Combining Risk Estimates for Binary Outcome Events in Project Management

b. Type of submission

- Refereed Research Paper—Abstract if rejected

c. Track that best fits the submission: Product/Process Innovation and Project Management

Stage today – Analysis Stage

Conference stage – Completed Paper

ABSTRACT

Managing risks is an important part of project management in particular and management in general. Part of risk management is assessing the probability of risk events occurring. One method of assessing risks is for project managers use historical databases and expert opinions to obtain probability estimates for events. When experts are used, best practice dictates probability distributions are obtained to represent the likelihood of a risk event occurring. However, obtaining probability distributions from multiple experts is not always feasible due to cost or time constraints. A less expensive alternative is to use more readily available single point probably estimates to derive a probability distribution. This study compares a new method for combining probability estimates to established methods. The new method proposed by Ranjan and Gneiting (2010) outperforms existing methods on multiple measures.

INTRODUCTION

Although there is a transition to relying on probability distributions for forecasting (Gneiting, 2008), single point estimates for common risk events, such as the weather, are cheaper and easier to obtain. Most project managers do not have the luxury of obtaining full probability distribution for risk events. For example, risk management decisions tied to weather, such as concrete pouring, asphalting, roofing, and outdoor events are situations where project managers rely on single point probability estimates obtained from multiple experts. Forecasts related to economic recessions, agriculture prices, currency exchange rates, and political elections are similar to the weather in that obtaining probability distribution from multiple experts is not always feasible. A common factor of these events is that underestimating an event’s true probability can result in costly rework while over estimating can result in lost revenue opportunities.

This paper introduces the field of project management to a method for combining probability forecasts proposed by Ranjan and Gneiting (2010). It holds the promise of improving risk management in a range of contexts. The sections below discuss risk estimation techniques, common methods for combining probability estimates, and the new method for combining probability estimates. The new method is compared to popular methods using a unique dataset that provides an opportunity to measure the effect of improved accuracy using measures that managers are familiar with, return on investment and total monetary earnings.

RISK MANAGEMENT

A common definition of a risk is an uncertain event or condition that, if it occurs, will affect one or more project objectives (PMI, 2008). Risks can be positive or negative events. Project risk management includes the processes of planning the risk management effort, identifying risks, prioritizing risks for further analysis, probability of risk events, developing risk response plans, monitoring risks, executing planned responses to risk, and identifying new risks.

Although all parts of the risk management process are important, this study focuses on the process of quantifying the probability of risk events. The estimates of risks to high-level project objectives are derived from the probabilities assigned to one or more risks. Often, subjective probability estimates from multiple experts are used to assign probabilities to individual risks. Two popular methods for developing risk estimates, event trees and simulations, are discussed below.

Event Trees

Event trees, decision trees, or probability trees, are used to develop estimates or risk events occurring (Pate-Cornell, 1984). Usually, these estimates are expressed as probabilities. The tree development process begins by placing a risk event (e.g., loss of electricity) at the top or left side of a chart. The second level of the chart includes an event that might result from the risk event (e.g., electric generator takes over or the generator fails) and the probabilities associated with the events occurring (e.g. 90% probability the generator takes over, 10% it fails). The next level of the chart lists possible outcomes for each of the events listed in the second level and estimated probabilities for the events occurring. The decomposition process continues until all likely effects of the root event are documented. By combining the probabilities of the associated branches in the chart, it is possible to compute a probability score for each identified outcome. A project manager incorporates the computed probabilities into the risk management plan. Ideally, the project manager derives the probabilities for each branch in the tree from historical databases. However, as projects are by definition unique events, subjective probability estimates from experts often are substituted for observed probabilities.

Simulations

Stochastic simulation models are computer simulations that take as input ranges of possible values for activity durations, an associated probability distribution (e.g., normal, beta, triangle, etc.), and sometimes a correlation coefficient to capture a relationship between two or more activities (Kwak & Ingall, 2007). These simulations often are called Monte Carlo simulations, because they use a random number generator to assign durations to each activity in the project network. A single experiment consists of the software selecting duration estimates for all activities and calculating the total project duration and costs. Thousands of experiments can provide insight into the probability of achieving high-level project objectives and help identify sensitive activities that require additional risk management.

Simulations and event trees often use subjective probabilities elicited from experts to find empirical probability distributions for high-level project objectives. Because the inputs to schedule simulations are framed as duration estimates instead of probability estimates, the dependency on probability estimates is not apparent. Scheduling simulations provide experts with a probability distribution, with Beta being the most common, and elicit duration estimates for points on the distribution curve: 5%, 50%, and 95% being the most common points. The relationship between event tree probability estimates and duration estimate entered into simulations is often overlooked. In practice, there is no difference between providing experts with a probability distribution and asking them to provide duration estimates to match points in the distribution and asking expert to provide a probability estimate for a risk event based on an assumed probability distribution. The next section discusses common methods for combining probability estimates provided by experts.

