Group 1 (BAL+CAL)	24	7.3	8.5
Group 2 (BAL)	21	10	5.6
Group 3 (CAL)	19	7.5	3.5
Group 4	15	8.8	–

In the following we investigate the impact of the BAL and CAL programs. Motivated both by the lower costs and the higher estimated effect of the tutorial program as compared to computer assisted learning we consider the choice whether or to introduce the tutorial program.

Notice that we do not have a balanced sample. We accomodate for this by repeatedly randomly dropping observations from the larger sample. In this case, as we are investigating the effect of the BAL treatment (group 2 versus group 4), we create a balanced sample by repeatedly randomly dropping 4 observations from group 4.

The table below shows that maximal regret of 2.4 can be ensured by introducing it to 90% of the schools (resulting from the correlated binomial average rule). However this recommendation is based on post test grades only. If we consider the change in performance then the increased range raises maximal regret and makes the recommendation less pronounced. Maximal regret in terms of grade points is now equal to 4.7, implemented by the recommendation to introduce tutorials to 74% of the schools (again using CBAR).

Next we assume that tutorial program has been introduced and decide whether or not to implement the computer treatment. Here we find that 72% should also receive the computers while the remaining 28% should only receive the tutorials.

Table 9: Decision Rules

Decision Rules	*N*	*Minimax Regret in Points*	*Treatment Probability (Bin. Average)*	*Treatment Probability (Corr. Bin. Avg)*	*ESR**
Tutorial Program (BAL) vs None					
- level	52	2.4	0.73	0.9	1(8.4)
- differences	52	4.7	0.59	0.74	1(17)
Tutorial and Computer Program (BAL + CAL) vs Tutorial Program only (BAL)					
- levels	52	2.4	$0.6^{\#}$	$0.72^{\#}$	1(8.4)
- differences	52	4.7	$0.56^{\#}$	$0.62^{\#}$	1(17)

** For the empirical success rule, upper bound on maximal regret given in brackets.*
 # Probability of recommending both treatments.

As these programs are costly and the recommendation for either package is not that pronounced we add a cost c of implementing the computer treatment. While we know the monetary cost per student of the computer program, the variable c has to be measured in units of grade points. Below we compare different values for c.

Again we compare implementing both treatments to implementing only the tutorials. We consider evidence based on post-test grades only due to the initial high levels of maximal regret for the analysis of change in performance. Notice that this cost increases the range to $[-c, 100]$.

Consider first the empirical success rule. Looking at the table above we see that if $c \geq 3$ then only the tutorials will be recommended to all. The associated upper bound on maximal regret in terms of grades is equal to $0.084 * (100 + c)$.

Consider now the correlated binomial average rule. We find that $c = 20$ yields 0.017 probability of recommending both treatments. So not recommending both yields maximal regret in terms of grade points of $(0.024 + 0.017) * 120 = 4.92$. For the same cost but using the binomial average rule we obtain 0.16 weight on both treatments which is too large to be able to replace the random recommendation by a deterministic one and still ensure sensible levels of maximal regret.

To summarize, only minimal cost will lead to dismissal of the computer program with maximal regret being bounded above by 9 grade points. This is reached by evaluating the empirical success rule. Lower maximal regret can be achieved with the correlated binomial average rule, however only substantial costs will result in completely abandoning the computer program. This reflects the hetergeneous recommendation instrinsic in the random nature of the correlated binomial average rule.

6 Four Nations play the Ultimatum Game

Roth et. al. (1991) tested experimentally the bargaining behavior of subjects from four different countries (Israel, Japan, USA and Yugoslavia) using the Ultimatum Game. The Ultimatum Game is a simple game in which two players have to agree on how to share a given amount of money T. In this game, one of the players (the proposer) proposes how much of this money x the other player (responder) should receive, keeping the rest $T - x$ to himself. Given this proposal the responder has to decide whether to accept or to reject this proposal. If the responder accepts then each player gets the amount resulting from the proposal, so the proposer gets $T - x$ while the responder gets x. If the respondent rejects the proposal then both players get no money.

In the experiment, the participants played this game anonymously ten times with the same partner via computers. The proposer remained proposer throughout and similarly the responder remained responder throughout.

Consider how selfish players interested only in their own payoffs should play the Ultimatum Game. It is in the best interest of the responder to accept any offer $x > 0$ from the proposer. Knowing this the proposer should only offer very little to the responder as she knows that the responder will accept anything. The prediction of Game Theory is for the proposer to offer 0 and the respondent to accept any offer.

