An insurance policy against unexpected data

Ivo W. Molenaar

Vakgroep Statistiek en Meettheorie, FSW
University of Groningen.

Eerder verschenen in de reeks:

Heymans Bulletins Psychologische Instituten R.U. Groningen
HB-79-447-EX

Voordracht voor de European Meeting
of the Psychometric society Groningen 1980

Trefwoorden: Bayesiaanse statistiek
gemengde a priori verdelingen

Summary

In a Bayesian analysis of the probability parameter Π of a binomial distribution, a prior distribution for Π would be combined with the sample data consisting of X=x successes in n independent trials. The choice of an informative prior of the natural conjugate (beta) form implies in most cases that some values of X have a very low predictive probability. The present paper argues that an investigator who would obtain such an unlikely value of X will feel inclined to distrust either his data or his informative prior. If the data are reliable, he may be led to an after-the-fact distrust of his own prior specification rather than to a revision according to Bayes' theorem.

This paper explores a strategy in which the original beta prior is replaced by a mixture in which this beta distribution, with a strong weight, is mixed with an uninformative beta distribution, with a small weight. Using this mixture prior, the predictive distribution for X no longer contains extremely unlikely values. The posterior distribution for Π given X, on the other hand, is hardly changed as long as X assumes values with moderate or high predictive probability under the original prior.

The mixture strategy thus acts like an insurance policy, in the sense that a small premium is paid in normal circumstances (a slight bias of the posterior estimate) in exchange for a large gain in the unlikely case of a sharp conflict between data and prior. Two reasons for adopting the strategy can be distinguished, in which the same algebra leads to a somewhat different interpretation. First, the investigator may be convinced that the mixture, with its heavier tails, is a better representation of his personal prior ideas than even the best fitting member of the class of beta densities. Second, when the best beta well represents his ideas, he may still want to protect the analysis against unforeseen data values sharply contradicting those ideas.

1. Introduction

It is seldom stressed that virtually all statistical procedures are conditional:"if assumptions A, B, C are correct, then the data D lead to conclusion Z". As is well known, a procedure which is optimal under such assumptions may perform badly when they are violated. Relevant amendments are found under keywords such as robustness, outlier detection, pretesting, testimators, distribution-free tests or adaptive inference. Among the reasons for the limited popularity of such alternatives are tradition, availability, appreciation of elegant and simple procedures, a firm belief that serious violations of the standard assumptions are rare, and overestimation of the loss of efficiency involved in the use of alternative procedures.

Many researchers seem to be unwilling to sacrifice a little under ideal circumstances, even when a lot could be thus gained in a less favorable situation. And yet, that is precisely what the present paper asks them to do, in the special context of prior specification for Bayesian inference. Contrary to what an occasional optimist might believe, assumptions like independent observations, homoscedasticity and a particular type of distribution, are generally just as vital to a Bayesian statistical analysis as they are to a frequentist one. At most we can hope that the use of prior information diminishes the impact of outliers etc. on the final results of a Bayesian analysis. This diminished impact has e.g. been demonstrated where the analysis of a small sample has been improved by a Bayesian method using samples from similar though not identical populations; examples are the extension of Kelley's formula for true score estimation (Novick & Jackson (1974, p.308-322))and m-group regression (Novick, Jackson, Thayer & Cole (1972); Molenaar & Lewis (1979)).

It has seldom been explicitly discussed, however, that prior information is only beneficial when there are no gross errors in its specification. Indeed Bayesian statistics has an additional robustness problem to the ones discussed above, namely: how sensitive are the conclusions to the choice of the prior? For dichotomous conclusions, Vijn (1980) introduces the "robustness region" consisting of all values of the hyperparameters governing the prior distribution, for which the conclusion remains unaltered. Here a different approach will be chosen: the prior that the investigator has specified will be mixed with a component indicating (almost)total uncertainty. This second component will serve as our "safety belt", or "insurance company" in the case of a violent collision

between the prior and the data. The adaptive characteristic of the proposed procedure will ensure that the posterior weight of the non-informative prior component will grow when the discrepancy between the data and the informative prior component increases. When the data agree with the informative prior, however, the weight of this informative component assumes a value close to one, and little precision is lost. The whole procedure thus resembles the payment of a small insurance premium against a rare but enormous loss.

2. The beta-binomial case: example of a dilemma

Let X be the number of successes in a series of n independent trials with known n and constant but unknown success probability Π. Our model density or likelihood function is thus binomial:

$$(2.1) \qquad P(X=x|\Pi) = \binom{n}{x} \Pi^x (1-\Pi)^{n-x} \qquad (x=0,1,\ldots,n) \ .$$

The natural conjugate prior density for the parameter Π is the beta distribution with parameters a and b, say, and with probability density

$$(2.2) \qquad f(\pi) = \frac{1}{B(a,b)} \pi^{a-1}(1-\pi)^{b-1} \qquad (0 < \pi < 1).$$

Here $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ and the gamma function satisfies $\Gamma(x+1) = x\Gamma(x)$. Formula (2.2) can describe rather different forms of prior beliefs by suitable choices of a and b. This flexibility is combined with mathematical elegance: the posterior distribution of Π after observing X=x is again beta with parameters a+x and b+n-x, and the prior predictive distribution for X is the beta-binomial or Polya distribution with parameters n, a, and b.

In a given research situation, a and b should now be chosen such that they formalize the investigator's prior knowledge on Π. This problem has been tackled by offering him bets (e.g. Raiffa, 1968) or by an interactive and adaptive questioning sequence embodied in a computer package (e.g. CADA as described in Isaacs & Novick, 1978).

The source of the prior knowledge may be previous random samples, data for similar experiments or unspecified beliefs; all we need here is that the investigator decides at a certain moment that his ideas on Π are well represented by a certain beta distribution, say with parameters a and b.

Example. The author interrogated himself on the fraction Π of female students among the roughly 17,000 students enrolled at the University of Groningen in the academic year 1979/1980. Seated behind a terminal, he used the CADA computer package to specify and reconsider his ideas about the quartiles, highest density regions, mode and graph of the density for Π until he felt satisfied that a beta distribution with $a = 20$ and $b = 40$ well represented his prior information.

He then confrontated himself with the question of how he would react if he was told that a random sample of $n = 100$ students had been drawn and had been found to contain only $X = 10$ females. The sample estimate for Π would thus be $X/n = 0.10$.

