

I^2SDS
The Institute for Integrating Statistics in Decision Sciences

Technical Report TR-2007-14
December 1, 2007

Choosing a Coverage Probability for Prediction Intervals

Joshua Landon
Institute for Integrating Statistics in Decision Sciences
The George Washington University, USA

Nozer D. Singpurwalla
Department of Statistics
The George Washington University, USA

Choosing a Coverage Probability for Prediction Intervals

Joshua Landon and Nozer D. Singpurwalla

The George Washington University, Washington, D.C. 20052

December, 2007

Abstract

Coverage probabilities for prediction intervals are germane to filtering, forecasting, previsions, regression, and time series analysis. It is a common practice to choose the coverage probabilities for such intervals by convention, or astute judgment. We argue here that coverage probabilities can be chosen by decision theoretic considerations. But to do so, we need to specify meaningful utility functions. Some stylized choices of such functions are given, and a prototype approach is presented.

Key Words: Confidence Intervals, Decision Making, Filtering, Forecasting, Previsions, Time Series, Utilities.

1 Introduction and Background

Prediction is perhaps one of the most commonly undertaken activities in the physical, the engineering, and the biological sciences. In the econometric and the social sciences, prediction generally goes under the name of *forecasting*, and in the actuarial and the assurance sciences under the label *life-length assessment*. Automatic process control, filtering, and quality control, are some of the engineering techniques that use prediction as a basis of their modus operandus.

Statistical techniques play a key role in prediction, with regression, time series analysis, and dynamic linear models (also known as state space models) being the predominant tools for producing forecasts. The importance of statistical methods in forecasting is underscored by Pearson (1920) who has claimed that prediction is the “fundamental problem of practical statistics”. Similarly, with de Finetti (1972, Chapters 3 and 4), who labels prediction as “prevision”, and makes it the centerpiece of his notion of “exchangeability” and a subjectivistic Bayesian development around it. In what follows, we find it convenient to think in terms of regression, time series analysis, and forecasting techniques as vehicles for discussing an important aspect of prediction.

We start by noting that inherent to the above techniques is an underlying distribution (or error) theory, whose net effect is to produce predictions with an uncertainty bound; the normal (Gaussian) distribution is typical. An exception is Gardner (1988) who uses a Chebychev inequality in lieu of a specific distribution. The result of the above is a *prediction interval* whose width depends on a coverage probability; see, for example, Box and Jenkins (1976), p.254, or Chatfield (1993). It has been a common practice to specify coverage probabilities by convention, the 90%, the 95%, and the 99% being typical choices. Indeed Granger (1996) states that academic writers concentrate almost exclusively on 95% intervals, whereas practical forecasters seem to prefer 50% intervals. The larger the coverage probability, the wider the prediction interval, and vice versa. But wide prediction intervals tend to be of little value [cf. Granger (1996), who claims 95% prediction intervals to be “embarrassingly wide”]. By contrast, narrow prediction intervals

tend to be risky in the sense that the actual values, when they become available, could fall outside the prediction interval. Thus the question of what coverage probability one should choose in any particular application is crucial.

1.1 Objective

The purpose of this article is to make the case that the choice of a coverage probability for a prediction interval should be based on decision theoretic considerations. This would boil down to a trade-off between the utility of a narrow interval versus the disutility of an interval that fails to cover an observed value. Hopefully, our approach endows some formality to a commonly occurring problem that seems to have been traditionally addressed by convention and judgment, possibly because utilities are sometimes hard to pin down.

1.2 Related Issues

Before proceeding further, it is important to note that in the context of this article, a prediction interval is not to be viewed as a *confidence interval*. The former is an estimate of a future observable value; the latter an estimate of some fixed but unknown (and often unobservable) parameter. Prediction intervals are produced via frequentist or Bayesian methods, whereas confidence intervals can only be constructed via a frequentist argument. The discussion of this paper revolves around prediction intervals produced by a Bayesian approach; thus we are concerned here with *Bayesian prediction intervals*. For an application of frequentist prediction intervals, the paper by Lawless and Fredette (2005) is noteworthy; also the book by Hahn and Meeker (1991, Section 2.3), or the paper of Beran (1990).