COMBINING PROBABILITY ESTIMATES

Because simulations, event trees, and other methods rely on subjective, single point probability estimates, best practice dictates combining estimates from multiple experts. However, there is little guidance in the project management literature for how to combine the subjective probability estimates. Next, we review methods for combining probability estimates and discuss a new method for addressing this problem. That discussion is followed by a comparison of the methods discussed.

Linear Combination

One of the simplest methods for combining expert opinions is to linearly combine the probabilities to create a simple average (i.e., $w_1,\ldots,w_k = 1/k$). This computationally simple process offers an improvement on relying on a single expert. However, this weighting scheme does not take into account additional knowledge of the historical accuracy, trustworthiness of the expert, or other relevant factors. Another approach is to use a unique weight for each expert, to create a linear opinion pool where,

$$p = w_1 p_1 + \ldots + w_k p_k, \quad w_1,\ldots,w_k \geq 0, \quad w_1 + \ldots + w_k = 1 \qquad (1)$$

Utilizing the linearly weighted combination of estimates can result in better outcomes than relying on a single expert, but it is not optimal. An issue that arises is that individual probabilities may be drawn from different distribution shapes, e.g., normal, uniform, monotonic, etc. While forecasts from each source can show a high degree of predictive accuracy, when they are combined into a single estimate, the accuracy can be reduced (Ranjan & Gneiting, 2010).

A combination method frequently discussed in the project management literature is the formula used in the Program Evaluation and Review Technique (PERT). PERT is a scheduling technique that allows manager to do more than develop an estimate for a single activity and it is by far the most discussed method for combining estimates. Three estimates are required for the PERT formula. The weightings consist of a pessimistic weight of 1/6, a most likely weight of 4/6, and an optimistic weight of 1/6 (Johnson, 1997; Malcom, Roseboom, Clark, & Fazar, 1959). However, the PERT distribution is not a linear combination at its core, but is derived from the Beta distribution (Johnson, 1997). The PERT formula attempts to reweight the expert estimate, on a static or a priori basis, based on the predictive capability of different point estimates.

Many researchers have documented the shortcomings of the PERT method. One of the commonly reported issues is an inherent optimistic bias (Trietsch & Baker, 2012). Another limitation of PERT is that its weightings do not take into account any additional knowledge that may be available to the project manager conducting the PERT analysis, such as the skills of the expert or the expert's risk propensity. A third limitation is that there is little research documenting the utility of the PERT formula for forecasting binary events. These limitations and others reduce the effectiveness and reliability of PERT's weighting formula.

Many previous studies have provided different versions of PERT to either increase its accuracy (Trietsch & Baker, 2012) or reduce the data load (Cottrell, 1999) by using two data points. These improvements to PERT can come at a price, such as computational complexity or reducing the number of data points used, which will likely reduce the accuracy of the forecast (Ranjan & Gneiting, 2010; Wallsten & Dierderich, 2001). The next section discusses a new method for combining expert estimate proposed by Ranjan and Gneiting (2010).

Beta Transformed Linear Opinion Pool

Ranjan and Gneiting's (2010) binary linear opinion pool (BLP) method is a new method for combining probability estimates obtained from multiple experts or models. For brevity, we do not include a detailed explanation of the BLP methodology. The following description provides a basis for discussing the BLP equation by showing the origin of the weightings and shape factors for BLP. We refer the reader to (Ranjan & Gneiting, 2010) for the complete formulation and background. The BLP method uses a parametric approximation by aggregating the individual probabilities, $p_1, \ldots, p_k$, into a linear opinion pool and applies a beta transformation to calibrate the weights, $w_1, \ldots, w_k$, and α and β where

$$p = H_{\alpha,\beta}\left(\sum_{i=1}^{k} w_i p_i\right) \tag{2}$$

$$H_{\alpha,\beta}(x) = B(\alpha,\beta)^{-1}\int_0^x t^{\alpha-1}(1-t)^{\beta-1}\,dt \qquad \text{for } x \in [0,1] \tag{3}$$

Equations (2) and (3) are transformed into a log-likelihood function in equation (4) to yield the weights and shape parameters alpha and beta.

$$\sum_{t=1}^{n} y_t \log\left\{H_{\alpha,\beta}\left(\sum_{i=1}^{k} w_i p_{it}\right)\right\} + \sum_{i=1}^{n}(1-y_t)\log\left\{1 - H_{\alpha,\beta}\left(\sum_{i=1}^{k} w_i p_{it}\right)\right\} \tag{4}$$

BLP, as is typical of most calibration methods, uses part of the dataset to estimate the weights and values of α and β. With the recalibrated weightings and modified shape parameters, BLP can be used to create calibrated probability estimates.

