Bayesian Ranking with Limited Sample Sizes

Andrei Mitrofanov[1] Timur Magzhanov[1,2,3]

Anna Ivanova[3] Anton Polous[3,4]

Abstract

Ranking groups by average member quality becomes unreliable when group sizes differ. The sample mean has high variance for small groups and makes comparisons unreliable. We study this problem in the context of a laboratory forecasting experiment, where students from different university departments must be compared after completing the same task. We propose a Bayesian shrinkage estimator instead of the sample mean and prove that it strictly dominates the sample mean under Bayesian risk (unconditional MSE). A practical obstacle arises because the estimator requires a prior distribution, but no historical data exist for this type of experiment. We address this by generating a synthetic sample of students from the MSU Faculty of Economics via a large language model. How well language models reproduce human forecasting behaviour remains an open question.

1 Introduction

The problem addressed in this paper arises in a laboratory experiment studying systematic forecasting biases. Participants, students from different university departments, forecast values of AR(1) processes and receive scores depending on their accuracy. At the conclusion of the experiment, the departments must be compared: which one performed best? Computing the average score for each department and ranking them seems sufficient, but it is not.

The difficulty is that departments participate with different numbers of students. A department with three participants and an average score of 80 is placed above a department with forty participants and an average of 54. Yet three observations constitute an extremely unreliable basis: such a result could easily have arisen by chance even for a mediocre department. The sample mean is unbiased, but with a small number of observations its variance is large, and ranking by it systematically promotes small groups, not because they are better, but because their estimates happened to deviate upward by chance. There is no mechanism that pulls them back toward a typical level.

We address this problem using a Bayesian estimator, which automatically accounts for the unreliability of estimates based on small groups. A further difficulty arises: Bayesian estimation requires prior information about the distribution of participants' scores. Since this experiment has no historical precedent, such information is unavailable. We therefore propose a synthetic-data procedure based on a large language model, which transforms

[1]Lomonosov Moscow State University. Email: *andrei.mitrofanov77@gmail.com*

[2]Bocconi University. Email: *timur.magzhanov@phd.unibocconi.it*

[3]HSE University. Email: *ivanova@hse.ru*

[4]University of Edinburgh. Email: *A.Polous@sms.ed.ac.uk*

qualitative departmental descriptions into an initial empirical prior before real observations become available.

To formally compare the Bayesian estimator with the sample mean, a quality criterion is needed. The standard choice, mean squared error, fails here: its value depends on the parameter we are trying to estimate. For some values of the true mean the sample mean looks better, for others it looks worse. We instead use Bayesian risk: MSE averaged over the distribution of true departmental means across the population. This quantity does not depend on any unknown parameter and permits an unambiguous comparison. Using Bayesian risk as the evaluation criterion, we show that the Bayesian estimator strictly dominates the sample mean. We further establish its optimality for the ranking problem studied in this paper.

The paper is organised as follows. Section 2 introduces the problem through a binary case of ranking football penalty-takers. Section 3 extends the construction to continuous departmental scores and derives the shrinkage formula. Section 4 estimates model parameters by the method of moments in closed form. Sections 5 and 6 introduce the Bayesian risk criterion and prove the optimality of the shrinkage estimator. Section 7 describes the synthetic sample procedure. Section 8 concludes.

2 Ranking penalty-takers

We begin with a simple version of the problem: ranking football penalty-takers by their true conversion rate. Outcomes are binary (goal or miss), which allows the entire Bayesian construction to be written out explicitly.

2.1 Sampling model

Let $p_j \in (0,1)$ denote the true probability that player j converts a penalty. Each attempt is an independent Bernoulli trial,

$$y_{ji} \;\sim\; \text{Bernoulli}(p_j), \qquad i = 1, \ldots, n_j, \tag{1}$$

where $y_{ji} = 1$ if the kick is converted. The total number of goals

$$k_j \;=\; \sum_{i=1}^{n_j} y_{ji} \;\sim\; \text{Binomial}(n_j,\, p_j) \tag{2}$$

is a sufficient statistic for p_j, and the likelihood is

$$\mathcal{L}(p_j \mid k_j,\, n_j) \;=\; \binom{n_j}{k_j} p_j^{k_j} (1-p_j)^{n_j - k_j}. \tag{3}$$

The naive estimator

$$\hat{p}_j \;=\; \frac{k_j}{n_j} \tag{4}$$

is unbiased: since $k_j \sim \text{Binomial}(n_j, p_j)$ and $\mathbb{E}[k_j] = n_j p_j$, we obtain

$$\mathbb{E}[\hat{p}_j] \;=\; \mathbb{E}\left[\frac{k_j}{n_j}\right] \;=\; \frac{\mathbb{E}[k_j]}{n_j} \;=\; \frac{n_j p_j}{n_j} \;=\; p_j. \tag{5}$$

The variance of the estimator is computed analogously: since $\mathrm{Var}(k_j) = n_j p_j(1-p_j)$ for the binomial distribution,

$$\mathrm{Var}(\hat{p}_j) \;=\; \mathrm{Var}\left(\frac{k_j}{n_j}\right) \;=\; \frac{\mathrm{Var}(k_j)}{n_j^2} \;=\; \frac{n_j p_j(1-p_j)}{n_j^2} \;=\; \frac{p_j(1-p_j)}{n_j}. \tag{6}$$

Despite its unbiasedness, the estimator is unreliable for small n_j: the variance (6) is inversely proportional to the number of observations and remains high for small n_j regardless of the true value of p_j. In other words, unbiasedness only guarantees correctness of the estimator on average over many repetitions, not in any particular sample, which is precisely where it is needed in practice. Consider two players:

Player	**Kicks** n_j	**Goals** k_j	**Sample mean** k_j/n_j	**Naive rank**
A	3	3	1.000	1
B	30	24	0.800	2

By sample conversion rate, player A ranks first with a perfect record. However, this result rests on only three kicks, whereas player B has thirty. Comparing sample means based on such different numbers of observations is invalid: a perfect record from three attempts may be coincidence rather than evidence of genuine superiority. To eliminate this distortion, we turn to the Bayesian estimator, which automatically accounts for the volume of data behind each result.

2.2 Prior distribution: the Beta family

The parameter p_j is a probability and takes values on $(0,1)$, so the prior distribution must be concentrated on this interval. We choose the Beta family because its density has the same functional form as the Binomial likelihood. The Binomial likelihood is proportional to $p_j^{k_j}(1-p_j)^{n_j-k_j}$, and the $\mathrm{Beta}(\alpha,\beta)$ density is proportional to $p^{\alpha-1}(1-p)^{\beta-1}$: both are of the form $p^a(1-p)^b$. This shared structure means that multiplying prior and likelihood yields a posterior in the same family, so the Beta distribution is the unique conjugate prior for the Binomial on $(0,1)$. The $\mathrm{Beta}(\alpha,\beta)$ distribution admits a transparent interpretation as a virtual sample of $\alpha+\beta$ kicks, of which α were scored and β missed. The parameter α plays the role of the number of virtual goals, β gives the number of virtual misses, and their sum $\alpha+\beta$ determines the total "volume" of prior information in units of actual kicks.