In the predictive Polya (n, a, b) distribution for X based on the prior, the event $X \leq 10$ has a probability of 0.001. In short, there is a violent discrepancy between the prior and the data.

What kind of action would I undertake, facing the collision between data and prior knowledge? The formal strategy, discussed first, would be to stay strictly within the Bayesian framework. My posterior density for Π given $X = 10$ out of $n = 100$ and a beta (20,40) prior is a beta (30, 130) distribution; it has mean 0.19, standard deviation 0.03,and 90 percent highest density region ranging from 0.14 to 0.24. If I felt enough worries I might try to get hold of another sample of 100 registered students. Intuitively I might be inclined to predict that this second sample contains either roughly 10 percent females (in which case my prior was too optimistic about emancipation) or just over 30 percent (in which case my first sample was a stroke of very bad luck).

Formally however, Bayes' theorem forces me to use a predictive distribution for the second sample in which the event $15 \leq X \leq 25$ has a probability of 0.71 (Polya distribution with n=100, a=30, b=130).

Informally, the data-prior collision makes me feel very tempted to solve my cognitive dissonance by an after-the-fact distrust of either data or prior. Let us criticise the data first. I might forward the suggestion that the data were drawn from a different universe (e.g., excluding the university's teachers' colleges, where female participation is high) than I envisaged when choosing my a=20 and b=40 prior. Alternatively, the registration of the sex may have contained clerical errors, or the sampling procedure may have been biased.

Such afterthoughts on the quality of the data are not unusual in actual research. It would be desirable, though difficult, to develop a formal statistical model dealing with the risks and error rates of a procedure in which unexpected outcomes lead to a thorough checking of the sampling,measurement and data processing, which extra check would be omitted when no surprising results would be found. Such a development would be outside the scope of this paper, which will follow the traditional assumption that the data (10 females among 100 students in our case) were carefully collected.

What remains, then, is a temptation to solve the data-prior collision by disavowing the beta (20, 40) prior. By "disavow" more is meant than "revise according to Bayes' theorem", because as was mentioned above, this had little intuitive appeal. My personal experience is that I suddenly recalled the slight uneasiness that I had felt at the completion of the interactive session leading to my choice for a=20 and b=40. I said at that time that I was satisfied with this choice. And yet, why did I pretend that I was so sure? Perhaps my impression about female university enrollment was biased by my familiarity with the Social Science students? And even knowing that fewer female students register for Natural Sciences, I have only rather vague ideas about the proportion of students that those two subgroups contribute to the whole student body. My ideas about female student registration for

Medicine or Arts may be based on the situation of several years ago. Come to think of it, there are lots of reasons why I should regret having been so specific about my prior beliefs.

3. A beta mixture as an insurance policy

The following proposal is made not only for the specific example discussed above, but for any situation in which an investigator has just decided that the beta distribution with parameters a and b best fits his prior information on an unknown binomial parameter Π. Choose a number λ $(0 < \lambda < 1)$ and another set of beta parameters c, d and consider the prior distribution which is a mixture of the two beta distributions:

(3.1) $\Pi \triangleq \lambda \text{beta}(a, b) + (1 - \lambda)\ \text{beta}(c, d)$,

where the symbol $\triangleq$ denotes "is distributed as". In our female student enrollment example, I decided to choose $\lambda = 0.80$, $c = d = 2$, keeping $a = 20$ and $b = 40$ as parameters of the first component.

The mixture prior for this case is graphed in figure 1.

---

figure 1 see page 71

This mixture is almost as peaked at 0.33 as the first component, but attributes a little more weight to values of Π close to 0 and 1. Within the beta family itself, the only way to allow for more uncertainty would be to diminish a + b, keeping a/b constant. The dotted line in figure 1 shows that this is less effective as an insurance policy than the mixture: if both have the same mean and variance, then the mixture has a higher kurtosis, cf. Molenaar and van Zwet (1966), and it was precisely this concentration of probability both in the peak and in the fat tails that mimicked my uncertainty better than the single beta with lower peak, high "shoulders" and still very little probability in the far tails.

The mixture with $\lambda = 0.95$, $c = d = 1$, graphed in the same figure, will turn out to give even more protection against outlying values of X. I dismissed it, however, because its prior density $f(\pi)$ and its predictive density $p_M^*(x)$ given in Table 1 were too far from my prior ideas. See also the discussion following Table 1.

We shall return to the aspect of mimicking uncertainty in section 6. First, however, we shall discuss the posterior distribution of Π, produced by combining the prior mixture (3.1) with the observation of X=x successes in n independent trials. For that discussion, it is more convenient to temporarily use the notation

$$(3.2) \qquad a_1=a,\ b_1=b,\ \lambda_1=\lambda,\ a_2=c,\ b_2=d,\ \lambda_2=1-\lambda.$$

Let $f_i(\pi|x)$ denote the beta $(a_i+x,\ b_i+n-x)$ posterior density that would be obtained by combining the i-th component of the prior with X=x:

$$(3.3) \qquad f_i(\pi|x) = \frac{1}{B(a_i+x,\ b_i+n-x)}\,\pi^{a_i+x-1}(1-\pi)^{b_i+n-x-1}, \qquad (0<\pi<1;\ i=1,2)$$

and let $p_i(x)$ denote the Polya (n, a_i, b_i) predictive density for X using the i-th component as a prior:

$$(3.4) \qquad p_i(x) = \binom{n}{x} B(a_i+x,\ b_i+n-x)/B(a_i,\ b_i) \qquad (x=0,1,\ldots,n;\ i=1,2).$$

The joint density of Π and X using the mixture prior (3.1) is

$$(3.5) \qquad g_M(\pi,x) = \sum_{i=1}^{2} \lambda_i \binom{n}{x} \pi^x(1-\pi)^{n-x}\,\pi^{a_i-1}(1-\pi)^{b_i-1}/B(a_i,b_i) =$$
$$= \sum_{i=1}^{2} \lambda_i\, p_i(x)\, f_i(\pi|x).$$

The conditional density of Π given X=x, which is the posterior based on the mixture prior, now equals

$$(3.6) \qquad f_M(\pi|x) = \frac{g_M(\pi,x)}{P(X=x)} = \sum_{i=1}^{2} w_i(x)\, f_i(\pi|x),$$

a mixture of the two component posteriors with weights

$$(3.7) \qquad w_i(x) \stackrel{\text{def}}{=\!=} \lambda_i\, p_i(x)/\{\lambda_1\, p_1(x) + \lambda_2\, p_2(x)\}.$$

The ratio of the two weights is thus $\{\lambda_1\, p_1(x)\} / \{\lambda_2\, p_2(x)\}$, which clearly shows the adaptive nature of the procedure: when a value X=x is observed which is very improbable under the first prior but not under the second, then the weight of the first component is decreased compared to λ_1, and vice versa.

