Calibrating Subjective Probabilities Using Hierarchical Bayesian Models

Edgar C. Merkle

Department of Psychology
Wichita State University
Wichita, KS 67260-0034
edgar.merkle@wichita.edu
http://psychology.wichita.edu/merkle

Abstract. A body of psychological research has examined the correspondence between a judge's subjective probability of an event's outcome and the event's actual outcome. The research generally shows that subjective probabilities are noisy and do not match the "true" probabilities. However, subjective probabilities are still useful for forecasting purposes if they bear some relationship to true probabilities. The purpose of the current research is to exploit relationships between subjective probabilities and outcomes to create improved, model-based probabilities for forecasting. Once the model has been trained in situations where the outcome is known, it can then be used in forecasting situations where the outcome is unknown. These concepts are demonstrated using experimental psychology data, and potential applications are discussed.

Key words: Subjective probability, confidence, Bayesian methods, calibration, expert judgment

1 Introduction

Subjective probability is commonly used to measure judges' certainties in decisions and forecasts. People are generally familiar with reporting such probabilities, making them a natural way to gauge certainty in many situations. This has led to a long line of psychology research devoted to understanding how individuals construct subjective probabilities, where it is often found that subjective probabilities tend to be larger than the true probabilities of the corresponding outcomes (e.g., [1–3]).

The intent of this paper is to study the use of a hierarchical logistic model for improving subjective probabilities. The model transforms individual subjective probabilities into predicted probabilities of an outcome's occurrence. Once the model has been fit to data with known outcomes, the model can be used to transform subjective probabilities and forecast unknown outcomes. A particularly-interesting aspect of the model is that it yields unique transformations for individual judges, accounting for individual differences in response styles while maintaining general trends present in the group of judges.

The model is related to the vast literature on combining expert judgments (see, e.g., [4, 5]), where the goal is to take many subjective judgments as input and yield a single, aggregated prediction as output. The current paper differs from this literature in that it examines how *individuals'* subjective probabilities are related to true probabilities of particular outcomes. Thus, the model in the current paper may be used to improve individual expert probabilities prior to aggregating the probabilities (an idea advanced by [6]). A special case of the model may also be used in situations where only a single expert reports a probability.

In the pages below, I first define measures of the correspondence between subjective probabilities and outcomes, along with the measures' use in applications. I then define the model that is used to transform subjective probabilities. Next, I demonstrate the utility of the approach using data from a visual discrimination experiment. Finally, I describe how the model can be used in applications and consider other statistical methods that could be relevant.

1.1 Correspondence Between Subjective Probability and Outcomes

Researchers have defined many measures of the correspondence between probabilistic forecasts and outcomes. One of the most intuitive measures is of the extent to which probabilistic forecasts match the long-term proportion of occurring outcomes. This can be defined mathematically as a measure of *bias.* Let $d_j \in \{0, 1\}$ be the outcome of event j $(j = 1, \dots, J)$,[1] and let f_j be a judge's subjective probability that $d_j = 1$. For the purposes of this paper, bias is then defined as:

$$\text{bias} = \overline{f} - \overline{d}, \tag{1}$$

where $\overline{f}$ is the mean of the f_j and $\overline{d}$ is the mean of the d_j $(j = 1, \dots, J)$. Biases close to zero reflect "good" forecasts, and biases far from zero reflect "bad" forecasts.

There exist a variety of other measures designed to examine other aspects of the correspondence between the f_j and the d_j; see [7]. Two of these measures are *slope* and *scatter*. Slope measures the extent to which forecasts differ for $d_j = 0$ and $d_j = 1$:

$$\text{slope} = \overline{f}_1 - \overline{f}_0, \tag{2}$$

where $\overline{f}_1$ is average subjective probability for events where $d_j = 1$ and $\overline{f}_0$ is average subjective probability for events where $d_j = 0$. Large slopes reflect good forecasts, and small slopes reflect bad forecasts.