Conjugacy guarantees a closed-form analytical posterior without numerical integration and makes Bayesian updating straightforward and explicitly computable. Bayesian updating upon observing k_j goals from n_j kicks reduces to simply adding the real and virtual counts: the posterior is again Beta with parameters $(\alpha+k_j,\ \beta+n_j-k_j)$.

The prior distribution

$$p_j \;\sim\; \mathrm{Beta}(\alpha,\ \beta) \tag{7}$$

has density

$$\pi(p_j) \;=\; \frac{p_j^{\alpha-1}(1-p_j)^{\beta-1}}{B(\alpha,\beta)}, \qquad p_j \in (0,1), \tag{8}$$

or, equivalently, up to a normalising constant:

$$\pi(p_j) \;\propto\; p_j^{\alpha-1}(1-p_j)^{\beta-1}, \tag{9}$$

where $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ is the normalising constant ensuring $\int_0^1 \pi(p_j)\, dp_j = 1$.

It is convenient to reparametrise as

$$\mu_0 \;=\; \frac{\alpha}{\alpha+\beta}, \qquad \kappa \;=\; \alpha+\beta, \tag{10}$$

so that $\alpha = \kappa\mu_0$ and $\beta = \kappa(1-\mu_0)$. Here μ_0 denotes the population-average conversion rate, and κ sets the size of the virtual sample: the larger κ, the more real data are required to shift the estimate away from μ_0.

2.3 Posterior distribution

The key tool for updating the estimate from data is Bayes' theorem. In general form: for a parameter θ with likelihood $\mathcal{L}(\theta) = p(\text{data} \mid \theta)$ and prior density $\pi(\theta)$, the posterior density is proportional to their product:

$$\pi(\theta \mid \text{data}) \;\propto\; \mathcal{L}(\theta) \cdot \pi(\theta). \tag{11}$$

Applying (11) to p_j with $p_j \sim \text{Beta}(\alpha, \beta)$ and likelihood (3):

$$\pi(p_j \mid k_j) \;\propto\; \mathcal{L}(p_j \mid k_j,\, n_j) \cdot \pi(p_j). \tag{12}$$

Substituting (3) and (8):

$$\pi(p_j \mid k_j) \;\propto\; p_j^{\alpha+k_j-1}(1-p_j)^{\beta+n_j-k_j-1}, \tag{13}$$

which is the kernel of $\text{Beta}(\alpha + k_j,\ \beta + n_j - k_j)$. The posterior mean equals

$$\hat{p}_j \;=\; \mathbb{E}[p_j \mid k_j] \;=\; \frac{\alpha + k_j}{\alpha + \beta + n_j} \;=\; \frac{\kappa\mu_0 + k_j}{\kappa + n_j}. \tag{14}$$

This formula implements an intuitive principle: when data are scarce (small n_j), we place more weight on the typical population level μ_0; when data are abundant, we favour the observed result k_j/n_j.

2.4 Numerical illustration

Example 1. Returning to the two players from Section 2.1, we apply formula (14). Set the prior parameters: $\mu_0 = 0.70$ and $\kappa = 10$, giving $\alpha = \kappa\mu_0 = 7$ and $\beta = \kappa(1-\mu_0) = 3$. This is equivalent to a virtual sample of ten kicks with a 70% conversion rate, supplementing the real data of each player. We compute the Bayesian estimates:

$$\hat{p}_A \;=\; \frac{\alpha + k_A}{\alpha + \beta + n_A} \;=\; \frac{7+3}{10+3} \;=\; \frac{10}{13} \;\approx\; 0.769,$$

$$\hat{p}_B \;=\; \frac{\alpha + k_B}{\alpha + \beta + n_B} \;=\; \frac{7+24}{10+30} \;=\; \frac{31}{40} \;=\; 0.775.$$

Player	n_j	k_j	k_j/n_j	**Naive rank**	$\hat{p}_j$	**Bayesian rank**
A	3	3	1.000	1	0.769	2
B	30	24	0.800	2	0.775	1

Bayesian ranking reverses the naive ordering. Player A's perfect record rests on only three kicks: the posterior mean pulls $\hat{p}_A$ from 1.000 toward $\mu_0 = 0.70$, reflecting the fact that three attempts provide too thin a basis for confident conclusions. Player B has thirty kicks with a stable conversion rate. The ten virtual observations from the prior barely influence the estimate, which stays close to the observed 0.800. The numerical difference between $\hat{p}_A$ and $\hat{p}_B$ is small, but the direction is correct.

In the numerical illustration, the parameters $\mu_0 = 0.70$ and $\kappa = 10$ were chosen as specific numbers for ease of exposition. In practice they are estimated from the same data: μ_0 is estimated as the sample mean of the observed conversion rates across all players, and κ is estimated from the spread of those conversion rates, corrected for sampling noise.

The penalty-taker example has given us a clean illustration of Bayesian logic: the posterior mean automatically accounts for the data volume and pulls unreliable estimates toward the center. We now pose the same problem in a more complex context: continuous scores and many groups simultaneously. This is where two new obstacles arise that were absent in the binary case.

3 Academic departments

In the penalty-taker example the task was to estimate the hidden parameter p_j for each player, based on a limited number of observations. The same problem arises when ranking academic departments, except that participant scores are continuous rather than binary. In this section we carry the Bayesian logic over to the department-ranking problem: we specify the statistical model, derive the shrinkage formula in closed form, and analyse its asymptotic behaviour. Substantive justification of the distributional assumptions is given in Section 7.

3.1 Model specification

Let $x_{j1}, \ldots, x_{jn_j}$ denote the independent scores of participants in department j, drawn from a distribution with true mean μ_j and within-group variance σ_j^2, which may differ across departments. The sample mean

$$\bar{x}_j \;=\; \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ji} \tag{15}$$

is an unbiased estimator of μ_j with variance

$$\operatorname{Var}(\bar{x}_j \mid \mu_j) \;=\; \frac{\sigma_j^2}{n_j}. \tag{16}$$

For small n_j this variance is large, and $\bar{x}_j$ may deviate substantially from μ_j in a given sample. Ranking by sample means ignores this difference in precision and places small groups at an unfair advantage.