This is illustrated in Table 1. Although only some selected values of x are listed, it is clear that the posterior weight $w_1(x)$ becomes small for all values which have a low predictive probability $p_1(x)$. Also tabled are the posterior means

$$(3.8) \qquad \mu_i(\Pi|x) = \int_0^1 \pi f_i(\pi|x)\, d\pi = (a_i+x)/(a_i+b_i+n) \quad (i=1,2)$$

for both components, and the posterior mean for the mixture, which is

$$(3.9) \qquad \mu_M(\Pi|x) = \int_0^1 \pi f_M(\pi|x)\, d\pi = \sum_{i=1}^{2} w_i(x)\, \mu_i(\Pi|x)\ .$$

It follows from a complete version of Table 1 that for $20 \leq X \leq 46$, which has a predictive probability under p_1 of 0.92, the posterior means μ_M and μ_1 differ by 0.01 or less.

table 1 see page 72

A next point illustrated in the table is that the predictive density

$$(3.10) \qquad p_M(x) = \sum_{i=1}^{2} \lambda_i\, p_i(x)$$

based on the mixture (3.1) does not assign extremely low probabilites

to very small or very large values of x, as does p_1.

This property holds even more strongly for the mixture containing 0.05 times the uniform as a second component; its predictive denstiy is denoted by $p_M^*(x)$ in Table 1. For such a mixture it is easily seen that $p_M^*(x) \approx 0.05\ p_2(x) = 0.05$ for all values of x for which $p_1(x)$ is negligible. As announced earlier, this may not meet the needs of an investigator who feels that the best fitting beta distribution has lighter tails than his personal prior.

Section 4 will elaborate the idea that replacement of the first component by a mixture does little harm as long as the data are in

harmony with the beta (a,b) prior. Sections 5 and 6 investigate how much can be gained by this replacement in the case of unexpected outcomes.

4. A small loss when no catastrophe occurs

In this section it is assumed that the investigator wants to estimate Π with a quadratic loss function, and that the true prior for Π is beta (a,b). The terminology "true prior" will be evident in situations where the unknown proportion Π is itself generated by some random mechanism. It is also defensible when the prior is a summary of an investigator's prior knowledge and/or prior beliefs, however: a "wrong" prior may be the one entertained by another investigator (adversary statistics, cf. Novick & Jackson, 1974, p.148), or the investigator himself might consider how much difference it would make if he replaced the original prior by the mixture.

Postponing the discussion of losses under other priors until the next sections, we shall now evaluate the expected losses of various estimators under the assumption of a beta (a,b) prior distribution for Π.

Any estimator, $\hat{\pi}$ say, has an expected loss given X=x of

$$(4.1) \qquad \int_0^1 \{\pi - \hat{\pi}\}^2 \, f_1(\pi|x) \, d\pi \quad .$$

This loss given x is clearly minimized by taking for $\hat{\pi}$ the posterior mean

$$(4.2) \qquad \mu_1(\Pi|x) = \int_0^1 \pi \, f_1(\pi|x) \, d\pi = (a+x)/(a+b+n) \; .$$

The loss of this quadratic estimator, which is unbiased, equals the variance

$$(4.3) \qquad \sigma_1^2(\Pi|x) = (a+x)(b+n-x)(a+b+n)^{-2}(a+b+n+1)^{-1},$$

by the standard expression for the variance of a beta distribution.

Averaged across x, which has the Polya (n,a,b) distribution $p_1(x)$ given by (3.4) in the notation (3.2), we shall denote this loss by AVE_1 VAR_1, where the first index tells us that the averaging is carried out assuming that the first component distribution beta (a,b) is the true distribution of Π, and the second index tells that the variance is also based on this first component distribution. For the evaluation of AVE_1 VAR_1 we multiply (4.3) by $p_1(x)$ obtained from (3.4) and rewritten by (3.2) in the original notation; next four out of the five factors occurring in (4.3) can be drawn into the beta function by using the line below (2.2). The result contains the sum of all probabilities of a Polya (n, a+1, b+1) distribution, and the result of the algebra is that

$$(4.4) \qquad AVE_1\ VAR_1 = \sum_{x=0}^{n} p_1(x)\ \sigma_1^2(\Pi|x) =$$

$$= \sum_{x=0}^{n} \binom{n}{x} \frac{B(a+x+1,\ b+n-x+1)}{B(a+1,\ b+1)} \frac{B(a+1,\ b+1)}{B(a,b)} \frac{1}{a+b+n} =$$

$$= \frac{ab}{(a+b)(a+b+1)(a+b+n)} .$$

If any other estimator $\hat{\pi}$ is used than the posterior mean (4.2), we may evaluate (4.1) by writing $\pi - \hat{\pi} = \pi - \mu_1(\Pi|x) + \mu_1(\Pi|x) - \hat{\pi}$. This leads to the usual split of the square, in which the cross product term gives no contribution because the second difference does not depend on π. The loss of any estimate $\hat{\pi}$ given X=x thus equals

$$(4.5) \qquad \int_0^1 \{\pi - \hat{\pi}\}^2\ f_1(\pi|x)\ d\pi = \sigma_1^2(\Pi|x) + \{\mu_1(\Pi|x) - \hat{\pi}\}^2 .$$

Averaged across x the loss for $\hat{\pi}$ thus is the sum of the minimal loss AVE_1 VAR_1 and the averaged squared bias to be denoted as $AVE_1 SQ\ BIAS_{\hat{\pi}}$; this parallels the well known sampling theory result on the mean squared error although our interpretation is quite different.