A decision theoretic approach for specifying the confidence coefficient of a confidence interval is not explored by us. All the same, it appears that some efforts in this direction were embarked upon by Lindley and Savage [cf. Savage (1962), p.173, who also alludes to some work by Lehmann (1958)]. By contrast, a decision theoretic approach for generating prediction intervals has been alluded to by Tay and Wallis (2000) and developed by

Winkler (1972). However, Winkler's aim was not the determination of optimal coverage probabilities, even though the two issues of coverage probability and interval size are isomorphic. Our focus on coverage probability is dictated by its common use in regression, time series analysis, and forecasting.

Finally, predictions and prediction intervals should not be seen as being specific to regression and time series based models. In general they will arise in the context of any probability models used to make previsions, such as the ones used in reliability and survival analysis [cf. Singpurwalla (2006), Chapter 5].

2 Motivating Example

Our interest in this problem was motivated by the following scenario. For purposes of exposition, we shall anchor on this scenario.

A university wishes to predict the number of freshman students that will be retained to their sophomore year. Suppose that N is the number of freshman students, and X is the number retained to the sophomore year; $X \leq N$. Knowing N , the university wishes to predict X . The prediction is to be accompanied by a prediction interval, and the focus of this article pertains to the width of the interval. The width of the interval determines the amount of funds the university needs to set aside for meeting the needs of the sophomore students. The wider the interval, the greater the reserves; however, large reserves strain the budget. By contrast, the narrower the interval the greater is the risk of the actual number of sophomores falling outside the interval. This would result in poor budgetary planning due to insufficient or excessive reserves. Thus a trade-off between the risks of over-budgeting and under-budgeting is called for.

The student retention scenario is archetypal because it arises in several other contexts under different guises. A direct parallel arises in the case of national defense involving an all volunteer fighting force. Meaningful predictions of the retention of trained personnel are a matter of national security. A more classical scenario is the problem of inventory control wherein a large volume of stored items ties up capital, whereas too little inventory may

result in poor customer satisfaction or emergency actions; see, for example, Hadley and Whitin (1960, Chapter 4). Another (more contemporary) scenario comes from the Basel II accords of the banking industry. Bank regulators need to assess how much capital a bank needs to set aside to guard against financial risks that a bank may face; see Decamps, Rochet and Roger (2004) for an appreciation. From the biomedical and the engineering sciences arises the problem of predicting survival times subsequent to a major medical intervention or a repair.

In all the above scenarios, the width of the prediction interval is determined by the nature of an underlying probability model and its coverage probability. This point is best illustrated by a specific assumption about the distribution of the unknown X ; this is done next. But before doing so, it is necessary to remark that neither the literature on inventory control, nor that on Basel II accords, addresses the issue of optimal coverage probabilities. In the former case, a possible reason could be the difficulties associated with quantifying customer dissatisfaction.

2.1 Distributional Assumptions

Suppose that the (posterior) predictive distribution of X obtained via a regression or a time series model is a normal (Gaussian) with a mean μ and variance σ^2 , where μ and σ^2 have been pinned down; the normal distribution is typical in these contexts. Then, it is well known [cf. De Groot (1970), p.228] that under a squared error loss for prediction error, μ is the best predictor of X . For a coverage probability of $(1 - \alpha)$, a prediction interval for X may be of the form $\mu \pm z_{\alpha/2}\sigma$. Here $z_{\alpha/2}$ is such that for some random variable W having a standard normal distribution, $P(W \geq z_{\alpha/2}) = \alpha/2$.

The question that we wish to address in this article is, what should α be? A small α will widen the prediction interval diminishing its value to a user. Indeed, $\alpha = 0$ will yield the embarrassing $(-\infty, +\infty)$ as a prediction interval. By contrast, with large values of α , one runs the risk of the prediction interval not covering the actual value (when it materializes). Thus, we need to determine an optimum value of α to use. To address the

question posed, we need to introduce utilities, one for the worth of a prediction interval, and the other, a disutility, for the failure of coverage.

3 Candidate Utility Functions

Utilities are a key ingredient of decision making, and the principle of maximization of expected utility prescribes the decision (action) to be taken; see, for example, Lindley (1985), p.71. Utilities measure the worth of a consequence to a decision maker, and disutilities the penalty (or loss) imposed by a consequence. With disutilities, a decision maker's actions are prescribed by the principle of minimization of expected disutilities. The unit of measurement of utilities is a "utile". However, in practice utilities are measured in terms of monetary units, such as dollars, and this is what we shall assume.

