Illustrating Frequentist and Bayesian Statistics in Oceanography[1]

George Casella
Cornell University

ABSTRACT

Both frequentist and Bayesian methodologies provides means for a statistical solution to a problem. However, it is usually the case that, for a given situation, one methodology is more appropriate. Using a number of oceanographic examples we explore the components of a statistical solution and illustrate the most appropriate methodology. We argue that the statistical consideration of utmost importance is the type of inference and conclusion to be made. In some examples it is more appropriate to make this inference as a Bayesian, and in some it is more appropriate to make this inference as a frequentist.

"Still, it is an error to argue in front of your data. You find yourself insensibly twisting them round to fit your theories."

Sherlock Holmes
The Adventure of Wisteria Lodge

1. INTRODUCTION

An alternate title for this paper might well be "Conditional and Unconditional Inference in Oceanographic Studies," as a fundamental difference between frequentist and Bayesian statistics is their resulting inference. A frequentist inference is unconditional, applying to a series of repeated experiments (most always an imagined series). In contrast, a Bayesian inference is conditional, applying to the data at hand, and not directly addressing the concept of repeatability.

This paper is an introduction to these methods, and illustrates their

[1]This paper was presented at the 'Aha Huliko'a Winter Workshop on "Probability Concepts in Physical Oceanography," January 12-15, 1993, Honolulu, Hawaii, and is technical report BU-1187-M, in the Biometrics Unit, Cornell University. This research was supported by National Science Foundation Grant No. DMS9100839 and National Security Agency Grant No. 90F-073.

uses with some oceanographic data sets. The primary message is that each statistical view has a lot to offer, and, depending on the problem, one methodology is probably more appropriate. We illustrate this through the examples.

A second goal of this paper is to try to explain to the oceanographic community how a statistician approaches a problem. The purpose of this endeavor is to provide a structured approach to dealing with problems involving data, from their inception to ending. In doing so, perhaps the task of dealing with the ever-increasing data bases can be made a little easier.

The remainder of the paper is arranged as follows. In Section 2 we give general outline of how to approach a problem statistically, illustrating this with an example in Section 3. Section 4 discusses the underlying differences between the frequentist and Bayesian approaches to statistics, and Sections 5 and 6 contain more examples illustrating these methodologies. Section 7 contains a concluding discussion.

2. COMPONENTS OF A STATISTICAL SOLUTION

In the best of all possible worlds, a problem is planned, statistically, from beginning to end. Chronologically, the steps of a solution can be listed as in Table 1.

Table 1: Components of a Statistical Solution
(Chronological Order)

1. **Model** the Process
2. **Design** the Experiment
3. **Collect** the Data
4. **Estimate** and Verify the Model
5. **Infer** and Conclude
6. **Implement** the Solution

Although the steps are performed in chronological order, they are best planned in reverse order. That is, when approaching any problem, the first consideration is "How will the knowledge we gain be implemented?"

For example, if a study is proposed to examine wave magnitude and direction in the North Atlantic, the first consideration should be the use of the resulting knowledge. Will it be used to plan routes for oil tankers? Will it be used to increase our basic knowledge of ocean dynamics? By answering this question first, the remainder of the steps of a statistical solution will fall into place, and the problem can be attacked in a very efficient fashion. Although this mechanism for solution is not usually taught in the classroom, it seems to be the one most preferred by statisticians. By concentrating on the final result, the entire study becomes focused.

With respect to frequentism or Bayesianism, the components of the statistical solution remain essentially unchanged. Of course, there are some differences in the approaches, with the major difference being in the modeling and inference stages. However, the overall attack is similar. This is illustrated in the next section.

3. AN EXAMPLE CONCERNING ICEBERGS

Defant (1961, page 278) presented the following data on the frequency of icebergs off Newfoundland.

Table 2: Frequency of Icebergs off Newfoundland south of 48°N (a) and south of the Grand Banks (b), for the period 1900-1926.