Before we turn to our mixture estimator $\mu_M(\Pi|x)$, let us consider two simple competitors. A non-Bayesian would use the sample fraction X/n as his estimator $\hat{\pi}$. Still assuming that the beta (a,b) distribution

is the valid prior, this means compared to the use of (4.2) an unnecessary additional loss of

$$(4.6)\qquad \mathrm{AVE}_1\ \mathrm{SQ.BIAS}_{X/n} = \sum_{x=0}^{n} \left\{ \frac{a+x}{a+b+n} - \frac{x}{n} \right\}^2 p_1(x) =$$

$$= \sum_{x=0}^{n} \left\{ \frac{a+b}{a+b+n} \right\}^2 \left\{ \frac{a}{a+b} - \frac{x}{n} \right\}^2 p_1(x) = \frac{a+b}{n}\ \mathrm{AVE}_1\ \mathrm{VAR}_1;$$

here it has been used that when X has the Polya distribution $p_1(x)$ then X/n has mean a/(a+b) and variance $\frac{ab(a+b+n)}{n\,(a+b+1)(a+b)^2}$. The sum above is thus a constant times this variance, which leads after a little algebra to the desired result.

The second competitor is the prior mean a/(a+b). An investigator neglecting the data and using this estimator suffers, compared to the use of (4.2), an unnecessary additional loss of

$$(4.7)\qquad \mathrm{AVE}_1\mathrm{SQ.BIAS}_{a/(a+b)} = \sum_{x=0}^{n} \left\{ \frac{a+x}{a+b+n} - \frac{a}{a+b} \right\}^2 p_1(x) =$$

$$= \sum_{x=0}^{n} \left\{ \frac{n}{a+b+n} \right\}^2 \left\{ \frac{x}{n} - \frac{a}{a+b} \right\}^2 p_1(x) = \frac{n}{a+b}\ \mathrm{AVE}_1.\mathrm{VAR}_1.$$

The results (4.6) and (4.7) clearly show how the unnecessary extra losses are related to the sample size n and the fictitious sample size a+b with which the prior knowledge is equivalent (see e.g. Novick & Jackson, 1974, p.126). These two quantities determine the weights that are attributed to X/n and a/(a+b) in the optimal estimator (4.2), which can be written in the form $\frac{a+b}{a+b+n}\,\frac{a}{a+b} + \frac{n}{a+b+n}\,\frac{x}{n}$ showing the compromise between prior and data. Results if this type are fairly common; for their use in philisophical inference models cf. e.g. Kuipers (1978, p. 56), on generalized Carnapian systems.

Let us now return to the effect of our mixture strategy on the posterior estimation of Π, still under the assumption that the beta (a,b) prior is the true one. This strategy implies the use of the posterior mixture mean $\mu_M(\Pi|x)$ defined in (3.9). This involves an unnecessary additional loss of

$$(4.8)\qquad \mathrm{AVE}_1\mathrm{SQ.BIAS}_M = \sum_{x=0}^{n} \{\mu_1(\Pi|x) - \mu_M(\Pi|x)\}^2\, p_1(x).$$

The expression between curly brackets can be rewritten as $w_2(x)\{\mu_1(\Pi|x) - \mu_2(\Pi|x)\}$, but substitution does not lead to a sum that can be written in an easy closed form, like above. We shall therefore have to use other means to analyze (4.8).

The claim to be substantiated is that the insurance premium is small, i.e. that the additional loss inflicted by the use of μ_M in a situation when μ_1 would be optimal is small, i.e. that the averaged squared bias is small compared to the averaged variance. Numerical support for this claim is found in Table 2. This is confirmed by analytical considerations, not worked out here, based on the idea that

a specific value of x gives rise to a sizeable contribution to the averaged squared bias only if simultaneously

(i) $p_1(x)$ is non-negligible,

(ii) $w_2(x)$ is non-negligible,

(iii) $(a+x)/(a+b+n)$ differs substantially from $\mu_2(\Pi|x) = (c+x)/(c+d+n)$.

Luckily it turns out that the values of x for which the two fractions in (iii) are far apart are in most cases ruled out by the influence of condition (i) for x/n far from a/(a+b) and condition (ii) for intermediate values. Exceptions occur mainly when n is small compared to a+b (details will not be given here).

Table 2 gives a condensation of our calculations for some 80 different parameter combinations all having $\lambda = 0.8$. It seems that the relative extra loss, defined as

$$(4.10) \qquad REL_1 = REL_1(a,b,c,d,\lambda,n) = \frac{AVE_1 SQ.BIAS_M}{AVE_1 VAR_1}$$

is typically less than 0.1 for $(a+b)/n \leq 1$, and 0.05 or less for $(a+b)/n \leq 0.5$, unless the ratios $a/(a+b)$ and $c/(c+d)$ differ by as much as 1/10 differs from 1/2, in which case the values of REL_1 can be twice as large. One could then decrease REL_1 by choosing $c < d$ for $a << b$; this is illustrated in the bottom part of Table 2 and it will be briefly discussed in section 7.

Table 2 see page 73

Very large relative losses, due to extreme forms of bias, obviously occur when n is very small compared to a+b: the combination of rather specific prior knowledge and only little data is a counter-indication against paying insurance premiums.

As a footrule one might infer from the complete dataset condensed in Table 2 that, for c=d=2 and $\lambda = 0.8$,

$REL_1 \approx 0.06\ (a+b)/n$ for $a/(a+b) = 1/2$,

$REL_1 \approx 0.08\ (a+b)/n$ for $a/(a+b) = 1/3$ or $2/3$,

$REL_1 \approx 0.18\ (a+b)/n$ for $a/(a+b) = 1/10$ or $9/10$.

Although deviations from this linear relationship occur, and seem to be growing with (a+b)/n, this footrule gives at least the order of magnitude of the relative extra loss REL_1 incurred by using the mixture prior when the original beta (a,b) prior was correct.

Compared to the use of X/n, for which the equivalent of REL_1 would be (a+b)/n by (4.6), the mixture is certainly very efficient under the conditions studied here. Table 2 leads us to the conclusion that the avaraged squared bias in using $\mu_M(\Pi|x)$ is small, both in absolute sense and relative to the average variance which is the minimum loss, as long as (a+b)/n is 1 or less; this becomes even more true when n=10 is ruled out on the grounds of being an unrealistically small sample size.

The last two columns of Table 2, marked by stars, again give the corresponding results for the mixture with $\lambda = 0.95$ and $c = d = 1$. It can be seen that REL_1^* is typically 1/3 of REL_1, or even less: mixing with 0.05 times a uniform prior has a very small effect on the quadratic loss estimation of π when the true prior is beta (a,b). The only non-negligible effects seem to occur when n is much less than a + b. The footrule

$$REL_1^* \approx .02\ \frac{a+b}{n}$$

seems to work well for $c = d = 1$ and $\lambda = 0.95$, although it is a little optimistic when a/b differs strongly from 1.