Scatter reflects noise in the f_j that is unrelated to the d_j:

$$\text{scatter} = \frac{(n_1 - 1)s^2_{f_1} + (n_0 - 1)s^2_{f_0}}{n_1 + n_0 - 2}, \tag{3}$$

[1] *Outcome* has multiple meanings. It could refer to whether or not an event occurs, in which case we have 0=event does not occur, 1=event does occur. Alternatively, *outcome* could refer to whether or not a judge's prediction of an event's occurrence matches the event's actual occurrence. In this case, we have 0=judge's prediction was incorrect, 1=judge's prediction was correct.

where $s^2_{f_1}$ is the variance of the f_j for which $d_j = 1$, n_1 is the number of events for which $d_j = 1$, and $s^2_{f_0}$ and n_0 are defined similarly. Small values of scatter reflect good forecasts, and larger values reflect bad forecasts.

1.2 Use of the Measures

Decision researchers have tended to focus on the bias measure: bias is generally intuitive, and observed bias may be immediately compared to the "perfect" bias value of 0. In a variety of experimental tasks, researchers tend to find biases greater than zero; that is, judges' subjective probabilities tend to be larger than they should [7–10]. This implies that, in applied situations, subjective probabilities are suboptimal for guiding decisions. Less research has focused on measures other than bias (though see, e.g., [6, 11]), which may be because it is generally impossible for judges to attain perfect values on these other measures. Thus, it is difficult to say whether a particular value of slope or scatter is good. It is still possible to compare relative magnitudes of slope and scatter, making them useful for comparing observed slope and scatter with model-predicted slope and scatter. These measures are used in the example that follows, but I first describe the specific model that is used to transform the subjective probabilities.

2 Model

Let i index judges and j index forecasts. To transform subjective probabilities, I consider a hierarchical logistic model with f_{ij} as a predictor variable and d_{ij} as a response variable. The basic model is given as:

$$\begin{aligned} d_{ij} &\sim \text{Bernoulli}(p_{ij}) \\ \log(p_{ij}/(1-p_{ij})) &= b_{0i} + b_{1i} f_{ij}, \end{aligned} \tag{4}$$

where p_{ij} is the probability that judge i is correct on forecast j. This probability is modeled using the judge's subjective probability, f_{ij}, as a predictor. The slope and intercept in the model vary for each judge i, allowing the relationship between p and f to differ from judge to judge. The hierarchical formulation of the model is obtained by assuming a joint normal distribution on the b_{0i} and b_{1i}:

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim \text{N}\left[\begin{pmatrix} B_0 \\ B_1 \end{pmatrix}, \mathbf{\Sigma}_b = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix} \right], \tag{5}$$

where B_0 is the mean of the intercepts, B_1 is the mean of the slopes, and $\mathbf{\Sigma}_b$ is the covariance matrix of the intercepts and slopes. The hierarchical normal distribution is traditionally used in this model, but a different distribution could be used if deemed useful or necessary. This flexibility in hierarchical distributions is an advantage of the Bayesian approach.

There are two other Bayesian advantages that led to the implementation of a Bayesian model here. First, the Bayesian model allows for incorporation of prior knowledge about base rates of correct forecasts or about relationships between p and f. This could be useful for specific applications. Second, the Bayesian model allows for calculation of posterior predictive distributions of the p_{ij}. As will be shown below, this allows us to systematically transform judges' reported probabilities into conservative and/or liberal probabilistic predictions. To complete the Bayesian model, we require prior distributions on B_0, B_1, and the associated covariance matrix $\mathbf{\Sigma}_b$. These are given as:

$$B_0 \sim \mathrm{N}(\mu_0, s_0^2) \tag{6}$$

$$B_1 \sim \mathrm{N}(\mu_1, s_1^2) \tag{7}$$

$$\mathbf{\Sigma}_b \sim \text{Inv-Wishart}(\text{df}, \mathbf{\Sigma}_0), \tag{8}$$

where $\mathbf{\Sigma}_b$ follows an inverse Wishart distribution with df > 1 and scale matrix $\mathbf{\Sigma}_0$. These parameters can be set based on prior knowledge about the forecasting scenario, or they can be set to reflect the absence of prior knowledge. I consider the latter situation in the following example.