In the context of prediction, we make the natural assumption that, in principle, one prefers a prediction interval of width zero over any other prediction interval. This makes the utility of any prediction interval of non-zero width a disutility. Similarly, the failure of any prediction interval to cover an observed value results in a disutility. Following Winkler (1972), the two disutilities mentioned above are assumed to be additive, though this need not be so. Thus for the scenario considered here, one endeavors to choose that value of α for which the total expected disutility is a minimum.

3.1 The Disutility of a Prediction Interval

The width d_α of a prediction interval of the type described in Section 2.1 is $d_\alpha = 2z_{\alpha/2}\sigma$; here the coverage probability is $(1 - \alpha)$. Let $c(d_\alpha)$ be the disutility (or some kind of a dollar penalty) associated with a use of d_α . Clearly $c(d_\alpha)$ should be zero when $d_\alpha = 0$, and $c(d_\alpha)$ must increase with d_α , since there is a disadvantage to using wide intervals. A possible choice for $c(d_\alpha)$ could be

$$c(d_\alpha) = d_\alpha^\beta, \tag{3.1}$$

for $\beta > 0$. When $\beta < 1$, $c(d_\alpha)$ is a concave increasing function of d_α , and when $\beta > 1$, $c(d_\alpha)$ is convex and increasing in d_α . The choice of what β must be depends on the application.

In certain applications, such as target tracking, $\beta < 1$ may be more desirable than $\beta > 1$; in others, such as econometric forecasting, a convex disutility function may be suitable. The choice of (3.1) for a disutility function is purely illustrative. The proposed approach is not restricted to any particular choice for $c(d_\alpha)$.

3.2 The Disutility of Non-Coverage

A possible function for the disutility caused by a failure of the prediction interval to cover x , a realization of X , can be prescribed via the following line of reasoning.

Suppose that $U_\alpha = \mu + z_{\alpha/2}\sigma$ is the upper bound, and $L_\alpha = \mu - z_{\alpha/2}\sigma$, the lower bound of the $(1 - \alpha)$ probability of coverage prediction interval. Let $L(d_\alpha, x)$ denote the disutility or penalty loss (in dollars) in using a prediction interval of width d_α when X reveals itself as x . Then $L(d_\alpha, x)$ could be of the form

$$L(d_\alpha, x) = \begin{cases} f_1(x - U_\alpha), & x > U_\alpha, \\ 0, & L_\alpha < x < U_\alpha, \\ f_2(L_\alpha - x), & x < L_\alpha, \end{cases} \quad (3.2)$$

where f_1 and f_2 are increasing functions of their arguments, which encapsulate the penalty of x overshooting and undershooting the prediction interval, respectively.

As illustrated in Figure 1, the said functions will generally be convex and increasing because a narrow miss by the interval will matter less than a large miss. Furthermore, these functions need not be symmetric. For example, as shown in Figure 1, the penalty for undershooting the interval is assumed to be more severe than that of overshooting.

3.3 The Expected Total Disutility

With $c(d_\alpha)$ and $L(d_\alpha, x)$ thus specified, there remains one caveat that needs to be addressed. When the α is chosen, the value of x is not known and thus $L(d_\alpha, x)$ needs to be averaged over the possible values that x can take. This is easy to do because the predictive distribution of X has to be specified. Accordingly, let

$$R(d_\alpha) = E_X [L(d_\alpha, x)], \quad (3.3)$$

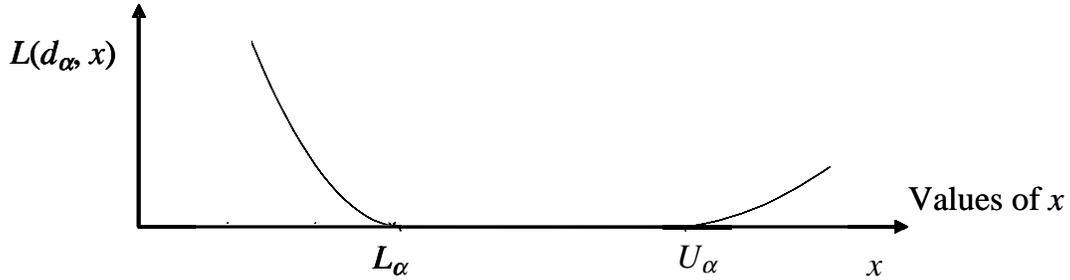


Figure 1: The Disutility of Non-Coverage

be the expected value of $L(d_\alpha, x)$. In decision theory, $R(d_\alpha)$ is known as the **risk function**; it is free of X . $R(d_\alpha)$ encapsulates the risk of non-coverage by an interval of width d_α , with $R(d_\alpha)$ decreasing in d_α .