5. Losses under the mixture prior

In section 2 an example was given of a violent collision between prior and data. Let us now investigate whether our "insurance policy" of replacing the beta (a,b) prior by the mixture offers protection against such collisions.

When Π really has a beta (a,b) distribution, collisions are rare by definition, and unavoidable when they occur: each random variable on rare occasions assumes very improbable values, and all statistical analyses can be rather misleading when such a rare phenomenon happens. The occurrence of values for X which are very improbable under the predictive distribution $p_1(x)$ given in (3.4) however, can also be viewed as an indication that the beta (a,b) prior was incorrect. When assessing the benefits of the mixture strategy, we shall assume in this section that not beta (a,b) but the mixture (3.1) is the true prior distribution of Π, and investigate the quadratic loss incurred by an investigator who continues to use beta (a,b) instead.

The best estimate for Π of course is $\mu_M(\Pi|x)$ defined by (3.9). It is under our new assumption unbiased, and its loss given X=x thus equals the posterior variance

$$(5.1) \qquad \sigma_M^2(\Pi|x) = \sum_{i=1}^{2} w_i(x)\, \sigma_i^2(\Pi|x) + \sum_{i=1}^{2} w_i(x)\, \{\mu_i(\Pi|x) - \mu_M(\Pi|x)\}^2.$$

Averaged across x this becomes

$$(5.2) \qquad \mathrm{AVE}_M \mathrm{VAR}_M = \sum_{x=0}^{n} p_M(x)\, \sigma_M^2(\Pi|x),$$

where $p_M(x)$ given in (3.10) is the predictive distribution for X based on the mixture (3.1).

By a reasoning analogous to that leading to (4.5), the additional loss, averaged across x, of using $\mu_1(\Pi|x)$ instead of $\mu_M(\Pi|x)$ equals

$$(5.3) \qquad \mathrm{AVE}_M \mathrm{SQUARED\ BIAS}_1 = \sum_{x=0}^{n} p_M(x)\, \{\mu_M(\Pi|x) - \mu_1(\Pi|x)\}^2 .$$

Let us investigate (5.2), (5.3) and their ratio, the relative extra loss

$$(5.4) \qquad \mathrm{REL}_M = \frac{\mathrm{AVE}_M \ \mathrm{SQ.BIAS}_1}{\mathrm{AVE}_M \mathrm{VAR}_M}$$

We may re-examine conditions (i) (ii) (iii) stated in section 4, but (i) is now replaced by the condition that $p_M(x)$ is non-negligible. This, however, occurs for far more values of x, as was discussed in section 3 and as follows from the expression (5.1) for its variance. We are led to expect a somewhat larger minimum loss than in section 4, but a much larger average squared bias.

The numerical results confirm this. Table 3 shows that REL_M is typically much larger than REL_1, for the case $\lambda = 0.8$, c=d=2. To the right of the vertical bar the table describes the $\lambda = 0.95$, c=d=1 case, again denoted by stars: here REL_M^* is much larger than REL_1^*. Low values of REL_M and REL_M^* occur only when a and b are both small, and not too different from each other: in such cases the mixture resembles the original beta (a,b) prior. The table does not show the new minimum loss $AVE_M VAR_M$. It is usually between 1.3 and 2 times $AVE_1 VAR_1$ which was given in Table 2, and almost equal to $AVE_1 VAR_1$ in the starred case.

6. Gains in case of a catastrophe

The conclusion from section 4 and 5 is that unless a+b exceeds n there is usually only little extra loss in using μ_M when μ_1 would be optimal, and quite a bit more loss when μ_1 is used in a typical mixture situation.

Our investigator, however, may tell us that this is not a convincing argument in favor of the mixture. He will reason that he considers the beta (a,b) prior as far more probable, indeed it is his best choice of a beta prior. Why then sacrifice it, just for the reason that if ever some other state of affairs were true, using μ_1 would lead to substantial biases?

Indeed the use of the quadratic loss function, very defensible in section 4, is not adequate when it comes to assessing the possible gains of the mixture strategy. We want to be insured against the embarrassment caused by a serious conflict between the data and the beta (a,b) prior. How do we measure the embarrassment caused by observing a value that the experimenter considered as very unlikely? Postponing the point that all individual probabilities tend to be small when n is large, we could roughly say for a person with predictive distribution $p_1(x)$ that an observation with

$p_1(x) = 0.01$ may cause a mild surprise, $p_1(x) = 0.001$ may be a source of worries, and $p_1(x)$ of 0.0001 or less may be viewed as very serious. This is purposively stated in such vague terms, because the amount of distrust in either data or prior would depend very much on external factors that are not incorporated in our model.

For our operationalization of embarrassment, we stipulate that our measure should go up when $p_1(x)$ decreases, but not linearly: a logarithmic scale is more adequate. One could even argue in favor of a threshold due to the fact that probabilities of 10^{-5} and 10^{-7} lead to the same embarrassment. For mathematical simplicity and in order to avoid the arbitrary choice of such a threshold value, we shall use $-\log p_1(x)$ as the embarrassment caused by observing x.

If p_1 is indeed the true distribution, our expected (minimum) embarrassment is thus $-\sum_{x=0}^{n} p_1(x) \log p_1(x)$. If our predictive distribution is p_M, and p_1 is true, we thus have an additional expected embarrassment of

$$(6.1) \qquad \text{AVE}_1 \text{ EMB}_M = \sum_{x=0}^{n} p_1(x) \log \{p_1(x)/p_M(x)\}.$$

The usefulness of the information discrimination statistic for our purpose depends very much on its asymmetry: when p_M describes the true state of affairs then someone predicting according to p_1 has an additional expected embarrassment of

$$(6.2) \qquad \text{AVE}_M \text{ EMB}_1 = \sum_{x=0}^{n} p_M(x) \log \{p_M(x)/p_1(x)\} ,$$

which is typically much larger in our application, because the combination of a small p_M and a large p_1 is rare, but a very small p_1 frequently accompanies a p_M that is moderate or large. The last columns of Table 1 provide a numerical illustration; there one obtains that (6.1) equals 0.099 but (6.2) equals 0.833 for $\lambda = 0.8$, c=d=2; for the $\lambda = 0.95$, c=d=1 case these numbers become 0.012 and 0.052 respectively. Values of (6.1) and (6.2) are displayed in Table 3 for various value combinations for a,b,n: in terms of embarrassment much is lost when prior 1 is used and the mixture is

true, and much less is lost when the reverse holds. Note, however, that (6.1) and (6.2) also seem to grow with the sample size.