	Month												
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	
(a)	3	10	36	83	130	68	25	13	9	4	3	2	386
(b)	0	1	4	9	18	13	3	2	1	0	0	0	51

For our example, we will look at the question of whether the yearly distribution of icebergs is the same in each location. A glance at Figure 1 will show that such a hypothesis is very likely, but for illustration we will step through both a Bayesian and frequentist approach to the problem. We take as the goal of our study to be the description of the distribution of icebergs off Newfoundland.

In both the Bayesian and frequentist approaches to this problem we assume that the data are distributed according to a multinomial distribution, and we wish to test the null hypothesis H_0: The distributions in locations (a) and (b) are the same. To test this as a frequentist we use a chi-squared test of association (see Snedecor and Cochran, 1989). The chi-squared test results in a p-value of .977, which is very strong evidence in favor of the null hypothesis.

To perform a Bayesian analysis a prior distribution must be specified, that is, a distribution that we subjectively believe describes the pattern of icebergs. We then use this distribution, in conjunction with the observed data, to assess the plausibility of the hypothesis. Since we really have no prior knowledge about the icebergs, we use a strategy that attempts to model this ignorance, and calculate the probability of every data table with the given marginal totals, using a hypergeometric distribution. This leads us to use Fisher's exact test (Fisher,1970) and assess the probability of the null hypothesis as .994. Again, this is very strong evidence in favor of this hypothesis. (Strictly speaking, Fisher's exact test is not a Bayesian procedure but a conditional procedure, as it is calculated conditionally on the observed data. However, the important feature is that it yields a conditional inference.)

We now can clearly see the distinction between Bayesian and frequentist inferences. The frequentist bases inference on a frequency interpretation. A formal conclusion would be of the form, "the statistical procedure used (here the chi-squared test) would result in an erroneous inference less than 5% of the time in repeated experiments." In contrast, the Bayesian inference is conditional on the observed data, and would formally conclude "based on the stated prior distribution and observed data, the probability is .994 that H_0 is true." We now look at these differences a bit more closely.

4. WHERE DOES THE RANDOMNESS COME FROM?

The most important part of any statistical investigation is the resulting inference. In fact, it may even be said that the main reason

for doing a statistical investigation is to produce a meaningful inference, since the inference applies to a wider population than is actually studied and measured. (For example, after measuring the activities of a number of waves in a certain area, we are then interested in making a statement (an inference) about all waves in that area.) To make this inference we need an underlying model of the phenomena, one that accounts for the randomness of the observations and allows an inference. Bayesians and frequentists have different approaches to this.

4.1 Frequency Randomness

The frequentist assumes repeatability of the experiment, that the experiment actually performed is one of an infinitely long sequence of identical experiments. If we denote this sequence of experiments E_1, E_2, ..., E_k, E_{k+1}, ..., then we make our inference to the entire sequence, even though only one experiment (say E_k) is actually performed. The rest of this imagined sequence builds the randomness into our model. We know that the results of each experiment (if performed) would be slightly different, and our inference will take these potential differences into account.

Thus, the frequentist inference is an unconditional one that applies to the entire sequence, and does not single out the experiment actually performed. It is important to realize that the inference is about the performance of the *procedure* over the entire sequence of experiments, such as, "The statistical procedure used will be correct in 95% of all experiments performed." The actual outcome of the observed experiment will not change this inference.

4.2 Bayesian Randomness

In a Bayesian analysis the data are assumed to be fixed, and inference is made conditional on their observed values. Thus, no randomness comes from the data. The randomness in a Bayesian inference comes from the subjective prior distribution. This randomness, together with the information in the data, are combined into the posterior distribution. The posterior distribution is then used for inference. Of course, different subjective prior distributions may result in different

inferences.