3 Example: Visual Discrimination

To demonstrate the potential applicability of these hierarchical models, I consider data from Experiment 1 of [12]. In this experiment, judges viewed images of asterisks randomly arranged in a 10 × 10 array. The number of asterisks was randomly drawn from one of two normal distributions, with the first distribution being $\mathrm{N}(45, \sigma = 5)$ and the second being $\mathrm{N}(55, \sigma = 5)$. For each of 450 trials, judges viewed an array and stated a probability that the asterisks arose from the second distribution. Judges were not told the distributions governing number of asterisks; they were required to learn the distributions by themselves. Reported probabilities were required to come from the set $\{.05, .15, .25, \ldots, .95\}$. Choices were inferred from the reported probabilities, and probabilities in the choices were then obtained (ranging from .55 to .95).

3.1 Model Details

The Bayesian hierarchical model described in the previous section was fit to data from 36 subjects across 40 experimental trials. The prior distributions on model parameters were taken to be noninformative:

$$\mu_0 \sim \mathrm{N}(1, 1.0\mathrm{E}5) \tag{9}$$

$$\mu_1 \sim \mathrm{N}(0, 1.0\mathrm{E}5) \tag{10}$$

$$\mathbf{\Sigma}_b \sim \text{Inv-Wishart}(2, \mathbf{I}), \tag{11}$$

where $\mathbf{I}$ is a 2 × 2 identity matrix. The model was estimated in OpenBugs [13] via Markov chain Monte Carlo methods, with relevant code appearing in the appendix. Three chains of parameters were sampled for $14,000$ iterations each, with the first $4,000$ iterations being discarded as burn-in.

OpenBugs was also used to obtain predicted probabilities for all judges across 20 trials that were not used during model fitting. To be specific, OpenBugs was used to sample from the posterior predictive distributions for these 20 trials. These distributions can be used to obtain conservative and/or liberal predicted probabilities. The extent to which this is useful is examined below.

3.2 Results

Results are presented in two parts. First, I make some general remarks about the fitted model and the probabilistic predictions. I then make detailed comparisons between the predicted probabilities and judges' reported probabilities.

Fitted Model Before examining the predicted probabilities, a fundamental issue involves the extent to which reported probabilities (f_{ij}) are related to accuracy (d_{ij}). Within the model (Equation (4)), the hierarchical distribution on the b_{1i} informs this issue. This distribution is estimated as $\text{N}(3.1, \widehat{\sigma_1^2} = 0.96)$, with a 95% posterior interval for the mean being $(2.24, 4.09)$. Because the interval is far from zero, we have evidence that the f_{ij} are indeed useful for predicting accuracy. Further evidence comes from the estimated b_{1i} for each judge. All 36 of these estimates are positive, with no 95% posterior intervals including zero.

Now that a predictive relationship between the f_{ij} and d_{ij} has been established, we can examine the extent to which the model's probabilistic predictions are an improvement over the f_{ij}.

Probabilistic Predictions In this section, the model's probabilistic accuracy predictions, $\widehat{p_{ij}}$, are compared to the f_{ij} for the 20 trials that were excluded from model estimation. Figure 1 displays the observed f_{ij} versus the $\widehat{p_{ij}}$ for all 36 judges. The diagonal line in the graph is the identity line, reflecting instances where $f_{ij} = \widehat{p_{ij}}$. Considerable variability is observed in the mapping from f_{ij} to $\widehat{p_{ij}}$ for different judges. Further, the f_{ij} and $\widehat{p_{ij}}$ differ primarily for large f_{ij}: in these cases, the $\widehat{p_{ij}}$ are smaller. Thus, the model compresses the range of probabilities.

As stated previously, the model's posterior predictive distributions of $\widehat{p_{ij}}$ were obtained for the 20 trials excluded from model estimation. We can summarize these distributions in various ways to obtain predictive probabilities. A common summary involves taking the means of the posterior distributions. Alternatively, if we want more conservative predictive probabilities, we can take the 25th percentile of these distributions, for example. Slope, scatter, and bias statistics were calculated for these two types of posterior summaries, along with statistics for the observed f_{ij}. These serve as measures of the extent to which the model predictions are improvements over the f_{ij}.