METHODOLOGY

The dataset used in this study was compiled from records of three experts' predictions for thoroughbred horse races in North America that took place between April 2008 and October 2010 at a single racetrack. This dataset was selected because it represented problems with similar characteristics to the problems facing risk managers in which multiple experts provide probability estimates for the same event (picking the winning horse). In addition, each event has a definitive outcome (if the selected horse won) tied to a revenue gain (win payout) or loss that is tied to the uncertainty surrounding the event.

The two of the experts charged the same amount for their day-of-race predictions ($10 for about 10 races) and their predictions for the previous day's races were free. The third expert was the track expert, who provided predictions were free. Expert 1 provided estimates for five horses in each race. Expert 2 provided probability estimates for between one and five horses. Expert 3, the track expert, provided probabilities for all the entrants. To filter the dataset to limit it to cases compatible with the planned combination procedures, each race needed to be associated with probability estimates from all three experts. To address this issue, experts were assessed on their ability to select the winning horse. From the dataset, Expert 1 correctly selected the winner 710 times out of a possible 2001 races. Expert 2 correctly selected the winner 654 times. Expert 3 was the track expert. This expert did not always select a favorite to win the race. The dataset was filtered to include the probabilities for the horse that Expert 1 predicted to win. If Expert 2 did not provide a probability for Expert 1's predicted winner, that case was excluded. This filtering process yielded 2001 usable cases.

In addition to evaluating combination methods using traditional scoring rules, this study set out to evaluate the combination methods using financial measures of total earnings and total return on investment. This required a method for selecting risks worth taking. The combined probability estimate was chosen as the selection mechanisms. The observed frequency of Expert 1's first selection winning the race across half of the cases, which were used as training cases to develop the weights for the BLP method, was chosen as the threshold for placing a wager. The threshold value was established by using Expert 1's correct prediction of the winner, 35.66%, in the training set. If the any combination method resulted in a probability estimate that exceeded 0.3566, no bet was placed under that method. See the formulations below and Table 1 for all of the combinations.

$$\text{If } f(E1_{1i}, E2_{1i},\ E3_{1i}) < T \qquad B_i = 1$$

$$\text{If } f(E1_{1i}, E2_{1i},\ E3_{1i}) > T \qquad B_i = 0$$

Where:

$E1_{1i}$ is Expert 1's winning horse selection for race *i*

$E2_{1i}$ is Expert 2's probability of $E1_{1i}$'s horse for race *i*

$E3_{1i}$ is Expert 3's probability of $E1_{1i}$'s horse for race *i*

T is the threshold value

B_i is the binary value of a bet being placed for race *i*

These values were then compared to actual horse that won the race, such that,

$$\text{If } E1_{1i} = W_i;\ P_i = 1$$

$$\text{If } E1_{1i} \neq W_i;\ P_i = 0$$

Where:

W_i is the winner of race *i*

P_i is the binary value of the race outcome

Table 1 Race Betting and Outcome Scores

	$E1_{1i}=W_i$	$E1_{1i}\neq W_i$
Bet $f(E1_{1i}, E2_{1i}, E3_{1i})<T$	$B_i=1$ $P_i=1$	$B_i=1$ $P_i=0$
No Bet $f(E1_{1i}, E2_{1i}, E3_{1i})>T$	$B_i=0$ $P_i=1$	$B_i=0$ $P_i=0$

The combining of the three experts' point estimates for BLP and PERT were compared using their cumulative Brier scores (Brier, 1950). This method of comparison was selected to remain consistent with Ranjan and Gneitning (2010).

$$Brier\ score = \frac{\sum_{i=1}^{N}(B_i - P_i)^2}{N}$$

Weightings

Expert predictions were combined using three methods: 1) a linear combination in which all experts have a weighing of 1/3, and 2) PERT, and 3) BLP method. For the PERT combination Expert 1's prediction was always the probability weighted as realistic, receiving 4/6 of the total with the other experts having 1/6 each. This weighting scheme was used because Expert 1 was determined to be the best predictor of the race winner, thus his predictions should be given the highest weighting. In addition, Expert 1 probability estimates resulted in the lowest Brier score (Brier, 1950) over the training cases. For the calibration of BLP, half the cases (1001) were used as training cases. The new weightings for the experts were: Expert 1, 0.5368733; Expert 2, 0.2449975; and Expert 3, 0.2181292. The α and β parameters were 1.171598and 1.543399 respectively. Using the above weightings, we simulated a series of decisions based on probabilities derived using the three methods being compared.

ANALYSIS

Each race in the dataset represents a risk event that is analogous to a project manager deciding to pursue or postponing weather-dependent work. Both of these decisions are based on events with binary outcomes and are made based on a combination of probability estimates provided by outside experts (i.e., weather forecasters). If the decision is to pursue work and the weather is favorable, the decision is a win for the project and the revenue flow is positive. Alternatively, if the decision is to work and the weather turns unfavorable, results in a loss of revenue. This is the same structure as the decision to place a bet on a horse or not.