As reported by Roth et. al. (1991) and many other experimental papers that deal with the Ultimatum Game, this prediction is almost never fulfilled by the agents. Typically around half the money is offered to the responder. The new aspect is that Roth et. al. (1991) found that there are differences across countries in the way players behave. However, due to small numbers they could not statistically compare mean proposals.

We do not statistically compare proposals either. Instead, we apply the methodolgy of this paper and formulate the country comparison in terms of a choice problem. If you want to select a country in which mean proposals are highest, i.e. where subjects are most generous, which country would you choose?

To asses this question, we make pairwise comparisons between countries selecting only proposals in the first round. Each of the two selected countries is associated to a method and the outcome measured is the proposal x.

We first present the descriptive data.

Table 10: Round 1 Data

	Israel	*Yugoslavia*	*Japan*	*USA*
Average Proposal on scale 0,..,1000	363	442	446	447
Standard Deviation	157	85.5	211	95.7

Consider the recommendation under the empirical success rule. USA has the highest mean proposal and hence is selected in any pairwise comparison. The regret

of this recommendation is bounded above by around $78 - 81$ units on the proposal scale. This does not seem to be a very effective recommendation in light of the fact that the point estimates for the difference between Israel and any other country is around 70 while for any pair not including Israel it is less than 4.4.

Tables 11 and 12 report the probability of choosing the row country under the binomial average rule and under the correlated binomial average rule. Maximal regret was either 0.022 or 0.023.

Table 11: Pairwise Comparison Using Binomial Average Rule

Prob. of choosing the row country	*Israel*	*Yugoslavia*	*Japan*	*USA*
Israel (30)	x	0.26	0.24	0.26
Yugoslavia (30)	0.74	x	0.5	0.49
Japan (29)	0.76	0.5	x	0.5
USA (27)	0.74	0.51	0.5	x

Notice that we choose sample size $N = 59$ for comparing any pairs not containing USA while we choose $N = 55$ when comparing USA with others.

Table 12: Pairwise Comparison Using Correlated Binomial Average Rule

Prob. of choosing the row country	*Israel*	*Yugoslavia*	*Japan*	*USA*
Israel	x	0.094	0.1	0.11
Yugoslavia	0.9	x	0.46	0.47
Japan	0.9	0.54	x	0.53
USA	0.89	0.53	0.47	x

In both cases we find a very similar estimate of the comparison of Israel with any other country (0.25 or 0.1) while choices are very close to 50% among the pairwise comparisons of Japan, Yugoslavia and USA.

The recommendation under both the binomial average rule and the correlated binomial average rule identifies two groups of countries which already shows up when looking at the descriptive statistics. On the other hand, the empirical success rule identifies a unique choice irrespective of the variance of the data.

7 Colonization

Following Acemoglu et al. (2001) we ask whether the large cross-country differences in income per capita (as of 1995) can be explained by differences in the quality of institutions. Institution quality is measured by how good property rights are

protected.[17] The problem in determining the impact of institutions on growth is that there is also a reverse causal effect, economic growth can also influence institutions.

To get around this issue we follow the authors and include in our investigation a third variable. Ideally this so-called instrumental variable should influence the quality of the institution but should not directly influence economic growth. In this paper the chosen instrumental variable is the mortality rate faced by European settlers between the 17th and the 19th century. The underlying argument is that the mortality rates faced by European settlers in early colonization times influences institutions set up then and carried over to today but clearly will not directly determine economic growth today.

Our objective is to investigate the influence of institution quality on economic growth. We formulate this in a choice problem using the third variable of mortality. We run a comparison between countries with high mortality and with low mortality as follows. The sample of 64 countries is separated above and below the median of the mortality rates faced by European settlers, defining the "high-mortality" method versus the "low mortality" method. We then consider a decision maker choosing between these two methods according to minimax regret. We consider two cases, first using economic growth as outcome, then using institution quality as outcome.

If high quality institutions tend to induce better economic performance then we should find similar recommendations in terms of mortality method for inducing high quality institution ad for inducing high economic growth.

In this application there are no exogenous bounds on log GDP. We assess an upper bound by considering the logarithm of twice the largest GDP among those countries that were not colonized (Germany). As lower bound we set log GDP to 1.

First we present the descriptive statistics.