Figure 2 contains histograms of the difference between each judge's observed statistics and model-predicted statistics (using the means of the posterior predictive distributions). Values greater than zero reflect instances where a judge's

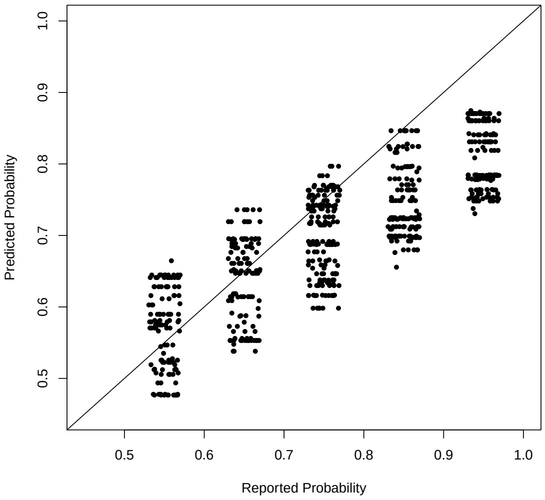

Fig. 1. Model mappings from reported probabilities to predicted probabilities. Points are jittered horizontally to reduce overlap.

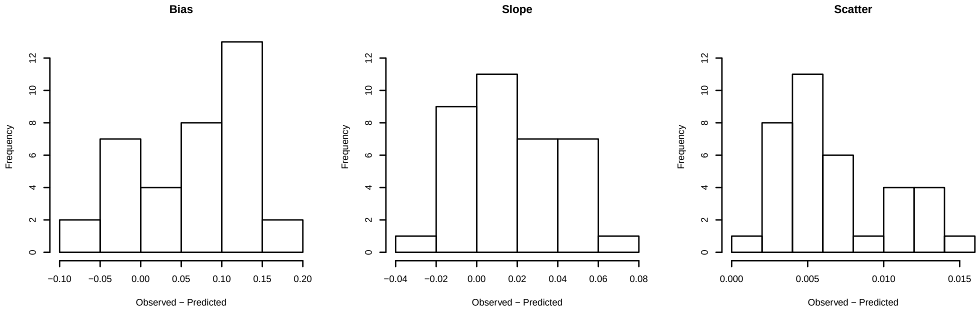

Fig. 2. Differences between observed probabilities and model predictions for each judge on the measures of bias, slope, and scatter.

observed statistic is greater than his/her model-predicted statistic. While there is judge variability, the graphs generally show that the model tends to yield smaller bias and scatter statistics. These trends are supported statistically: 95% confidence intervals for the mean difference in bias and scatter are $(.039, .084)$ and $(.005, .007)$, respectively. While these are positive results for the model, the slope statistics reflect a negative result: the observed slopes tend to be larger than the model-predicted slopes, with the 95% confidence interval for the mean difference being $(.008, .022)$. I address this negative result in more detail below.

For model predictions using the 25^{th} posterior percentiles, results are similar: the predictions yield reductions in both bias and scatter, but they do not yield increases in slope. Comparing the two types of predictions (25^{th} percentile predictions and mean predictions), both slope and scatter are virtually the same, with mean differences of .001 and .0003, respectively. The conservatism of the 25^{th} percentile predictions is reflected in the bias statistic. Mean bias for the 25^{th} percentiles is $-.036$ and mean bias for the means is .006, with a 95% confidence for interval for the mean difference being $(-.043, -.041)$. One may argue that the conservative predictions are too conservative, as the mean predictions display near-perfect bias statistics.

3.3 Discussion

The hierarchical logistic model transformed judges' subjective probabilities into predicted probabilities that were better calibrated (i.e., bias closer to zero) and contained less noise (i.e., reduced scatter). Importantly, the predictions were made on trials that were excluded from the model estimation. Thus, fitted models of this type can be used to predict probabilities of unknown outcomes, an attribute that is important for applications. More details appear in the General Discussion.