More precisely, suppose there are data, X, which vary according to a probability distribution $f(x|\theta)$, a distribution indexed by an unknown parameter θ. (For example $f(\cdot|\theta)$ may be a Gaussian distribution with unknown mean θ.) We then assume that the parameter θ varies according to a prior distribution $\pi(\theta)$. This probability distribution reflects our knowledge about the parameter θ before observing the new data x. (In keeping with convention, an upper case X denotes an unseen random variable while a lower case x denotes an observed value. Thus the equation "$X = x$" means that we have observed the value x of the random variable X.) Using the laws of probability (or sometimes called Bayes rule) we calculate the posterior distribution of θ given $X = x$, $g(\theta|x)$, as

$$g(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta},$$

where the integral is over all values of θ. (For more detail on such calculations, see Casella and Berger, 1990.) Our inference is then based on $g(\theta|x)$, which only considers the experiment actually performed, not any repeated sequence. For example, one might infer "Based on the specified $\pi(\theta)$ and observed x, we conclude that $\theta \geq 0$ with probability 95%." This inference would follow if it were the case that $\int_0^\infty g(\theta|x)d\theta = .95$.

4.3 The Appropriate Inference

As mentioned before, the purpose of this paper is not to make value judgments as to which of Bayesianism or frequentism is better. Rather, the purpose is to illustrate situations where one method is more appropriate. It then follows that the more appropriate methodology, and inference, is the one to use. From the previous two subsections, we see that the frequentist inference is more appropriate if repeatability is important, while the Bayesian inference is more appropriate if the inference is to be made conditional on the observed data. Returning to the iceberg data, it seems that the Bayesian inference is more appropriate, as we are faced with a data set that is unrepeatable, and we

are interested in an inference conditional on that data set. (Interestingly, it was argued during discussions at the workshop that one could consider the observed 26-year period as one of a sequence of 26-year periods, in which case the frequentist inference maybe more appropriate.) If it may be argued that either interpretation is valid, and hence either inference is appropriate, there is no problem. As long as the methodology is chosen to appropriately answer the question of interest, phrased in the manner of interest, the statistics have served their purpose.

5. AN EXAMPLE CONCERNING BREAKING WAVES

Hwang, Hsu and Wu (1990) report on an experiment concerning average height of breaking waves, H_B, measured as a function of RMS surface displacement, η. The data are presented in Figure 2. They conclude that $H_B < H_S$, the significant wave height, where $H_S = 4\eta$, and state, "In a random wave field, waves that break due to local instabilities are not necessarily the highest waves." Statistically, we can think of this as testing the hypotheses

$$H_0: \ H_B \leq 4\eta \qquad \text{vs.} \qquad H_1: \ H_B > 4\eta \ .$$

It seems here that frequency considerations are important, in that conclusions should apply to repetitions of the experiment. This concern seems implicit in the above quoted conclusion of Hwang et al. Thus, a frequentist analysis is more appropriate. Using a standard linear regression model with Gaussian errors, we obtain a p-value of .999 for the hypothesis H_0: $H_B \leq 4\eta$, showing that there is overwhelming evidence to support this hypothesis. (In fact, the hypothesis H_0: $H_B \leq 3\eta$ yields a p-value of .911, demonstrating extremely good support for this even stronger claim.)

Of course, a Bayesian analysis could also be performed, but the inference would not apply to a sequence of experiments. The conclusions would be conditional on the observed data. To do the Bayesian analysis we again use a standard linear regression model with Gaussian errors, but

we also assume that $H_B = b\eta$, where b is a parameter with a specified prior distribution. We specify the prior to also be Gaussian, and we take the prior mean to be equal to the hypothesized value. (Thus, for testing H_0: $H_B \leq 4\eta$ we specify a Gaussian prior with mean 4. This strategy of centering the prior at the hypothesized value gives equal prior weight above and below the value, and may be considered an impartial prior specification.)

Combining our prior specification with the observed data, we calculate $\Pr(b \leq 4 \mid \text{data}) = .999$ and $\Pr(b \leq 3 \mid \text{data}) = .623$. That is, for the specified priors and conditional on the observed data, b is less than 4 with probability .999 and less than 3 with probability .623. Quantitatively, these conclusions are similar to those of the frequentist, and show overwhelming support for the null hypotheses. The only difference is in the scope of the inference.