Example 2. Consider three departments with $n_A = 3$, $n_B = 10$, and $n_C = 40$ participants and sample means $\bar{x}_A = 80$, $\bar{x}_B = 72$, $\bar{x}_C = 54$.

Department	**Participants** n_j	**Sample mean** $\bar{x}_j$	**Rank**
A	3	80	1
B	10	72	2
C	40	54	3

Department A ranks first, but its lead with three participants may be the result of sampling chance rather than genuine superiority. The Bayesian answer to this problem is the same as for penalty-takers: instead of the sample mean, use the posterior mean, which automatically accounts for the unreliability of estimates based on small groups.

To obtain the posterior mean in closed form, a conjugate pair of distributions is required. For continuous scores the natural choice is the normal-normal pair, and we adopt it as the working basis of the model. First, for sufficiently large n_j, by the Central Limit Theorem the sample mean is approximately normally distributed:

$$\bar{x}_j \mid \mu_j \overset{\text{approx.}}{\sim} \mathcal{N}\left(\mu_j,\ \frac{\sigma_j^2}{n_j}\right). \tag{17}$$

Second, the latent departmental means are treated as realisations of a normal distribution over the population:

$$\mu_j \ \sim\ \mathcal{N}(\mu_0,\ \tau^2). \tag{18}$$

Both assumptions are working assumptions; their substantive justification and the mechanism for obtaining the parameters μ_0 and τ^2 in the absence of historical data are provided in Section 7. With the model specified, we proceed to derive the posterior mean in closed form.

3.2 Shrinkage formula

The likelihood (17) and the prior (18) are both normal. We apply Bayes' theorem (11):

$$p(\mu_j \mid \bar{x}_j) \ \propto\ p(\bar{x}_j \mid \mu_j) \cdot p(\mu_j).$$

This is a normal-normal conjugate pair, playing the same role as the Beta-Binomial pair in Section 2.

The densities of both factors are:

$$p(\bar{x}_j \mid \mu_j) \ =\ \frac{1}{\sqrt{2\pi\,\sigma_j^2/n_j}}\,\exp\left\{-\frac{n_j(\bar{x}_j-\mu_j)^2}{2\sigma_j^2}\right\}, \tag{19}$$

$$p(\mu_j) \ =\ \frac{1}{\sqrt{2\pi\tau^2}}\,\exp\left\{-\frac{(\mu_j-\mu_0)^2}{2\tau^2}\right\}. \tag{20}$$

The normalising constants do not depend on μ_j and are absorbed by the $\propto$ sign. Therefore the only quantity of interest in the product is the sum of the exponents:

$$p(\mu_j \mid \bar{x}_j) \ \propto\ \exp\left\{-\frac{n_j(\bar{x}_j-\mu_j)^2}{2\sigma_j^2}-\frac{(\mu_j-\mu_0)^2}{2\tau^2}\right\}. \tag{21}$$

Expanding the squares in (21) and grouping by powers of μ_j, discarding terms that do not depend on μ_j:

$$\begin{aligned}
&-\frac{n_j}{2\sigma_j^2}(\bar{x}_j^2-2\bar{x}_j\mu_j+\mu_j^2)-\frac{1}{2\tau^2}(\mu_j^2-2\mu_0\mu_j+\mu_0^2)\\
&\quad=\ -\underbrace{\left(\frac{n_j}{2\sigma_j^2}+\frac{1}{2\tau^2}\right)}_{A}\mu_j^2+\underbrace{\left(\frac{n_j}{\sigma_j^2}\bar{x}_j+\frac{1}{\tau^2}\mu_0\right)}_{B}\mu_j+\text{const.}
\end{aligned}$$

Denote the coefficient of μ_j^2 by A and the coefficient of μ_j by B:

$$A \;=\; \frac{n_j}{2\sigma_j^2} + \frac{1}{2\tau^2}, \qquad B \;=\; \frac{n_j}{\sigma_j^2}\,\bar{x}_j + \frac{1}{\tau^2}\,\mu_0. \tag{22}$$

The exponent equals $-A\mu_j^2 + B\mu_j + \text{const}$. Completing the square in μ_j:

$$-A\mu_j^2 + B\mu_j \;=\; -A\left(\mu_j - \frac{B}{2A}\right)^2 + \frac{B^2}{4A}, \tag{23}$$

whence

$$p(\mu_j \mid \bar{x}_j) \;\propto\; \exp\left\{-A\left(\mu_j - \frac{B}{2A}\right)^2\right\}. \tag{24}$$

The kernel of a normal distribution $\mathcal{N}(\mu, \sigma^2)$ has the form $\exp\{-(x-\mu)^2/(2\sigma^2)\}$, i.e. $\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$. Comparing with (24): the coefficient in front of the parenthesis equals $A = 1/(2\sigma^2)$, so $\sigma^2 = 1/(2A)$, and the center of the parenthesis is $\mu = B/(2A)$. Consequently, the posterior distribution is normal:

$$\mu_j \mid \bar{x}_j \;\sim\; \mathcal{N}\big(\mu_j^{\text{post}},\, \text{Var}_j^{\text{post}}\big)\,, \qquad \mu_j^{\text{post}} = \frac{B}{2A}, \quad \text{Var}_j^{\text{post}} = \frac{1}{2A}. \tag{25}$$

Substituting A from (22):

$$\text{Var}_j^{\text{post}} \;=\; \frac{1}{2A} \;=\; \frac{1}{\dfrac{n_j}{\sigma_j^2} + \dfrac{1}{\tau^2}} \;=\; \frac{\sigma_j^2\tau^2}{n_j\tau^2 + \sigma_j^2}. \tag{26}$$

Substituting A and B from (22):

$$\mu_j^{\text{post}} \;=\; \frac{B}{2A} \;=\; \frac{\dfrac{n_j}{\sigma_j^2}\,\bar{x}_j + \dfrac{1}{\tau^2}\,\mu_0}{\dfrac{n_j}{\sigma_j^2} + \dfrac{1}{\tau^2}}. \tag{27}$$

Introduce $\lambda_j = \sigma_j^2/\tau^2$ as the ratio of within-group to between-group variance for department j. Then $1/\tau^2 = \lambda_j/\sigma_j^2$, and dividing numerator and denominator of (27) by $1/\sigma_j^2$:

$$\mu_j^{\text{post}} \;=\; \frac{n_j\,\bar{x}_j + \lambda_j\,\mu_0}{n_j + \lambda_j} \;=\; \underbrace{\frac{n_j}{n_j + \lambda_j}}_{\text{data weight}}\,\bar{x}_j \;+\; \underbrace{\frac{\lambda_j}{n_j + \lambda_j}}_{\text{prior weight}}\,\mu_0. \tag{28}$$