Since $c(d_\alpha)$ is devoid of unknown quantities – indeed d_α is a decision variable – the matter of taking an expectation of $c(d_\alpha)$ is moot. We may now combine $c(d_\alpha)$ and $R(d_\alpha)$ to obtain the **total expected disutility function** as

$$D(d_\alpha) = c(d_\alpha) + R(d_\alpha). \quad (3.4)$$

As mentioned before, the additive choice, albeit natural, is not binding. We choose that value of α for which $D(d_\alpha)$ is a minimum. This is described next.

4 Choosing an Optimum Coverage Probability

To make matters concrete, suppose that $c(d_\alpha) = \sqrt{d_\alpha}$, so that the β of Equation (3.1) is $1/2$. Also, since $d_\alpha = 2z_{\alpha/2}\sigma$, $U_\alpha = \mu + z_{\alpha/2}\sigma$ can be written as $U_\alpha = \mu + d_\alpha/2$; similarly, $L_\alpha = \mu - d_\alpha/2$.

For the f_1 and f_2 of Equation (3.2), we let $f_1(x - U_\alpha) = (x - U_\alpha)^2/40$ and $f_2(L_\alpha - x) = (L_\alpha - x)^2/10$. These choices encapsulate a squared-error disutility, and make f_1 and f_2 asymmetric with respect to each other. Writing U_α and L_α in terms of d_α , we have $f_1(x - U_\alpha) = (x - \mu - d_\alpha/2)^2/40$, and $f_2(L_\alpha - x) = (\mu - d_\alpha/2 - x)^2/10$.

To compute the risk function of Equation (3.3) we need to specify μ and σ^2 of the normal distribution of X . Based on a Bayesian time series analysis of some student retention data, these were determined to be $\mu = 2140$ and $\sigma^2 = 396$. With the above in place, we may compute the total expected disutility as

$$D(d_\alpha) = \sqrt{d_\alpha} + R(d_\alpha),$$

where

$$R(d_\alpha) = \int_{\mu+d_\alpha/2}^{\infty} \frac{(x - \mu - d_\alpha/2)^2}{40} f(x) dx + \int_{-\infty}^{\mu-d_\alpha/2} \frac{(\mu - d_\alpha/2 - x)^2}{10} f(x) dx,$$

where $f(x)$ is the probability density at x of a normally distributed random variable with mean μ and variance σ^2 .

The computation of $R(d_\alpha)$ has to be done numerically, and a plot of $D(d_\alpha)$ versus d_α , for $d_\alpha \geq 0$, is shown in Figure 2.