Bayesian conclusions are, of course, dependent on the prior specification, and sometimes there might be concern about oversensitivity to this specification. Such a concern is easily addressed, however, by calculating posterior probabilities over a range of prior specifications. This is illustrated in Figure 3, where we display the posterior probabilities over a wide range of standard deviations. (The standard deviation of the data is .082, and the graph shows the prior standard deviation up to twice this value.) The figure shows that, for this range of prior standard deviations, the conclusions from the Bayesian analysis are relatively stable in their support of H_0.

6. AN EXAMPLE CONCERNING BUBBLE POPULATIONS

The distribution of bubble populations is also investigated by Hwang, et al. (1990). They collected data on bubble populations as a function of depth and wind velocity, as presented in Figure 4. For a given depth, Z, (cm) and wind velocity, u, (m/s), the logarithm of the bubble population, N(Z), (log cm^3) is modeled as

$$N(Z) = a_u + b_u Z + \epsilon \qquad u = 10,\ 11,\ \ldots,\ 15$$

where ϵ represents random error, and is assumed to have Gaussian distribution with mean 0 and variance σ^2.

A question of interest is whether the distribution of bubbles is the same at each depth. After some thought, it seems that the appropriate inference here is the frequentist inference. Concern about the repeatability of the inference leads to this conclusion, as we would like to be able to describe the bubble populations at a given depth and wind velocity when such conditions are again realized.

6.1 A Standard Frequentist Inference

A standard approach to this problem is to decide if the slopes are the same at each wind velocity, so we would test the null hypothesis H_0: $b_{10} = b_{11} = \ldots = b_{15}$. Doing so leads to a p-value of .063, which suggests rejection of H_0. Thus a standard frequentist analysis would lead us to fit separate regression lines for each wind velocity. So for each wind velocity we would use a separate regression equation to predict the bubble population. See Table 3 and Figure 5.

6.2 An Empirical Bayes Analysis

The bubble population data is ideal for an empirical Bayes analysis—a mixture of frequency and Bayesian analyses that combines the best features of each. Here we will only briefly explain the methodology, for a more detailed introduction see the articles by Casella (1985, 1992).

Table 3: Coefficients for the standard regression analysis (frequentist) and empirical Bayes analyses of the bubble populations.

Wind Velocity	n	intercept	slope	std. dev.	empirical Bayes slope
10	4	.666	−.084	.011	−.076
11	5	.924	−.040	.013	−.042
12	4	1.594	−.080	.008	−.073
13	5	1.669	−.050	.017	−.050
14	4	1.698	−.031	.029	−.035
15	4	1.635	−.0009	.027	−.011

To perform an empirical Bayes analysis we start with the frequentist model and inference structure. We append a Bayes model to the slopes

$$b_u \sim \text{Gaussian } (b, \tau^2) \ , \ u = 10, \ 11, \ \ldots, \ 15,$$

that is, that the slopes come from a common Gaussian population with unknown mean b and variance τ^2.

The "empirical" part of empirical Bayesian is to now estimate these unknown parameters b and τ^2 from the data. (A standard Bayesian analysis would specify values for these parameters.) Using these estimated values allows the data to assess the tenability of the *submodel*, that the b_u's come from a common population. The empirical Bayes slope estimates are a convex combination of the common overall slope ($-.048$) and the individual least squares slopes, given by

$$\begin{matrix}\text{empirical Bayes} \\ \text{slope}\end{matrix} = (.221)(-.048) + (.779)\left(\begin{matrix}\text{least squares} \\ \text{slope}\end{matrix}\right).$$

The weighting factor .221 (and $.779 = 1 - .221$) are data based estimates. The empirical Bayes slope estimates are valid under the model of frequentist repeatability. In fact, they are superior to the frequentist estimates using a criterion of expected mean squared error. Thus, on the average, the empirical Bayes estimates will be closer to the true values than the standard frequentist estimates. They combine the best features of Bayesian modeling and frequentist inference.

Figure 5 also shows the empirical Bayes regression lines. Although they are not very different from the standard frequentist lines, they do display a movement toward the common slope value. The empirical Bayes analysis has uncovered a small amount of common structure, and has used this in improving each of the estimates.