This is the *shrinkage formula*:

$$\boxed{\hat{\mu}_j \;=\; \frac{n_j}{n_j + \lambda_j}\,\bar{x}_j \;+\; \frac{\lambda_j}{n_j + \lambda_j}\,\mu_0.} \tag{29}$$

Denoting the data weight

$$w_j \;=\; \frac{n_j}{n_j + \lambda_j} \;\in\; (0, 1), \tag{30}$$

the shrinkage formula takes the compact form:

$$\hat{\mu}_j = w_j\,\bar{x}_j + (1 - w_j)\,\mu_0. \tag{31}$$

3.3 Asymptotic behaviour

The shrinkage formula (29) contains a single free parameter, the data weight $w_j \in (0,1)$. Analysis of limiting cases shows that this weight behaves exactly as a sensible estimator should: it tends to one when the departmental data are informative, and to zero when they are noisy or when departments are indistinguishable. Thus formula (29) is not an arbitrary convex combination but an adaptive rule that is consistent with intuition in all extreme cases.

Limiting case	**Condition**	w_j	$\hat{\mu}_j$
Many observations	$n_j \to \infty$	$\to 1$	$\to \bar{x}_j$
Low noise	$\sigma_j^2 \to 0$	$\to 1$	$\to \bar{x}_j$
High noise	$\sigma_j^2 \to \infty$	$\to 0$	$\to \mu_0$
Homogeneous groups	$\tau^2 \to 0$	$\to 0$	$\to \mu_0$
Heterogeneous groups	$\tau^2 \to \infty$	$\to 1$	$\to \bar{x}_j$

When $\hat{\mu}_j \to \bar{x}_j$, Bayesian ranking coincides with naive ranking: the departmental data are sufficiently informative, or the inter-departmental heterogeneity is too large for the prior to contribute anything. When $\hat{\mu}_j \to \mu_0$, all departments receive the same estimate: the department's own data are so noisy, or the departments so homogeneous, that the observed differences between them are pure noise. Shrinkage operates in the range between these extremes.

4 Parameter estimation: Method of moments

The shrinkage formula (29) is expressed in terms of three parameters: μ_0, σ_j^2, and τ^2, which are unknown in practice. They must be estimated from the data. We use the *method of moments*: it yields closed-form formulas without iterative computation and does not require full specification of the distributions. The approach whereby prior parameters are estimated from the same data is called *empirical Bayes*.

4.1 Method logic and estimation formulas

The idea of the method of moments is simple: model (17)–(18) uniquely links the three unknown parameters (μ_0, σ_j^2, τ^2) to three characteristics of the observed data. We equate the theoretical expressions to their sample analogues and obtain three equations from which the three unknowns are directly recovered.

The three theoretical moment conditions are:

$$\mathbb{E}[\bar{x}_j] \;=\; \mu_0, \tag{32}$$

$$\mathbb{E}\big[(x_{ji} - \bar{x}_j)^2\big] \;=\; \sigma_j^2 \cdot \frac{n_j - 1}{n_j}, \tag{33}$$

$$\mathrm{Var}(\bar{x}_j) \;=\; \tau^2 + \frac{\sigma_j^2}{n_j}. \tag{34}$$

We derive each from the model structure. The first follows from the law of total expectation:

$$\mathbb{E}[\bar{x}_j] \;=\; \mathbb{E}_{\mu_j}\big[\mathbb{E}[\bar{x}_j \mid \mu_j]\big] \;=\; \mathbb{E}_{\mu_j}[\mu_j] \;=\; \mu_0,$$

where the first step uses the unbiasedness of $\bar{x}_j$ at fixed μ_j, and the second uses $\mathbb{E}[\mu_j] = \mu_0$. The second condition relies on a standard result from estimation theory: at fixed μ_j, observations x_{ji} are independent with variance σ_j^2, and the expected sum of squared deviations from the sample mean equals $(n_j - 1)\sigma_j^2$, so

$$\mathbb{E}\big[(x_{ji} - \bar{x}_j)^2 \mid \mu_j\big] \;=\; \frac{n_j - 1}{n_j}\,\sigma_j^2;$$

averaging over μ_j does not change the right-hand side since σ_j^2 does not depend on μ_j. The third condition follows from the law of total variance:

$$\mathrm{Var}(\bar{x}_j) \;=\; \mathbb{E}_{\mu_j}[\mathrm{Var}(\bar{x}_j \mid \mu_j)] + \mathrm{Var}_{\mu_j}(\mathbb{E}[\bar{x}_j \mid \mu_j])\,,$$

where the first term corresponds to sampling noise σ_j^2/n_j and the second to the true between-group heterogeneity τ^2.

Substituting sample analogues into each of the three conditions yields the estimators in closed form. The within-group variance is estimated from condition (33):

$$\hat{\sigma}_j^2 \;=\; \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(x_{ji} - \bar{x}_j)^2. \tag{35}$$

The population center is estimated from (32):

$$\hat{\mu}_0 \;=\; \frac{1}{J}\sum_{j=1}^{J}\bar{x}_j. \tag{36}$$

The between-group variance is estimated from (34):

$$\hat{\tau}^2 \;=\; \max\left\{0,\; \frac{1}{J-1}\sum_{j=1}^{J}(\bar{x}_j - \hat{\mu}_0)^2 - \frac{1}{J}\sum_{j=1}^{J}\frac{\hat{\sigma}_j^2}{n_j}\right\}. \tag{37}$$

We truncate it at zero because variance is non-negative by definition, but the sample estimate of the difference of two quantities can be negative if the observed spread of sample means is smaller than the expected noise contribution. The method of moments is particularly well-suited to our problem. The three unknown parameters (μ_0, σ_j^2, τ^2) are exactly identified by the three moment conditions (32)–(34): each parameter governs exactly one observable characteristic of the data (μ_0 governs the level, σ_j^2 the within-department spread, and τ^2 the between-department spread). The system has a unique closed-form solution without any optimisation. The estimators (35)–(37) are consistent for any distribution F_j with finite variance.

4.2 Empirical shrinkage formula

Substituting the estimators (35)–(37) into the shrinkage formula (29) in place of the true parameters, introduce the empirical coefficient:

$$\hat{\lambda}_j \;=\; \frac{\hat{\sigma}_j^2}{\hat{\tau}^2}, \tag{38}$$

playing the same role as $\lambda_j = \sigma_j^2/\tau^2$ in the theoretical formula: the "virtual sample size" from the prior for department j. The empirical Bayes estimator is:

$$\boxed{\hat{\mu}_j^{\mathrm{EB}} \;=\; \frac{n_j}{n_j + \hat{\lambda}_j}\,\bar{x}_j \;+\; \frac{\hat{\lambda}_j}{n_j + \hat{\lambda}_j}\,\hat{\mu}_0.} \tag{39}$$

The structure is the same as in the theoretical formula (29): a weighted average of the department's sample result and the population center. The only difference is that all three parameters are replaced by their data-based estimates. Formula (39) is what is applied in practice.