In order to give the reader some insight in the meaning of the numerical values of $AVE_M\ EMB_1$, the eighth column of Table 3 gives

(6.3) $\quad p_M\{p_1(X) \leq 0.00015\}$;

the probability, evaluated under p_M, of obtaining some value of X for which the predictive density $p_1(X)$ would be 0.00015 or less. Of course embarrassment should not be a constant for any $p_1(X) \leq 0.00015$ and zero for all X with larger predictive density, and the bound itself is arbitrarily chosen. Still it may help to see that such a probability is below 0.05 when (6.2) is below 0.5, while it is below 0.10 if (6.2) does not exceed one. The highest values of (6.3) occur predominantly with the highest value of (6.2), but there are now some violations of monotonicity.

We conclude that our mixture is successful in preventing or diminishing embarrassment caused by the very low probabilities of some extreme values of X. Let us repeat, however, that the investigator should start by a critical evaluation of the data gathering and processing which might provide a clue for such an extreme value.

7. Discussion

This is a first draft of a model incorporating doubt about the validity of the model itself. The style may have irritated the reader: the author switched on many occasions between the role of an investigator and the role of someone looking over the investigator's shoulder. It was already observed that the mixture prior can be viewed as an improved prior, curing the light-tailedness of the beta that best fits the middle region, but also as an insurance policy in which someone willfully accepts a misrepresentation of his/her prior beliefs in order to secure protection against unexpected outcomes. The second approach places itself outside of the strict Bayesian context in which "the" prior should be revised by the data into "the" posterior; the present author doubts whether this normative view continues to be a fair description of what a serious investigator would do in cases of surprising outcomes.

The tentative character of this paper lies not only in this interpretative dilemma. Let us mention a few other points deserving further attention.

The weight of the first component was taken to be either 0.8 with c=d=2 or 0.95 with c=d=1. If the experimenter is firmly convinced that values of X very close to 0 or n are extremely unlikely, he might try c=d=3 for the second component instead of c=d=2. A trial for three parameter combinations suggested that the premium as operationalized by REL_1 does not decrease substantially, whereas less is won in cases where the mixture prior is true. Similarly , a few trials with $\lambda = 0.9$ and c=d=2 did not look very promising: the insurance premium is almost halved compared to $\lambda = 0.8$, but gains in preventing loss or embarassment decrease more than proportionally.

With a minor exception in section 4 (p.13-14), the original and the mixture strategy have only been compared in situations where one of them is correct. Their behavior under other circumstances should be studied is some detail. That will also shed more light on the question whether it is useful to replace c=d by $c < d$ when $a \ll b$, as was done in the last lines of Table 2.

Point estimation with quadratic loss is certainly not the only possible end product of a beta-binomial analysis. When highest density regions for Π are desired, the analysis is complicated by the fact that an investigator using a wrong prior mis-specifies not only the centering but also the width of the interval. This interval will thus not only have a wrong probability content, but also miss some values on one side which have a higher true posterior density than some wrongly admitted values on the other side. Is there a loss function that satisfactorily compares two such intervals?

This paper has little to offer to the Bayesian who firmly believes that his original beta (a,b) distribution is absolutely correct, and should go on combining it with the sample via Bayes' theorem. It also has little value for a sampling theorist who believes that any decision should be based on the sample result alone. He will, however, get beaten by better predictions from Bayesian analysis in all cases where the prior information that he chooses to neglect is somewhat substantial and somewhat reliable.

What remain, then, are analysts who have some reliable substantive knowledge, but who feel that the best beta distribution does not include the long tails that mimick their conviction that outlying sample values may occur more frequently than indicated by the Polya predictive distribution that flows from the chosen beta prior. In a pilot experiment jointly with Charles Lewis, the author has confronted six subjects, who had just finished deciding on a beta prior for a well specified problem, with a sample result which was an outlier in this sense. Most subjects had some doubts on the validity of the sample result itself, but all of them expressed embarrassment and also seemed convinced that a straightforward application of Bayes' theorem would not be the best thing to do. The present paper hopefully is a first step toward the development of a useful alternative strategy.

Hofstee (1977, 1980) has advanced his betting paradigm for scientific investigation: a scientific statement is essentially a prediction about the outcome of an experiment, and could be formulated as a betting offer on this outcome. In that context, the present author offers a bet that investigators using mixture priors will win more bets than investigators using beta priors, let alone investigators using no priors at all.

Acknowledgment

Several readers of a draft version provided valuable suggestions on the presentation of this material. The tentative character of its methods and results implies that the author would be very grateful for more suggestions or criticisms.

References

Hofstee, W.K.B., De weddenschap als methodologisch model, Ned. Tijdschrift voor de Psychologie, 32 (1977) 203-217.

Hofstee, W.K.B., De empirische discussie, Meppel, Boom, 1980.

Kuipers, T.A.F., Studies in inductive probability and rational expectation, thesis, University of Groningen, 1980.

References

Isaacs, G.I. & Novick, M.R., Manual for the Computer-Assisted Data Analysis monitor, 1978, University of Iowa, Dept. of Educ.Statistics.

Molenaar, I.W. & Lewis, C., Bayesian m-group regression: A survey and an improved model, Methoden en Data Nieuwsbrief, SWS/Ver. voor Statistiek, 4, nr.1, 1979, pp.62-72.

Molenaar, I.W. & Van Zwet, W.R., On mixtures of distributions, Ann. of Math.Statistics, 37, 1966, pp.281-283.

Novick, M.R. & Jackson, P.H., Statistical methods for educational and psychological research, New York etc.: McGraw-Hill, 1974.

Novick, M.R. , Jackson, P.H., Thayer, D.T. & Cole, N.S., Estimating multiple regressions in m-groups: a cross-validation study. The British Journal of Mathematical and Statistical Psychology, 1972, 25, pp.30-50.

Raiffa, H., Decision analysis: Introductory lectures on choices under uncertainty, Reading, Massachusetts, Addison-Wesley, 1968.

Vijn, P., Prior Information in Linear Models, thesis University of Groningen, 1980.