7. CONCLUSIONS

The statistical methodology to be used, whether Bayesian or frequentist, should be selected according to the type of inference that is desired (and is appropriate). The frequentist methodology is

appropriate for inference over a series of repeated experiments, while the Bayesian methodology is appropriate for inference specific to the experiment that was done. This article has given examples and provided discussion of situations where each methodology is appropriate.

There is no brick wall between Bayesianism and frequentism. The methodologies are not at odds with one another, they are complementary to one another. When approaching a statistical problem "opportunism" is best. With that in mind, the appropriate analysis and inference can be chosen from all available statistical methodologies.

Both Bayesianism and frequentism are built on a set of assumptions, some more palatable than others. For a user of frequentist methods, perhaps the assumption most difficult to believe is that the process (including parameter values) remains constant over the imagined series of experiments. For user of Bayesian methods, perhaps the assumption most difficult to believe is that the prior distribution is correct. These assumptions, however, can sometimes be checked and and maybe even relaxed. Moreover, their reasonableness in any particular situation may also form a basis for choosing an appropriate methodology. (See Berger 1985, who discusses robust Bayesian analysis, which addresses these concerns). Lastly, there is an enormous amount of research being done in statistics, and some of it is aimed at relaxing these assumptions. Such research has already given us techniques like empirical Bayes analysis, a synthesis of both Bayesian and frequentist methodologies which can often provide superior solutions.

REFERENCES

Berger, J.O. (1985): *Statistical Decision Theory and Bayesian Analysis*, Second Edition. New York: Springer-Verlag.

Casella, G. (1985): An Introduction to Empirical Bayes Data Analysis. *The American Statistician*, 39, 83-87.

Casella, G. (1992): Illustrating Empirical Bayes Methods. *Chem. and Intell. Lab. Sys.*, 16, 107-125.

Casella, G. and Berger, R.L. (1990): *Statistical Inference*, Pacific Grove: Wadsworth and Brooks/Cole.

Defant, A. (1961): *Physical Oceanography*, Volume I. New York: Pergamon Press.

Fisher, R.A. (1970): *Statistical Methods for Research Workers*, Fourteenth Edition. New York: Hafner (Reissued by Oxford University Press, 1990).

Hwang, P.A., Hsu, Y.-H.L., and Wu, Jin. (1990): Air Bubbles Produced by Breaking Wind Waves: A Laboratory Study, *J. Phys. Oceanography, 20*, 19-28.

Snedecor, G.W. and Cochran, W.G. (1989): *Statistical Methods*, Eighth Edition. Ames: Iowa State University Press.

Figure 1: Relative frequencies of icebergs off (a) Newfoundland (black squares) and (b) Grand Banks (white squares).