5 Loss function and quality criterion

To compare the sample mean and the Bayesian estimator and select the better one, a precise quality criterion is needed. In this section we construct three interrelated tools for building that criterion.

We first fix the form of the *loss function*, a numerical measure of the penalty for an estimation error. We then define the *mean squared error* (MSE), i.e. the expected loss upon resampling at fixed μ_j, which characterises the typical magnitude of the estimation error. Finally, we introduce the *Bayesian risk*, i.e. the MSE averaged over the distribution of the unknown parameter μ_j. The Bayesian risk serves as the final criterion for comparing estimators in our problem, where the true departmental parameters are unknown.

5.1 Loss function

Let $\hat{a}$ denote a point estimator of μ_j, and let

$$e \;=\; \hat{a} - \mu_j \tag{40}$$

denote the estimation error. The loss function $L(e)$ describes the cost of an error of magnitude e. We impose four natural properties on it. First, *symmetry* ($L(-e) = L(e)$): overestimating and underestimating by the same amount are equally costly. Second, *convexity*: large errors are penalised more than proportionally, if the error doubles, the loss grows faster than twofold. Third, *minimum at zero*: $L(0) = 0$, i.e. there is no loss when there is no error. Finally, *smoothness*: $L(e)$ is twice differentiable and $L''(0) > 0$. The sign of $L''(0)$ determines the character of the function near zero: if $L''(0) < 0$, the function has a local maximum at zero, contradicting the minimum; if $L''(0) = 0$, the loss grows slower than e^2 and the function is insensitive to small errors. Only $L''(0) > 0$ guarantees convexity near zero.

Expand an arbitrary $L(e)$ from this class in a Taylor series about zero:

$$L(e) \;=\; L(0) + L'(0)\cdot e + \tfrac{1}{2}L''(0)\cdot e^2 + \tfrac{1}{6}L'''(0)\cdot e^3 + \dots \tag{41}$$

Our three conditions progressively simplify this expansion.

First, $L(0) = 0$, so the constant term vanishes.

Second, differentiating both sides of $L(-e) = L(e)$ with respect to e:

$$\frac{d}{de}L(-e) = \frac{d}{de}L(e) \quad\Rightarrow\quad L'(-e)\cdot(-1) = L'(e) \quad\Rightarrow\quad L'(-e) = -L'(e).$$

The first derivative is anti-symmetric. Setting $e = 0$:

$$L'(0) = -L'(0) \quad \Rightarrow \quad L'(0) = 0.$$

By the same logic all odd-order derivatives at zero vanish: $L'''(0) = 0$, $L^{(5)}(0) = 0$, and so on.

Third, since all odd derivatives at zero vanish, all odd-order terms in the Taylor expansion disappear:

$$L(e) \;=\; L(0) + 0 \cdot e + \tfrac{1}{2}L''(0) \cdot e^2 + 0 \cdot e^3 + \ldots \;=\; \tfrac{1}{2}L''(0) \cdot e^2 + o(e^2),$$

where $o(e^2)$ denotes higher-order terms (of fourth degree and above), negligible for small errors.

Without loss of generality we set $\frac{1}{2}L''(0) = 1$. This yields the canonical quadratic loss function:

$$L(e) \;=\; e^2. \tag{42}$$

Thus, quadratic loss is a universal second-order approximation to any smooth symmetric loss function near zero. The convexity of e^2 has direct meaning in a ranking problem: large errors cost more. Quadratic loss formalises precisely this aversion to gross mistakes. The loss function is now fixed. The next step is to construct the expected loss - the mean squared error.

5.2 Mean squared error (MSE)

Having adopted the quadratic loss function $L(e) = e^2$. Here $\hat{a}$ is as defined in Section 5: the specific number we report as an estimate of μ_j. Suppose μ_j is fixed and known. The randomness resides only in the sample: with each new sample of students from the same department, the quality of estimator $\hat{a}$ changes. To characterise the accuracy of $\hat{a}$ on average, rather than the luck of a particular realisation, we average the loss over all possible samples at fixed μ_j. This is the mean squared error (MSE):

$$\mathrm{MSE}(\hat{a} \mid \mu_j) \;=\; \mathbb{E}\big[e^2 \mid \mu_j\big] \;=\; \mathbb{E}\big[(\hat{a} - \mu_j)^2 \mid \mu_j\big]\,. \tag{43}$$

The MSE decomposes into the sum of two meaningful components. Let $a := \mathbb{E}[\hat{a} \mid \mu_j]$. At fixed μ_j this is a constant. Adding and subtracting a inside the square:

$$\begin{aligned}(\hat{a} - \mu_j)^2 \;&=\; \big[(\hat{a} - a) + (a - \mu_j)\big]^2 \\ &=\; (\hat{a} - a)^2 + 2(\hat{a} - a)(a - \mu_j) + (a - \mu_j)^2.\end{aligned}$$

Taking expectations of each term.

First term: $\mathbb{E}[(\hat{a} - a)^2 \mid \mu_j] = \mathrm{Var}(\hat{a} \mid \mu_j)$, by definition of variance.

Second term: $(a - \mu_j)$ is a constant at fixed μ_j and can be taken outside the expectation:

$$\mathbb{E}[2(\hat{a} - a)(a - \mu_j) \mid \mu_j] \;=\; 2(a - \mu_j) \cdot \mathbb{E}[\hat{a} - a \mid \mu_j] \;=\; 0,$$

since $\mathbb{E}[\hat{a} - a \mid \mu_j] = a - a = 0$.

Third term: $(a - \mu_j)^2$ is a constant at fixed μ_j, independent of the sample, since $a = \mathbb{E}[\hat{a} \mid \mu_j]$ is determined by the model and μ_j alone, not by a particular realisation:

$$\mathbb{E}[(a - \mu_j)^2 \mid \mu_j] \;=\; (a - \mu_j)^2 \;=\; \mathrm{Bias}(\hat{a} \mid \mu_j)^2.$$

In total:

$$\mathrm{MSE}(\hat{a} \mid \mu_j) \;=\; \mathrm{Var}(\hat{a} \mid \mu_j) + \mathrm{Bias}(\hat{a} \mid \mu_j)^2. \tag{44}$$

The MSE decomposes into two independent components: *variance* (how unstable the estimator is upon resampling) and the square of the *bias* (how systematically the estimator deviates from the truth). The second term vanishes for unbiased estimators, so variance and bias contribute to the MSE independently. An estimate can be improved by reducing variance, reducing bias, or both.