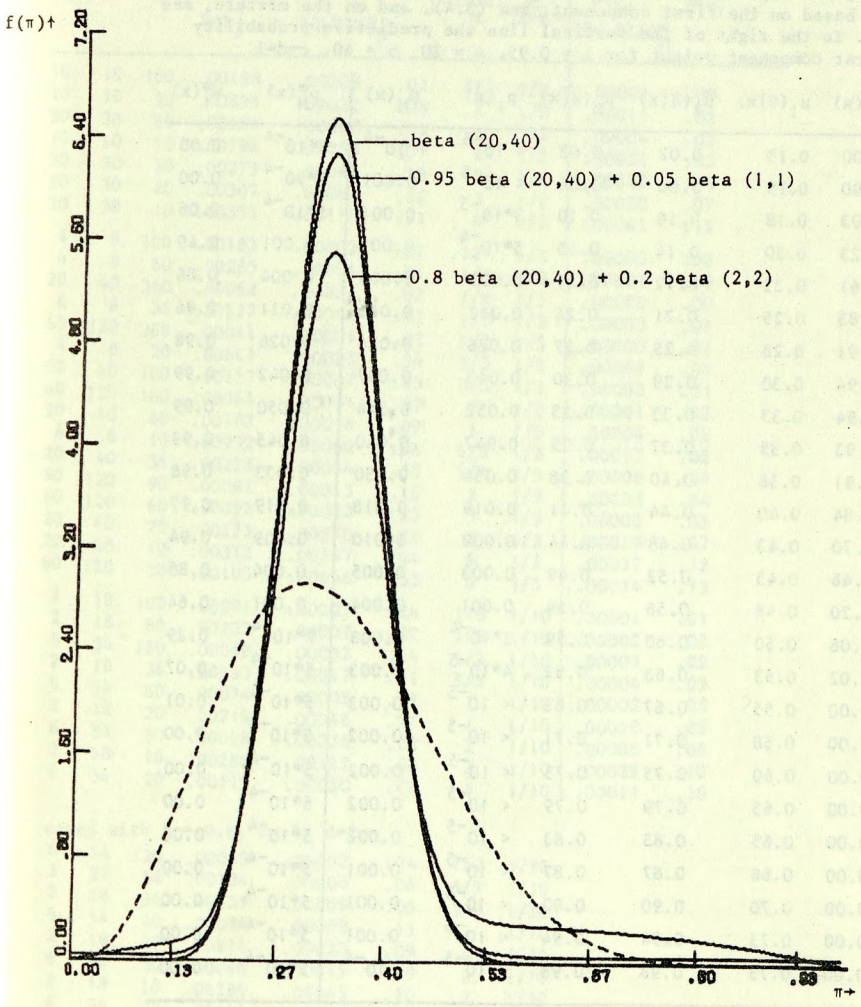

Figure 1. Densities of beta (20,40), two mixtures and beta (4,8) (dotted).

Table 1.
Posterior weights and means after obtaining x successes in n=100 trials, for selected values of x, when the prior for Π is the mixture (3.1) with $\lambda = 0.8$, a = 20, b = 40, c=d=2. Also given are the predictive probabilities of X=x based on the first component, see (3.4), and on the mixture, see (3.10). To the right of the vertical line the predictive probability and first component weight for $\lambda = 0.95$, a = 20, b = 40, c=d=1.

x	$w_1(x)$	$\mu_1(\Pi\|x)$	$\mu_2(\Pi\|x)$	$\mu_M(\Pi\|x)$	$p_1(x)$	$p_M(x)$	$p_M^*(x)$	$w_1^*(x)$
0	0.00	0.13	0.02	0.02	$< 10^{-5}$	10^{-4}	$5*10^{-4}$	0.00
4	0.00	0.15	0.06	0.06	$< 10^{-5}$	0.001	$5*10^{-4}$	0.00
8	0.03	0.18	0.10	0.10	$3*10^{-5}$	0.001	$5*10^{-4}$	0.06
12	0.23	0.20	0.14	0.15	$5*10^{-4}$	0.002	0.001	0.49
16	0.61	0.23	0.17	0.21	0.003	0.004	0.004	0.86
20	0.83	0.25	0.21	0.24	0.012	0.011	0.011	0.96
24	0.91	0.28	0.25	0.27	0.026	0.024	0.026	0.98
28	0.94	0.30	0.29	0.30	0.043	0.037	0.042	0.99
32	0.94	0.33	0.33	0.33	0.052	0.044	0.050	0.99
36	0.93	0.35	0.37	0.35	0.047	0.040	0.045	0.99
40	0.91	0.38	0.40	0.38	0.034	0.030	0.033	0.98
44	0.84	0.40	0.44	0.41	0.019	0.018	0.019	0.97
48	0.70	0.43	0.48	0.44	0.009	0.010	0.009	0.94
52	0.46	0.45	0.52	0.49	0.003	0.005	0.004	0.86
56	0.20	0.48	0.56	0.54	0.001	0.004	0.001	0.64
60	0.06	0.50	0.60	0.59	$2*10^{-4}$	0.003	$7*10^{-4}$	0.29
64	0.02	0.53	0.63	0.62	$4*10^{-5}$	0.003	$5*10^{-4}$	0.07
68	0.00	0.55	0.67	0.67	$< 10^{-5}$	0.003	$5*10^{-4}$	0.01
72	0.00	0.58	0.71	0.71	$< 10^{-5}$	0.002	$5*10^{-4}$	0.00
76	0.00	0.60	0.75	0.75	$< 10^{-5}$	0.002	$5*10^{-4}$	0.00
80	0.00	0.63	0.79	0.79	$< 10^{-5}$	0.002	$5*10^{-4}$	0.00
84	0.00	0.65	0.83	0.83	$< 10^{-5}$	0.002	$5*10^{-4}$	0.00
88	0.00	0.68	0.87	0.87	$< 10^{-5}$	0.001	$5*10^{-4}$	0.00
92	0.00	0.70	0.90	0.90	$< 10^{-5}$	0.001	$5*10^{-4}$	0.00
96	0.00	0.73	0.94	0.94	$< 10^{-5}$	0.001	$5*10^{-4}$	0.00
100	0.00	0.75	0.98	0.98	$< 10^{-5}$	10^{-4}	$5*10^{-4}$	0.00

Table 2.