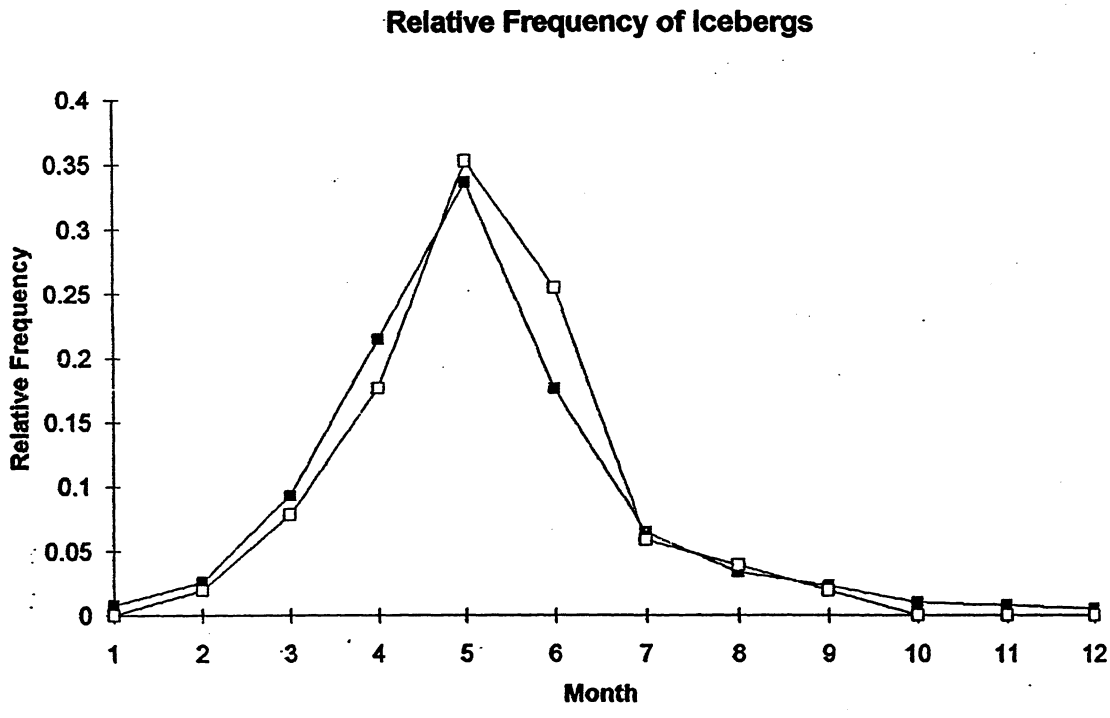

Figure 2: Averaged height of breaking waves, H_B, as a function of RMS water surface displacement, η. The line shown is the least squares line, with equation $H_B = .102 + 2.89\eta$ ($r^2 = .994$).