While the model improved bias and scatter, it did not improve slope. Stated differently, the model predictions were unable to better discriminate between correct and incorrect outcomes. This is partly due to the fact that the observed range of model predictions is smaller than the observed range of the f_{ij} (as shown in Figure 1). The result is also impacted by the fact that the $\widehat{p_{ij}}$ are increasing functions of the f_{ij}. This implies that the $\widehat{p_{ij}}$ follow the same ordering as the f_{ij}, which does not leave much room for improvement in slope. In the General Discussion, other statistical methods are considered that may improve slope.

4 General Discussion

Model-based corrections to subjective probabilities, such as those described in this paper, have the potential to be useful in many applied situations. Further, there exist many other statistical methods/models that can produce probabilistic predictions and that may yield improvements in slope. Both of these topics are considered below.

4.1 Applications

The model considered in this paper is applicable to situations where judges report many subjective probabilities in response to stimuli from the same domain, such as medical diagnoses and aircraft inspections. Focusing on the latter application, inspectors examine many areas of the aircraft and report the existence of defects. These reports are often made with considerable uncertainty, especially in nondestructive testing situations (e.g., [14, 15]). For example, to check for cracks in bolt holes, inspectors sometimes rely on eddy current technology. An electrical charge is sent through the material around the bolt hole, and inspectors rely on a digital monitor to diagnose cracks. This occurs across a large number of bolt holes on the aircraft.

If inspectors report probabilities of cracks in each bolt hole, the hierarchical logistic model can be used to improve the reported probabilities of individual inspectors. In such a scenario, inspectors may first complete test inspections where the existence of a crack is known. The model can then be fit to these inspections, and the fitted model used to improve reported probabilities for cases where the existence of a crack is unknown.

4.2 Other Statistical Methods

The primary disadvantage of the hierarchical logistic model is that it fails to yield improvements in slope. This is likely to be a problem with any statistical model whose predictions are a linear function of the f_{ij}, because the ordering among the predictions will be the same as the ordering among the f_{ij}. As a result, it may be useful to study models or algorithms that utilize nonlinear functions of f_{ij}. There are at least two classes of methods that one may consider: statistical learning algorithms (e.g., [16]) and psychological/psychometric models of subjective judgment (e.g., [11, 12, 17–19]).

The main focus of statistical learning algorithms, such as boosting, is prediction. The algorithms can make predictions that are nonlinear functions of the inputs, meaning that they may more easily yield improvements in the slope measure (as opposed to the logistic model). A possible problem with the use of these algorithms is lack of data: the algorithms are often suited to data containing thousands of observations and hundreds of predictor variables. The applications considered here may contain hundreds of observations and two predictor variables (subjective probability, judge who reported the probability). In such cases, it is unclear whether the algorithms will result in improvements over more traditional statistical models.

Psychological models may also be used to transform subjective probabilities. These models often posit psychological processes contributing to the construction of subjective probabilities. As a result, the models often treat subjective probability as a response variable instead of a predictor variable. This may make it difficult to use subjective probabilities to predict accuracy. On the other hand, the psychological models may be modified to obtain distributions of accuracy conditioned on subjective probability. If the models truly describe psychological processes contributing to subjective probability, then their accuracy predictions may be better than the more general models/algorithms described earlier. In any case, Bayesian model formulations and Markov chain Monte Carlo are likely to be useful tools for studying these psychological models.