5.3 Bayesian risk

The main problem with MSE is that it depends on the unknown μ_j. For different values of μ_j one estimator may yield a smaller MSE while another yields a larger one. To compare two estimators by MSE, one needs to know μ_j, but that is precisely what we are trying to estimate.

The answer is the *Bayesian risk*: the double expectation of the squared error, first over the sample at fixed μ_j (this is the MSE), then over μ_j from distribution G:

$$R_B(\hat{a}) \;=\; \mathbb{E}_{\mu_j}\big[\,\mathbb{E}\big[(\hat{a}-\mu_j)^2 \mid \mu_j\big]\big] \;=\; \mathbb{E}_{\mu_j}[\mathrm{MSE}(\hat{a} \mid \mu_j)]\,. \tag{45}$$

By the law of total expectation the two nested expectations combine into a single unconditional one:

$$R_B(\hat{a}) \;=\; \mathbb{E}\big[(\hat{a}-\mu_j)^2\big]\,. \tag{46}$$

Using the decomposition (44), the Bayesian risk also decomposes into two components:

$$R_B(\hat{a}) \;=\; \mathbb{E}_{\mu_j}[\mathrm{Var}(\hat{a} \mid \mu_j)] + \mathbb{E}_{\mu_j}\big[\mathrm{Bias}(\hat{a} \mid \mu_j)^2\big]\,. \tag{47}$$

These are the same variance and bias as in (44), but averaged over distribution G. The Bayesian risk is a single number independent of the unknown μ_j, and is therefore the uniquely valid criterion for comparing estimators in our problem.

6 Superiority of Bayesian risk

The previous section introduced the Bayesian risk as a quality criterion. In this section we rigorously establish the superiority of the Bayesian shrinkage estimator $\hat{\mu}_j^B$ (29) over the sample mean $\hat{\mu}_j^M = \bar{x}_j$ under this criterion. We first analyse the bias of both estimators and then compare the conditional MSE and show that it provides no definitive answer: the result depends on the unknown μ_j. We then move to Bayesian risk and prove that $R_B(\hat{\mu}_j^B) < R_B(\hat{\mu}_j^M)$ for all n_j, σ_j^2, and τ^2. Finally, we establish the stronger result: under quadratic loss the posterior mean minimises Bayesian risk among all possible estimators.

6.1 Conditional and unconditional bias of estimators

The conditional bias of estimator $\hat{\mu}_j$ is the systematic deviation of its expected value from the true parameter at fixed μ_j:

$$\mathrm{Bias}(\hat{\mu}_j \mid \mu_j) \;:=\; \mathbb{E}[\hat{\mu}_j \mid \mu_j] - \mu_j.$$

An estimator is called conditionally unbiased if this expression equals zero for all values of μ_j.

For the sample mean $\hat{\mu}_j^M = \bar{x}_j$, unbiasedness is obvious: since $x_{ji} \mid \mu_j$ are independent with $\mathbb{E}[x_{ji} \mid \mu_j] = \mu_j$,

$$\mathbb{E}[\hat{\mu}_j^M \mid \mu_j] \;=\; \mathbb{E}\left[\frac{1}{n_j}\sum_{i=1}^{n_j} x_{ji} \;\middle|\; \mu_j\right] \;=\; \frac{1}{n_j}\sum_{i=1}^{n_j}\mathbb{E}[x_{ji} \mid \mu_j] \;=\; \mu_j,$$

i.e. $\mathrm{Bias}(\hat{\mu}_j^M \mid \mu_j) = 0$.

The Bayesian estimator $\hat{\mu}_j^B = w_j\bar{x}_j + (1-w_j)\mu_0$ is conditionally biased. Since w_j and μ_0 are constants at fixed μ_j and $\mathbb{E}[\bar{x}_j \mid \mu_j] = \mu_j$:

$$\mathbb{E}[\hat{\mu}_j^B \mid \mu_j] \;=\; w_j\mu_j + (1-w_j)\mu_0,$$

whence

$$\mathrm{Bias}(\hat{\mu}_j^B \mid \mu_j) \;=\; (1-w_j)(\mu_0 - \mu_j). \tag{48}$$

The estimator systematically pulls every department toward the population center μ_0. However, if this bias is averaged over distribution G from which μ_j is drawn, it vanishes:

$$\mathbb{E}\big[\hat{\mu}_j^B - \mu_j\big] \;=\; (1-w_j)\,\mathbb{E}[\mu_0 - \mu_j] \;=\; (1-w_j)\,(\mu_0 - \mathbb{E}[\mu_j]) \;=\; 0, \tag{49}$$

since by (18) $\mathbb{E}[\mu_j] = \mu_0$. Thus the Bayesian estimator is conditionally biased at fixed μ_j but unconditionally unbiased: errors for strong and weak departments cancel each other out. This bias is the price paid for variance reduction; how advantageous this trade-off is will be shown in the next section.

6.2 Variance and conditional MSE

The Bayesian estimator $\hat{\mu}_j^B = w_j\bar{x}_j + (1-w_j)\mu_0$ is a linear function of $\bar{x}_j$, with w_j and μ_0 constant at fixed μ_j. Therefore:

$$\mathrm{Var}(\hat{\mu}_j^B \mid \mu_j) \;=\; \mathrm{Var}(w_j\bar{x}_j + (1-w_j)\mu_0 \mid \mu_j) \;=\; w_j^2\,\mathrm{Var}(\bar{x}_j \mid \mu_j) \;=\; \frac{w_j^2\,\sigma_j^2}{n_j}.$$

Since $w_j < 1$, this is strictly less than the variance of the sample mean:

$$\mathrm{Var}(\hat{\mu}_j^B \mid \mu_j) \;=\; \frac{w_j^2\,\sigma_j^2}{n_j} \;<\; \frac{\sigma_j^2}{n_j} \;=\; \mathrm{Var}(\hat{\mu}_j^M \mid \mu_j). \tag{50}$$

The Bayesian estimator always has strictly smaller variance than the sample mean. However, smaller variance does not by itself guarantee smaller MSE: from decomposition (44) we see that the conditional bias (48) contributes an additional term to the error. Let us compare the conditional MSEs directly:

$$\mathrm{MSE}(\hat{\mu}_j^M \mid \mu_j) \;=\; \frac{\sigma_j^2}{n_j}, \tag{51}$$

$$\mathrm{MSE}(\hat{\mu}_j^B \mid \mu_j) \;=\; \frac{w_j^2\,\sigma_j^2}{n_j} + (1-w_j)^2\,(\mu_j - \mu_0)^2 \;=\; \frac{n_j\,\sigma_j^2 + \lambda_j^2(\mu_j - \mu_0)^2}{(n_j + \lambda_j)^2}. \tag{52}$$