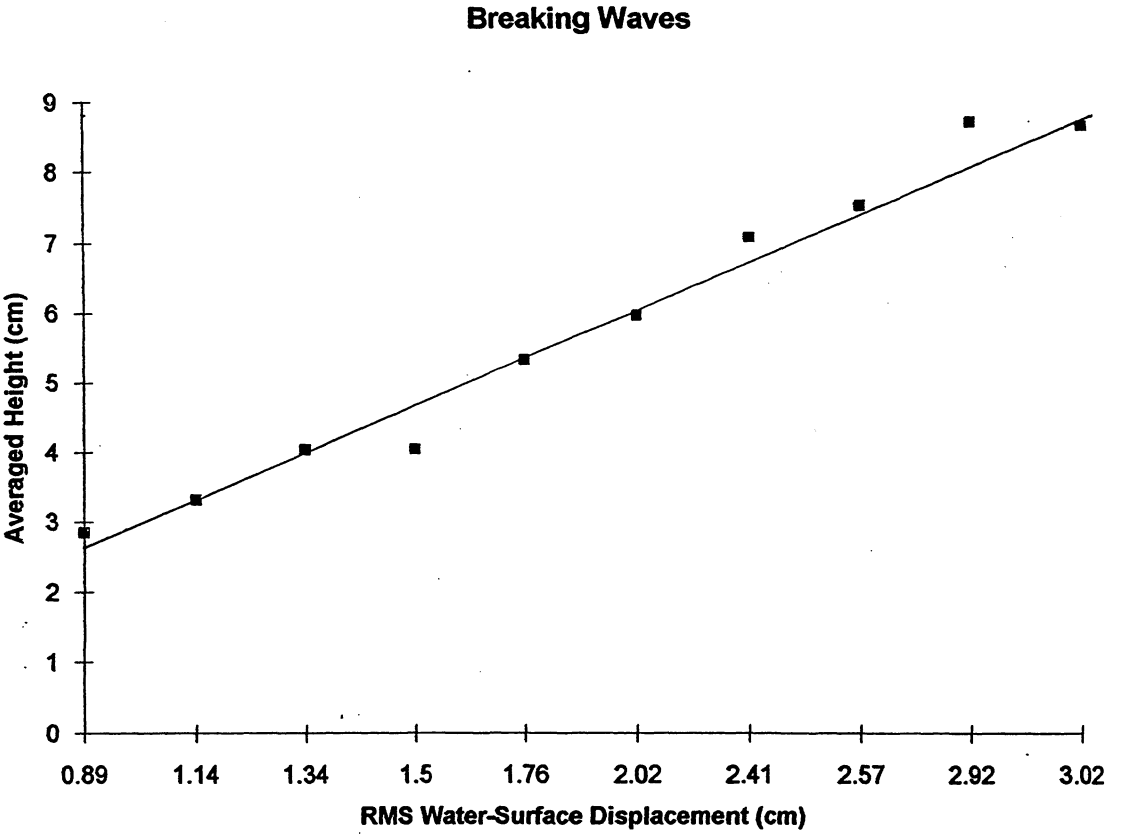

Figure 3: Posterior probabilities for the null hypotheses $H_0: b \leq 4$ (solid lines) and $H_0: b \leq 3$ (dashed lines), as a function of the prior standard deviation.