Table 13: Descriptive Statistrics

	Mean Risk	*Std Dev Risk*	*Mean Log GDP*	*Std Dev log GDP*
Low Mortality	7.097	1.42	8.618	0.990
High Mortality	5.934	1.28	7.483	0.764

This means that low mortality would be chosen to maximize expected risk or to maximize log GDP under the empirical success rule with an associated maximal regret in both cases bounded above by 7.6% in terms of the range.

[17] The index comes from the Political Risk Services, and it ranges from 0 (low protection against expropriation) to 10 (high protection against expropriation).

Table 14: Decision Rules

Outcome	*N*	*Minimax Regret in Percentage*	*Treatment Probability (Bin. Average)*	*Treatment Probability (Corr. Bin. Avg)*
Log GDP	32	2.13%	0.8302	0.9726
Institution	32	2.13%	0.8439	0.9683

The table gives the probability of choosing countries with low mortality

Whether the outcome of interest is income or institutions, according to the correlated binomial average rule the decision maker will choose the "low mortality" countries with a very high probability. For the given sample size of $N = 32$, these decision rules yield an upper bound on regret of 2.13%. Moreover, if low mortality is chosen with certainty then maximal regret in terms of institution is bounded above by $0.0213 + 1 - 0.9683 = 0.053$ and on log GDP by 0.049 (in terms of range).

We consider a decision maker who tries to uncover the relationship between institution on the one side and high income on the other by comparing countries with low and high mortality. The result is that choosing a random country with low mortality is a good way to both select a high quality institution and to achieve high income with maximal regret bounded above by 5.3%. This gives a nonparametric indication that good economic performance can be achieved with good institutions. Hopefully this thought experiment will trigger more research investigating minimax regret when there is an issue of causality.

8 Conclusion

There has been a growing theoretical interest in minimax regret treatment choice with recent results presenting minimax regret rules. The main value of this paper lies in the demonstration of how to bring these rules to the data.

In addition we present new methodologies for the analysis. The first and important innovation is the introduction of the correlated binomial average rule that is based on a very similar transformation of the data as the binomial average rule. This alternative minimax regret rule involves less variance in the data transformation which typically leads to "less random" recommendations. The second innovation concerns how to create deterministic recommendations. Recommendations that are almost deterministic are easily transformed into deterministic ones, increasing maximal regret by the weight on the treatment chosen less often.

Finally, Hodges and Lehmann (1950) estimates are included to round off the picture. These are useful as minimax regret rules give no indication of the magnitude of treatment response.

References

[1] Acemoglu, D., Johnson, S. and J.A. Robinson (2001), "Colonial Origins of Comparative Development: An Empirical Investigation," *Amer. Econ. Rev.* **91(5)**, 1369-1401.

[2] Banerjee, A., S. Cole, E. Duflo and L. Linden (2005), "Remedying Eduction: Evidence frmo Two Randomized Experiments in India," NBER Working Paper **11904**.

[3] Hodges, J.L.Jr. and E.L. Lehmann (1950), "Some Problems in Minimax Point Estimation," *Ann. Math. Stat.* **21(2)**, 182-197.

[4] Manski, C. (2004), "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica* **72**(**4**), 1221-1246.

[5] Milnor, J. (1954), Games Against Nature. In Decision Processes, ed. R.M. Thrall, C.H. Coombs & R.L. Davis. New York: John Wiley & Sons.

[6] Roth, A., Prasnikar, V. , Okuno-Fujiwara, M. and S. Zamir (1991), "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh and Tokyo: An Experimental Study", *Amer. Econ. Rev.* **81(5)**, 1068-1095.

[7] Sabeti-Aschraf, M., Dorotka, R., Goll, A. and K. Trieb, (2005), "Extracorporeal Shock Wave Therapy in the Treatment of Calcific Tendinitis of the Rotator Cuff," *Amer. J. Sports. Med.* **33(9)**, 1-4.

[8] Savage, L. J. (1951), "The Theory of Statistical Decision," *J. Amer. Stat. Assoc.* **46(253)**, 55-67.

[9] Schlag, K.H. (2006a), *Designing Non-Parametric Estimates and Tests for Means*, European University Institute, Mimeo.

[10] Schlag, K.H. (2006b), *Eleven - Tests needed for a Recommendation*, European University Institute Working Paper ECO **2006-2**, January 17.

[11] Wald, A. (1950), *Statistical decision functions*, New York: John Wiley & Sons.