Comparing the two expressions directly. The difference

$$\mathrm{MSE}(\hat{\mu}_j^M \mid \mu_j) - \mathrm{MSE}(\hat{\mu}_j^B \mid \mu_j) \;=\; \frac{\sigma_j^2}{n_j} - \frac{n_j\,\sigma_j^2 + \lambda_j^2(\mu_j - \mu_0)^2}{(n_j + \lambda_j)^2}$$

can be either positive or negative, depending on how far μ_j is from μ_0. When μ_j is close to the population center μ_0, the second term is small and the Bayesian estimator wins in conditional MSE. When μ_j is far from μ_0, the quadratic penalty for bias (48) outweighs the variance gain and the Bayesian estimator loses.

Thus, the conditional MSE provides no definitive answer to the question of which estimator is better: the answer depends on the true μ_j, which is precisely the unknown. This is the fundamental limitation of a conditional criterion, removed by moving to Bayesian risk in the next section.

6.3 Strict dominance under Bayesian risk

We move from conditional MSE to Bayesian risk, i.e. unconditional MSE averaged over the distribution $\mu_j \sim G$. We compute it for each of the two estimators.

Bayesian risk of the sample mean. Since $\hat{\mu}_j^M$ is conditionally unbiased (Section 6.1), its conditional MSE equals simply its variance: $\mathrm{MSE}(\hat{\mu}_j^M \mid \mu_j) = \sigma_j^2/n_j$ for all μ_j. Averaging over μ_j:

$$R_B(\hat{\mu}_j^M) \;=\; \mathbb{E}_{\mu_j}\left[\frac{\sigma_j^2}{n_j}\right] \;=\; \frac{\sigma_j^2}{n_j}. \tag{53}$$

Bayesian risk of the Bayesian estimator. From the MSE decomposition (44) and the conditional MSE formula (52):

$$R_B(\hat{\mu}_j^B) \;=\; \frac{w_j^2\,\sigma_j^2}{n_j} + (1-w_j)^2\,\mathbb{E}[(\mu_j-\mu_0)^2] \;=\; \frac{w_j^2\,\sigma_j^2}{n_j} + (1-w_j)^2\,\tau^2,$$

where $\mathbb{E}[(\mu_j-\mu_0)^2] = \tau^2$ by model (18). Substituting $w_j = n_j/(n_j+\lambda_j)$ and $1-w_j = \lambda_j/(n_j+\lambda_j)$ with $\lambda_j = \sigma_j^2/\tau^2$:

$$\begin{aligned}\frac{w_j^2\,\sigma_j^2}{n_j} + (1-w_j)^2\,\tau^2 \;&=\; \frac{n_j^2}{(n_j+\lambda_j)^2}\cdot\frac{\sigma_j^2}{n_j} + \frac{\lambda_j^2}{(n_j+\lambda_j)^2}\cdot\tau^2 \\ &=\; \frac{n_j\sigma_j^2+\lambda_j^2\tau^2}{(n_j+\lambda_j)^2} \;=\; \frac{n_j\sigma_j^2+\sigma_j^4/\tau^2}{(n_j+\lambda_j)^2}.\end{aligned}$$

Multiplying numerator and denominator by τ^2 and simplifying:

$$R_B(\hat{\mu}_j^B) \;=\; \frac{\sigma_j^2\tau^2}{n_j\tau^2+\sigma_j^2}. \tag{54}$$

Both Bayesian risks are now explicit. We compare them.

Proposition 1 (Strict Dominance)**.** For any $n_j \geq 1$ and any $\sigma_j^2, \tau^2 > 0$:

$$R_B(\hat{\mu}_j^B) \;<\; R_B(\hat{\mu}_j^M).$$

Proof. Substituting (54) and (53), we need to show:

$$\frac{\sigma_j^2\tau^2}{n_j\tau^2+\sigma_j^2} \;<\; \frac{\sigma_j^2}{n_j}.$$

Multiplying both sides by $n_j(n_j\tau^2+\sigma_j^2) > 0$:

$$n_j\sigma_j^2\tau^2 \;<\; \sigma_j^2(n_j\tau^2+\sigma_j^2) \;=\; n_j\sigma_j^2\tau^2+\sigma_j^4.$$

Cancelling $n_j\sigma_j^2\tau^2$ from both sides gives $0 < \sigma_j^4$, which holds for any $\sigma_j^2 > 0$. □

Proposition 1 is the central result of this section: the Bayesian risk of the Bayesian estimator is strictly less than the Bayesian risk of the sample mean for any $n_j \geq 1$ and $\sigma_j^2, \tau^2 > 0$. This means that switching from the sample mean to the shrinkage estimator is guaranteed to reduce the unconditional squared error, regardless of the value of the unknown μ_j and without any asymptotic assumptions.

Remark 1. Proposition 1 is proved for the theoretical estimator with known parameters μ_0, σ_j^2, τ^2. For the empirical version (39) with plug-in estimates, dominance is restored for a sufficiently large number of groups J: by the consistency of the method of moments, $\hat{\mu}_0 \to \mu_0$ and $\hat{\tau}^2 \to \tau^2$ as $J \to \infty$, whence the empirical estimator converges to the theoretical one.

6.4 Optimality of Bayesian estimator

Proposition 1 showed that the Bayesian estimator is strictly better than the sample mean under Bayesian risk. This is, however, only a partial result: we compared two specific estimators. A natural question arises: can a still better estimator be found? The answer is negative: the posterior mean minimises Bayesian risk among all possible estimators.