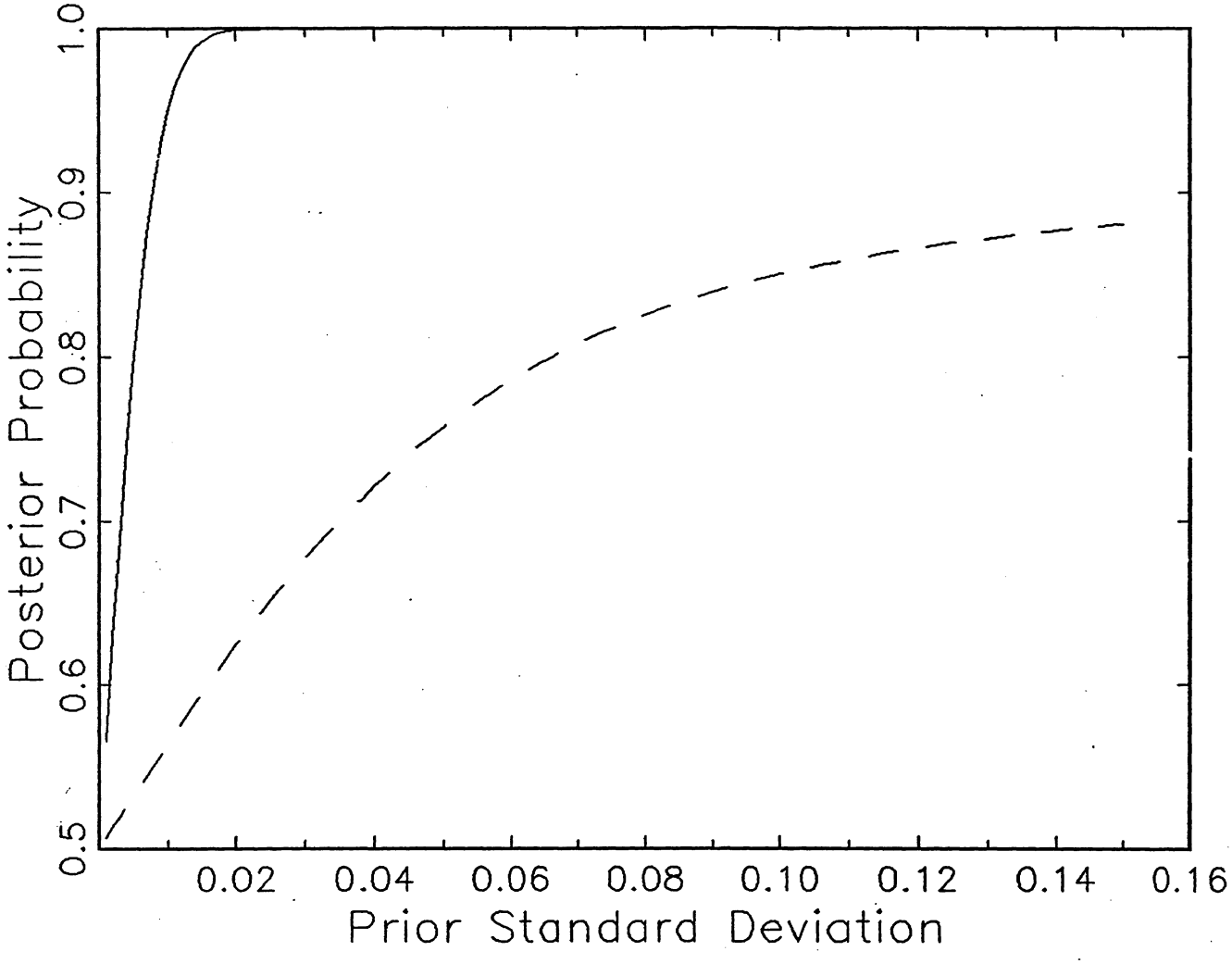

Figure 4: Data from Hwang et al. (1990) on bubble populations. The six groups are each at a different wind velocity, from 10 to 15 m/s in steps of 1. The groups are in order from 10 m/s (lowest) to 15 m/s (highest), and are denoted by black squares (10 m/s), white squares (11 m/s), black diamonds (12 m/s), white diamonds (13 m/s), black triangles (14 m/s) and white triangles (15 m/s). The data are connected merely to aid viewing.

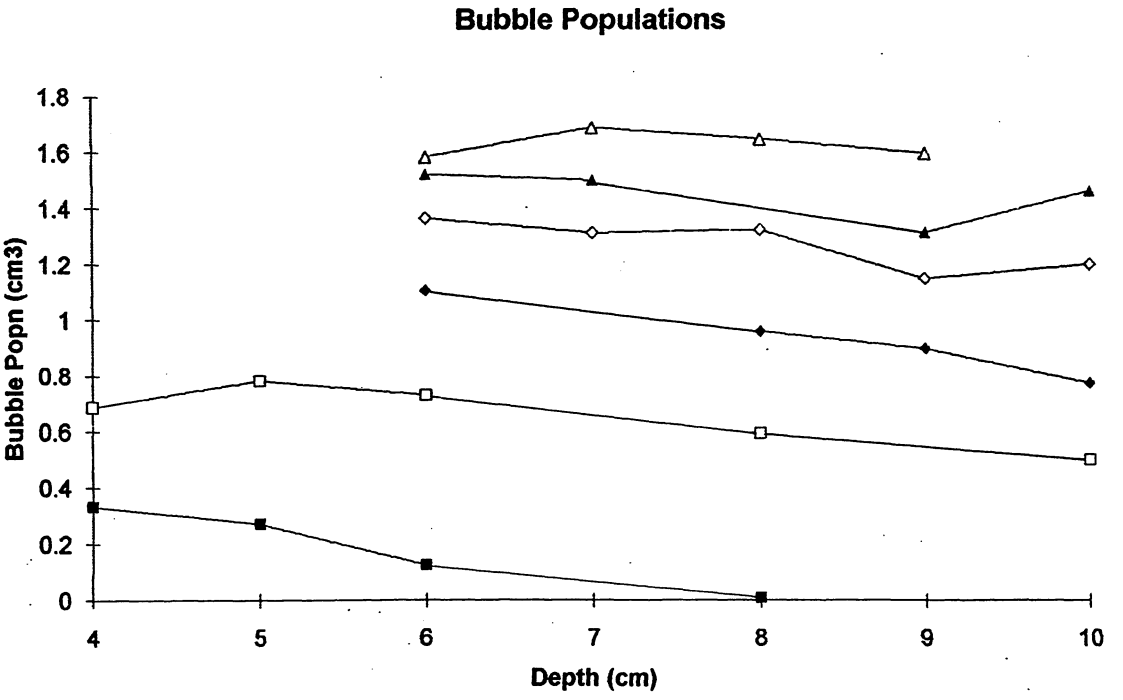

Figure 5: Standard frequentist (solid lines) and empirical Bayes (dashed lines) fits to the bubble data, coded as in Figure 4. The empirical Bayes lines (whose slopes are pulled toward −.048) are under the least squares lines for 11, 14 and 15 m/s, and above the least squares lines for 10 and 12 m/s. The lines are virtually identical for 13 m/s.