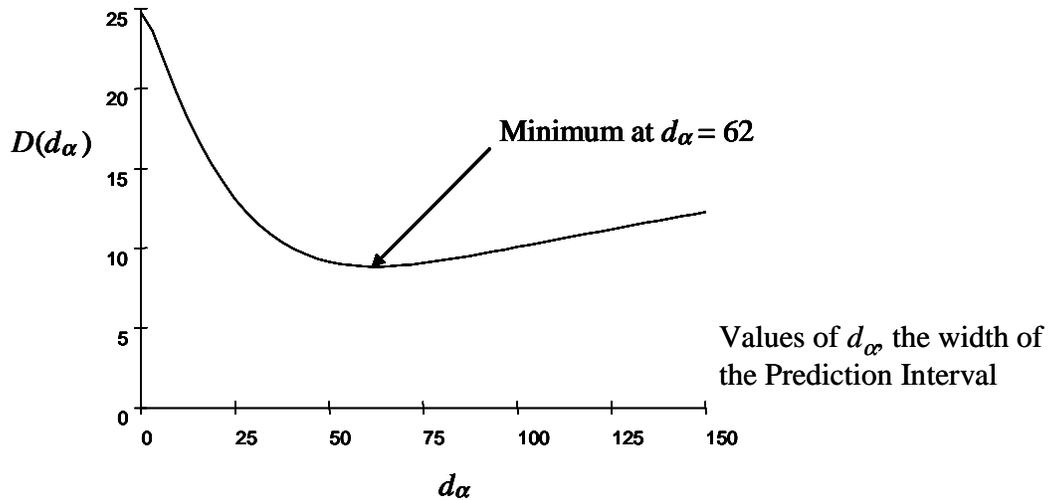


Figure 2: Total Expected Disutility versus d_α

An examination of Figure 2 shows that $D(d_\alpha)$ attains its minimum at $d_\alpha = 62$. This suggests, via the relationship $d_\alpha = 2z_{\alpha/2}\sigma$ with $\sigma^2 = 396$, and a table look up in the standard normal distribution, that the optimal coverage probability for this scenario is 0.88. Using coverage probabilities other than $0.88 \approx 0.90$, say the conventional 0.95 or

0.99 would yield a wider interval but the utility of such intervals would be less than that provided by the 0.90 coverage probability.

5 Generality of the Approach and Some Caveats

The proposed approach is general because it rests on the simple principle of minimizing $D(d_\alpha)$, the total expected disutility function – Equation (3.4). If $D(d_\alpha)$ attains a unique minimum, then a unique optimal coverage probability can be arrived upon. If the minimum is not unique then several optimal coverage probabilities will result, and the user is free to choose any one of these. There could be circumstances under which $D(d_\alpha)$ will not attain a minimum, and the method will fail to produce an answer. The optimality conditions which ensure a minimum value of $D(d_\alpha)$ is a matter that needs to be formally addressed, but with $c(d_\alpha)$ monotonic and concave (or convex), and with $L(d_\alpha, x)$ U -shaped as shown in Figure 1, $D(d_\alpha)$ will indeed attain a minimum. The choice of $L(d_\alpha, x)$ prescribed in Equation (3.1) is quite general. It is easily adaptable to one sided intervals, and also to the inventory and banking scenarios mentioned before. Furthermore, it is conventional in life-length prediction studies and in statistical inference wherein square error loss is a common assumption.

The assumed distribution of X with specified parameters plays two roles. One is to average out $L(d_\alpha, x)$ to produce the risk function $R(d_\alpha)$. In this role the choice of the distribution of X is not restrictive because its purpose here is to merely serve as a weighting function. Any well known distribution can be used, especially when $R(d_\alpha)$ is obtained via numerical methods, as we have done with the normal. By contrast, frequentist prediction intervals that entail pivotal methods limit the choice of distributions. The second role played by the distribution of X , is to facilitate a relationship between d_α and α . In the case of the normal distribution with mean μ and variance σ^2 , $d_\alpha = 2z_{\alpha/2}\sigma$; here μ does not matter. This type of relationship will arise with any symmetrical distribution, such as the Student's- t , the triangular, the uniform, the Laplace, etc. A relationship between d_α and α in the case of the exponential with scale λ turns out to be quite straightforward, indeed

more direct than that encountered with the normal; specifically $d_\alpha = \frac{1}{\lambda} \log[(2 - \alpha)/\alpha]$. By suitable transformations, the case of other skewed distributions such as the lognormal, the Weibull, and the chi-squared can be similarly treated. A difficult case in point is the Pareto distribution (popular in financial mathematics) wherein $P(X > x; \psi, \beta) = (\psi/(\psi + x))^\beta$. Here $d_\alpha = \psi[(1 + \alpha/2)^{-1/\beta} - (\alpha/2)^{-1/\beta}]$, and the relationship between d_α and α is involved for the method to be directly invoked.

Finally, besides the caveat of $D(d_\alpha)$ not having a minimum, the other caveat is the dependence of an optimal coverage probability on data. Specifically, the use of a posterior distribution of X to obtain $R(d_\alpha)$ makes this latter quantity depend on the observed data with the consequence that in the same problem one could conceivably end up using a different coverage probability from forecast to forecast. Unattractive as this may sound, it is the price that one must pay to ensure coherence. However, this dependence on the data becomes of less concern once the posterior distribution of X converges, so that the effect of the new data on the posterior diminishes. The same situation will also arise when the distribution of X is specified via a frequentist approach involving a plug in rule.