Proposition 2 (Optimality of the Posterior Mean)**.** Under quadratic loss (42), for any estimator $\hat{\mu}_j$:

$$R_B(\hat{\mu}_j^B) \;\leq\; R_B(\hat{\mu}_j),$$

with equality if and only if $\hat{\mu}_j = \hat{\mu}_j^B$ almost surely. In other words, $\hat{\mu}_j^B = \mathbb{E}[\mu_j \mid \bar{x}_j]$ is the Bayes estimator, i.e. the solution to

$$\operatorname*{argmin}_{\hat{\mu}_j} \; R_B(\hat{\mu}_j).$$

Proof. Fix an arbitrary realisation $\bar{x}_j$ and denote $m := \mathbb{E}[\mu_j \mid \bar{x}_j]$, the posterior mean. Consider an arbitrary estimator $\hat{\mu}_j$ and write its expected loss at the given $\bar{x}_j$:

$$f(\hat{\mu}_j) \;:=\; \mathbb{E}\big[(\hat{\mu}_j - \mu_j)^2 \mid \bar{x}_j\big]\,.$$

Adding and subtracting m: $(\hat{\mu}_j - \mu_j) = (\hat{\mu}_j - m) + (m - \mu_j)$, whence

$$f(\hat{\mu}_j) \;=\; (\hat{\mu}_j - m)^2 + 2(\hat{\mu}_j - m)\underbrace{\mathbb{E}[m - \mu_j \mid \bar{x}_j]}_{=\,0} + \underbrace{\mathbb{E}[(m - \mu_j)^2 \mid \bar{x}_j]}_{=\,\mathrm{Var}(\mu_j \mid \bar{x}_j)} \;=\; (\hat{\mu}_j - m)^2 + \mathrm{Var}(\mu_j \mid \bar{x}_j).$$

The second term does not depend on $\hat{\mu}_j$; the first, $(\hat{\mu}_j - m)^2 \geq 0$, vanishes if and only if $\hat{\mu}_j = m$. But $m = \mathbb{E}[\mu_j \mid \bar{x}_j]$ coincides exactly with formula (29), so $m = \hat{\mu}_j^B$. Consequently, $f(\hat{\mu}_j^B) \leq f(\hat{\mu}_j)$ for every realisation $\bar{x}_j$. Averaging over $\bar{x}_j$:

$$R_B(\hat{\mu}_j^B) \;=\; \mathbb{E}_{\bar{x}_j}\big[f(\hat{\mu}_j^B)\big] \;\leq\; \mathbb{E}_{\bar{x}_j}[f(\hat{\mu}_j)] \;=\; R_B(\hat{\mu}_j)$$

for any estimator $\hat{\mu}_j$, as required. □

This is the main result of the section. The Bayesian estimator, the posterior mean $\hat{\mu}_j^B = \mathbb{E}[\mu_j \mid \bar{x}_j]$, is the optimal estimator under the Bayesian risk criterion in an absolute sense: no other estimator, however ingenious, can achieve a smaller Bayesian risk. Section 6.3 established that the Bayesian estimator dominates the sample mean; the present section establishes that it dominates all others.

7 Synthetic sample via LLM

In Sections 3–4, normality of the likelihood and the prior was adopted as a working assumption. In our experiment, participants forecast values of AR(1) processes via a web application. This type of task has never previously been conducted in such a format: there exists neither a historical database nor an academic precedent from which prior parameters could be estimated directly. This section explains why this creates specific difficulties and describes the mechanism for resolving them.

Bayesian estimator in principle requires no specific distributional form and operates in full generality. However, to obtain the posterior distribution in closed form, a conjugate pair is required. For continuous scores, the natural choice is the normal-normal model. But justifying normality of both components in real data at the outset is not possible.

At the start of the experiment, the number of participants per department is small, and the Central Limit Theorem has not yet taken effect: the sample mean is far from normally distributed. Simultaneously, the true departmental means are unknown, and their collective distribution can be empirically verified only with a sufficient number of departments. For small J, such verification is impossible.

However, both problems disappear with sufficient data volume. The solution is to generate this volume artificially before the actual ranking begins, that is, to create a synthetic sample that ensures normality in closed form and permits reliable estimation of the prior parameters.

7.1 Synthetic sample via LLM as a solution

The idea is to generate, before the actual ranking begins, a synthetic sample large enough that both problems described above cease to be problems. The generation tool is the language model GPT-5.1. The language model simulates the behaviour of a specific participant in the experiment, reproducing the actual mechanism by which results are formed. A synthetic participant's score is computed from their forecasts in the same way as for a real participant.

Each synthetic participant is described by a profile vector: university, department, degree level, year of study, and academic performance level. The profile is passed to the language model as context and specifies the type of person being simulated. Since the initial experimental population consists primarily of students from the MSU Faculty of Economics, the synthetic sample targets this population exclusively. Profile parameters are varied systematically: degree levels (bachelor years 1–4, master, doctoral) and academic performance (above-average and below-average GPA) are represented in approximately equal proportions, ensuring realistic within-department variation. In total, $n^{\text{synth}} = 100$ synthetic participants are generated.

Each query specifies the round index, the set of participant profiles, and the series history observed by each participant up to that point. The model is instructed to produce forecasts at horizons T+1, T+2, T+4, T+5, T+7, and T+8, returning a single point forecast for each horizon. The forecasting behaviour is specified to reflect typical patterns observed in human forecasting tasks: predictions are anchored to the most recent observations, near horizons partially continue the visible local trend, and distant horizons are smoothed toward the historical mean of the series. Individual variation is introduced through the profile, but its influence is specified as secondary to the series history, so that the synthetic scores reflect the structure of the time series rather than systematic dif-

ferences across academic backgrounds. In particular, no systematic relationship between academic performance and forecast accuracy is imposed.

The synthetic data are used exclusively for estimating the prior parameters μ_0 and τ^2 by the method of moments and for justifying the normality of the model. Bayesian estimators of departments are built from the real participants' forecasts. The parameter $n^{\text{synth}} = 100$ comfortably exceeds the standard CLT applicability threshold of 30: with one hundred observations, normality of the likelihood is ensured by construction, and the sample variance $\hat{\sigma}_j^2$ is sufficiently stable. The true mean of each department μ_j is shaped by the combined influence of many independent factors: the curriculum, the quality of teaching, the student composition. When such factors are numerous and independent, their aggregate effect tends toward a normal distribution.

8 Conclusion

In this paper, we study the problem of ranking groups when the number of observations differs across them. The sample mean is not a reliable criterion in this setting, and we propose a Bayesian shrinkage estimator in its place. We first consider a binary setting where the goal is to rank football penalty-takers by conversion rate. This case admits a fully explicit Bayesian solution and illustrates the core idea: the posterior mean pulls each group's estimate toward the population center, with the degree of shrinkage determined by the available data volume. We then extend the construction to continuous scores and derive the shrinkage formula in closed form for university departments. The three unknown parameters of the model are estimated from the data by the method of moments.

To evaluate the estimator, we introduce Bayesian risk as the comparison criterion. Under Bayesian risk we prove two results: the shrinkage estimator strictly dominates the sample mean for any group size and variance, and it is optimal among all possible estimators.

A separate practical difficulty arises at the start of the experiment, when real data are scarce and the prior parameters cannot be estimated from observations. We propose resolving this by generating a synthetic sample via a large language model. Whether language models can adequately reproduce human forecasting behaviour in this setting is not fully understood and remains an open question, which we regard as the main limitation of the current approach and a direction for future research.