The first measure of comparison is the Brier score (Brier, 1950), a measure of mean square error of probability forecasts. It is a commonly used measure of a prediction method's performance. Scores range from 0 for a perfect score to 1 for the worse forecast. While the Brier score is widely used, the importance of small improvement is not apparent. There is not a widely accepted test for determining significant difference between two probability estimation methods. As an additional measure of the differences among the combination methods, two financial measures were used: total return on betting and return on investment (ROI). Total return on betting is the total amount earned based on a simulated $2.00 bet for the key horse (Expert 1's first selection) to win the race, less the amount bet. A winning bet is when B_i=1 and P_i=1 and the losing bet is when B_i=1 and P_i=0. The simulated bets were limited based the frequency of Expert 1's first selection winning in the 1,000 races used to train the BLP algorithm. The third measure used to compare the measure of difference was a simple return on investment (ROI) percentage.

RESULTS

Table 2 contains the results of applying the three combination techniques used in our analysis and a comparison to the probabilities extracted from the pari-mutuel market at the last possible point someone could have chosen to place a bet. Although it is unrealistic that someone could obtain and use the market-derived probabilities for Expert 1's first section, but it is included as a comparison. BLP outperformed averaging and PERT combination techniques as evidenced by a lower Brier score, fewer losses, a higher total return on betting, and a greater return on investment. BLP outperformed the Pari-mutuel market on all measure except the number of losses.

Table 2 Linear Combination, PERT, BLP, and Pari-mutuel Market Comparison

	Linear Combination	PERT	**BLP**	Pari-mutuel Market
# of Losses	542	484	**354**	307
# of Bets Placed	831	744	**534**	432
Brier Score	0.2380	0.2361	**0.2319**	0.2325
Total Return on Betting	-$68.20	-$12.80	**$28.20**	$16.20
Return on Investment	-4.10%	-0.86%	**2.64%**	1.87%

DISCUSSION AND CONCLUSION

This article addressed the topic of combining probability estimates in a project management context. It started with a discussion of event trees and simulations, two widely used techniques for quantifying the probability of a risk event occurring. Because it is common for managers to receive multiple probability estimates for the same event, two established methods for combining probability estimates were discussed. In addition, a new method for combining probability estimates was described. An analysis comparing the three combination methods was conducted. The results provided evidence that the new method produced probability estimates superior to the established methods.

Managers who rely on probability estimates from multiple sources such as historical databases and outside experts benefit from improved methods for combining estimates. Weather, commodity prices, interest rates, policy changes, and political events are events where managers must make decisions based on probability estimates from multiple sources. Under ideal circumstances, decision makers can acquire and combine probability distributions representing the likelihood of events. Acquiring full probability distributions is not always feasible due to time or cost constraints. In these cases, managers must rely on single point probability estimates provided by independent experts. The BLP method can help managers improving the accuracy of combined estimates, which can improve the financial performance of the project and the overall organization.

Additional research is required to understand the limits of the BLP method. In particular, research is needed to address questions related to the number of training cases required and how to handling missing data. Adapting the method for use in controlling schedule variance is another area where research can shed light on the applicability of the new method. Those interested can contact the author to obtain a copy of the programming code and data used in this study. These steps should assist the research community in replicating and extending this work.

REFERENCES

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1–3.

Cottrell, W. D. (1999). Simplified program evaluation and review technique. *Journal of Construction Engineering and Management,* 125(1), 16-22.

Johnson, D. (1997). The triangular distribution as a proxy for the beta distribution in risk analysis. *The Statistican,* 46, 387-398.

Kwak, Y. H., & Ingall, L. (2007). Exploring monte carlo simulation applications for project management. *Risk Management*, 9, 44-57.

Malcom, D. G., Roseboom, C. E., Clark, C. E., & Fazar, W. (1959). Application of a technique for research and development program evaluation. *Operations Research,* 7, 646-649.

Pate-Cornell, M, (1984). Fault tree vs. event trees in reliability analysis. *Risk Analysis*, 4(3), 177–186.

Project Management Institute. (2008). A guide to the project management body of knowledge (PMBOK Guide) (Fourth ed.). Newton Square, PA: Project Management Institute.

Ranjan, R., & Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 72, Part 1, 71-91.

Trietsch, D., & Baker, K. R. (2012). PERT 21: Fitting PERT/CPM for use in the 21st century. *International Journal of Project Management,* 30, 490-502.

Wallsten, T. S., & Dierderich, A. (2001). Understanding pooled subjective probability estimates. *Mathmatical Social Sciences,* 41(1), 1-18.