References

1. Dawes, R.M.: Confidence in intellectual vs. confidence in perceptual judgments. In Lantermann, E.D., Feger, H., eds.: Similarity and choice: Papers in honor of Clyde Coombs. Bern: Han Huber (1980) 327–345
2. Lichtenstein, S., Fischhoff, B.: Do those who know more also know more about how much they know? The calibration of probability judgments. Organizational Behavior and Human Performance **20** (1977) 159–183
3. Wallsten, T.S., Budescu, D.V.: Encoding subjective probabilities: A psychological and psychometric review. Management Science **29** (1983) 152–173
4. Cooke, R.M.: Experts in uncertainty: Opinion and subjective probability in science. New York: Oxford University Press (1991)
5. O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., Rakow, T.: Uncertain judgements: Eliciting experts' probabilities. Hoboken: Wiley (2006)
6. Wallsten, T.S., Budescu, D.V., Erev, I., Diederich, A.: Evaluating and combining subjective probability estimates. Journal of Behavioral Decision Making **10** (1997) 243–268
7. Yates, J.F., Curley, S.P.: Conditional distribution analyses of probabilistic forecasts. Journal of Forecasting **4** (1985) 61–73
8. Keren, G.: On the ability of monitoring non-veridical perceptions and uncertain knowledge: Some calibration studies. Acta Psychologica **67** (1988) 95–119
9. Lichtenstein, S., Fischhoff, B., Phillips, L.D.: Calibration of probabilities: The state of the art to 1980. In Kahneman, D., Slovic, P., Tversky, A., eds.: Judgment under uncertainty: Heuristics and biases. Cambridge, England: Cambridge University Press (1982) 306–334
10. Price, P.C.: Effects of a relative-frequency elicitation question on likelihood judgment accuracy: The case of external correspondence. Organizational Behavior and Human Decision Processes **76** (1998) 277–297
11. Dougherty, M.R.P.: Integration of the ecological and error models of overconfidence using a multiple-trace memory model. Journal of Experimental Psychology: General **130** (2001) 579–599
12. Merkle, E.C., Van Zandt, T.: An application of the Poisson race model to confidence calibration. Journal of Experimental Psychology: General **135** (2006) 391–408
13. Thomas, A., O'Hara, B., Ligges, U., Sturtz, S.: Making BUGS open. R News **6** (2006) 12–17
14. Swets, J.A.: Assessment of NDT systems–Part I: The relationship of true and false detections. Materials Evaluation **41** (1983) 1294–1298
15. Swets, J.A.: Assessment of NDT systems–Part II: Indices of performance. Materials Evaluation **41** (1983) 1299–1303
16. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning. New York: Springer (2001)

17. Batchelder, W.H., Romney, A.K.: Test theory without an answer key. Psychometrika **53** (1988) 71–92
18. Erev, I., Wallsten, T.S., Budescu, D.V.: Simultaneous over- and underconfidence: The role of error in judgment processes. Psychological Review **101** (1994) 519–527
19. Ratcliff, R., Starns, J.: Modeling confidence and response time in recognition memory. Psychological Review **116** (2009) 59–83
20. Gelman, A., Hill, J.: Data analysis using regression and multilevel/hierarchical models. New York: Cambridge (2007)

Appendix: OpenBugs Code for the Hierarchical Logistic Model

The code below takes a 36×40 matrix of accuracy data (0=incorrect, 1=correct) and a 36×60 matrix of confidence data, where rows reflect judges and columns reflect items. It simultaneously fits the hierarchical logistic model to 40 trials of data from each judge and yields posterior predictions for the final 20 columns in `corr`. The data file (not shown) contains the accuracy data matrix (`corr`), the confidence data matrix (`conf`), and a 2×2 identity matrix (`Iden`). More details on Bayesian hierarchical logistic models is found in, e.g., [20].

```
model{
  for (i in 1:36){
    for (j in 1:40){
      corr[i,j] ~ dbern(p[i,j])

      logit(p[i,j]) <- b[i,1] + b[i,2]*conf[i,j]
    }
    # Hierarchical distribution on bs
    b[i,1:2] ~ dmnorm(mu.b[], invS[,])
  }

  # Posterior predictions for new confidence judgments
  for (i in 1:36){
    for (j in 41:60){
      newp[i,(j-40)] <- b[i,1] + b[i,2]*conf[i,j]
    }
  }

  # Priors
  mu.b[1] <- dnorm(0,1.0E-5)
  mu.b[2] <- dnorm(0,1.0E-5)
  invS[1:2,1:2] ~ dwish(Iden[,],2)
}
```