6 Summary and Conclusions

The thesis of this article is to argue that choosing coverage probabilities for prediction intervals should be based on decision theoretic considerations. The current practice is to choose these by convention or astute judgment. Prediction intervals are a part of the essentials of regression, time series, and state space models. They also occur in conjunction with previsions based on probability models entailing the judgment of exchangeability. Furthermore, the principles underlying the construction of prediction intervals share some commonality with those involving inventory planning and banking reserves.

The decision theoretic approach boils down to the minimization of total expected disutility. This disutility consists of two components. One is a disutility associated with the width of the interval and the other is associated with the failure of an interval to cover the observed value when it reveals itself. The proposed approach is illustrated via a con-

sideration of stylized utility functions. It can be seen as a prototype for approaches based on other utility functions. The approach also entails a use of the normal distribution to describe the uncertainties. Again, this distributional assumption is not essential; other distributions will work equally well.

We emphasize that the material here pertains to prediction intervals, not confidence intervals. It would be interesting to develop a decision theoretic approach for choosing the confidence coefficient of a confidence interval. To the best of our knowledge, this remains to be satisfactorily done.

Acknowledgements

Supported by ONR Contract N00014-06-1-037 and the ARO Grant W911NF-05-1-0209. The student retention problem was brought to our attention by Dr. Donald Lehman. The detailed comments of three referees and an Associate Editor have broadened the scope of the paper. Professor Fred Joust made us aware of the papers by Granger, and by Tay and Wallis.

References

- [1] Beran, R. (1990), “Calibrating Prediction Regions,” *Journal of the American Statistical Association*, 85, 715–23.
- [2] Box, G. E. P., and Jenkins, G. M. (1976), *Time Series Analysis: Forecasting and Control*, San Francisco, CA: Holden-Day.
- [3] Chatfield, C. (1993), “Calculating Interval Forecasts,” *Journal of Business and Economic Statistics*, 11, 121-135.
- [4] Decamps, J. P., Rochet, J. C., and Roger, B. (2004), “The Three Pillars of Basel II: Optimizing the Mix,” *Journal of Financial Intermediation*, 13, 132-155.
- [5] de Finetti, B. (1972), *Probability, Induction and Statistics*, New York: Wiley.
- [6] De Groot, M. H. (1970), *Optimal Statistical Decisions*, New York: McGraw-Hill.
- [7] Gardner, Jr., E. S. (1988), “A Simple Method of Computing Prediction Intervals for Time Series Forecasts,” *Management Science*, 34, 541-546.
- [8] Granger, C. W. J. (1996), “Can We Improve the Perceived Quality of Economic Forecasts?” *Journal of Applied Econometrics*, 11, 455-473.
- [9] Hadley, G., and Whitin, T. M. (1963), *Analysis of Inventory Systems*, Englewood Cliffs, NJ: Prentice-Hall.
- [10] Hahn, G. J., and Meeker, W. Q. (1991), *Statistical Intervals: A Guide for Practitioners*, New York: Wiley.
- [11] Lawless, J. F., and Fredette, M. (2005), “Frequentist Prediction Intervals and Predictive Distributions,” *Biometrika*, 92, 529-542.
- [12] Lehmann, E. L. (1958), “Significance Level and Power,” *The Annals of Mathematical Statistics*, 29, 1167-1176.
- [13] Lindley, D. V. (1985), *Making Decisions* (2nd ed.), New York: Wiley.

- [14] Pearson, K. (1920), “The Fundamental Problem of Practical Statistics,” *Biometrika*, 13, 1-16.
- [15] Savage, L. J. (1962), “Bayesian Statistics,” in *Recent Developments in Information and Decision Processes*, eds. R. E. Machol and P. Gray, New York: The Macmillan Company, pp. 161-194.
- [16] Singpurwalla, N. D. (2006), *Reliability and Risk: A Bayesian Perspective*, England: Wiley.
- [17] Tay, A. S., and Wallis, K. F. (2000), “Density Forecasting: A Survey,” *Journal of Forecasting*, 19, 235–254.
- [18] Winkler, R. L. (1972), “A Decision-Theoretic Approach to Interval Estimation,” *Journal of the American Statistical Association*, 67, 187-191.