Calculation of REL_I defined by (4.12). Unless stated otherwise, c=d=2 and $\lambda = 0.8$, but to the right of the vertical line c=d=1 and $\lambda = 0.95$

a	b	n	(4.4) AVE_I / VAR_I	(4.8) AVE_I / $SQ.BIAS_M$	(4.10) REL_I	$\frac{a+b}{n}$	$\frac{a}{a+b}$	(4.8) AVE_I / $SQ.BIAS^*_M$	(4.10) REL^*_I
10	10	100	.00198	.00003	.02	1/5	1/2	.00001	.00
10	10	20	.00595	.00030	.05	1	1/2	.00011	.02
30	30	60	.00205	.00018	.09	1	1/2	.00004	.02
10	10	10	.00794	.00053	.07	2	1/2	.00022	.03
30	30	30	.00273	.00044	.16	2	1/2	.00012	.04
30	30	20	.00307	.00068	.22	3	1/2	.00020	.07
30	30	10	.00351	.00110	.31	6	1/2	.00041	.12
4	8	100	.00183	.00002	.01	.12	1/3	.00000	.00
4	8	60	.00285	.00004	.01	1/5	1/3	.00001	.00
20	40	360	.00052	.00001	.02	1/6	1/3	.00000	.00
4	8	36	.00427	.00010	.02	1/3	1/3	.00003	.01
60	120	360	.00041	.00001	.02	1/2	1/3	.00000	.01
4	8	20	.00641	.00025	.04	3/5	1/3	.00007	.01
20	40	100	.00137	.00007	.05	3/5	1/3	.00002	.01
60	120	180	.00061	.00004	.07	1	1/3	.00001	.02
20	40	60	.00182	.00016	.09	1	1/3	.00004	.02
4	8	10	.00932	.00060	.06	6/5	1/3	.00017	.02
20	40	36	.00228	.00034	.15	5/3	1/3	.00008	.04
60	120	90	.00082	.00013	.16	2	1/3	.00003	.04
60	120	60	.00092	.00023	.25	3	1/3	.00005	.05
20	40	20	.00273	.00070	.26	3	1/3	.00018	.07
20	40	10	.00312	.00137	.44	6	1/3	.00037	.12
60	120	30	.00105	.00056	.53	6	1/3	.00014	.13
2	18	100	.00071	.00003	.04	1/5	1/10	.00001	.01
2	18	60	.00107	.00007	.07	1/3	1/10	.00002	.02
6	54	120	.00049	.00002	.04	1/2	1/10	.00001	.02
2	18	36	.00153	.00017	.11	5/9	1/10	.00004	.03
6	54	60	.00074	.00008	.11	1	1/10	.00002	.03
2	18	20	.00214	.00046	.21	1	1/10	.00010	.05
6	54	30	.00098	.00026	.26	2	1/10	.00006	.06
2	18	10	.00286	.00127	.44	2	1/10	.00029	.10
6	54	20	.00111	.00050	.46	3	1/10	.00011	.10
cases with $\lambda = 0.8$, c=.5, d=2									
6	54	120	.00049	.00002	.04	1/2	1/10		
3	27	54	.00104	.00006	.06	5/9	1/10		
2	18	36	.00153	.00009	.06	5/9	1/10		
6	54	60	.00074	.00008	.11	1	1/10		
2	18	20	.00214	.00020	.09	1	1/10		
6	54	30	.00098	.00019	.19	2	1/10		
2	18	10	.00286	.00045	.16	2	1/10		
6	54	20	.00111	.00030	.27	3	1/10		

Table 3.

Comparison of relative extra losses (sections 4 and 5) and embarassment (section 6). Quantities without a star denote the choice c=d=2, $\lambda = 0.8$, and quantities with a star c=d=1, $\lambda = 0.95$; for (6.3) see text.

a	b	n	(4.10) REL_1	(5.4) REL_M	(6.1) AVE_1 EMB_M	(6.2) AVE_M EMB_1	(6.3)	(4.10) REL_1^*	(5.4) REL_M^*	(6.1) AVE_1 EMB_M^*	(6.2) AVE_M^* EMB_1
10	10	100	.02	.06	.039	.108	.022	.01	.04	.015	.071
10	10	20	.05	.10	.020	.031	.004	.02	.06	.008	.015
30	30	60	.09	.65	.070	.309	.042	.02	.36	.022	.159
10	10	10	.07	.08	.010	.012	0	.03	.04	.003	.004
30	30	30	.16	.61	.051	.148	.022	.04	.38	.017	.071
30	30	20	.22	.51	.038	.084	.014	.07	.34	.013	.037
30	30	10	.31	.29	.018	.025	0	.12	.18	.006	.008
4	8	100	.01	.04	.043	.117	.021	.00	.02	.012	.052
4	8	60	.01	.06	.041	.102	.015	.00	.03	.011	.042
20	40	360	.02	.52	.113	1.335	.101	.00	.19	.029	.578
4	8	36	.02	.09	.037	.083	.011	.01	.04	.010	.031
60	120	360	.02	3.49	.144	3.543	.130	.00	1.40	.036	1.433
4	8	20	.04	.11	.031	.058	.007	.01	.05	.008	.019
20	40	100	.05	1.13	.099	.833	.077	.01	.48	.026	.334
60	120	180	.07	4.55	.134	2.473	.115	.02	2.00	.034	.973
20	40	60	.09	1.31	.090	.604	.060	.02	.59	.024	.233
4	8	10	.06	.12	.022	.033	0	.02	.05	.005	.008
20	40	36	.15	1.33	.077	.397	.045	.04	.65	.020	.146
60	120	90	.16	4.69	.119	1.496	.048	.04	2.39	.030	.573
60	120	60	.25	4.27	.108	1.037	.092	.05	2.40	.028	.390
20	40	20	.26	1.14	.060	.214	.021	.07	.61	.015	.072
20	40	10	.44	.77	.038	.085	.008	.12	.43	.009	.023
60	120	30	.53	2.99	.084	.488	.052	.13	1.99	.022	.174
2	18	100	.04	1.01	.152	1.172	.117	.01	.38	.029	.336
2	18	60	.07	1.41	.148	1.028	.108	.02	.55	.028	.285
6	54	120	.04	5.86	.178	3.348	.155	.02	2.36	.035	.961
2	18	36	.11	1.78	.142	.853	.088	.03	.75	.027	.227
6	54	60	.11	7.27	.171	2.469	.140	.03	3.25	.034	.687
2	18	20	.21	2.03	.130	.632	.066	.05	.93	.024	.158
6	54	30	.26	7.00	.159	1.604	.123	.06	3.68	.031	.430
2	18	10	.44	1.87	.110	.387	.041	.10	.96	.020	.087
6	54	20	.46	6.08	.148	1.168	.107	.10	3.55	.029	.305