Methods for aggregating probability judgments

Table of Contents

Introduction

Below we outline several methods for mathematically aggregating subjective probability Judgments from individuals of a group. Individuals in this project evaluate the replicability of findings (‘claims’) from the social and behavioural sciences. Specifically, we ask them to estimate the “probability that direct replications ofthis study would find a statistically significant effect in the same direction as the originalclaim”.

We elicit this with a modified-Delphi approach for eliciting judgements from groups, called the IDEA protocol (Hemming et al. 2018). For each research claim assessed, participants follow this process:

1) Read the claim and scan the paper it came from (individuals)
2) Provide an anonymous estimate of replicability, including the (i) lower bound, (ii) upper bound and (iii) best estimate of the probability that the claim would successfully replicate, together with their justifications for their estimates (as individuals)
3) Receive feedback about how their individual estimates differ from others’ and from the average group response (as groups)
4) Discuss differences in opinion on the claim with the rest of their group, and ‘consider the opposite’ (reasons why a claim *may* or *may not* successfully replicate)
5) Provide a second anonymous estimate of replicability that incorporates insights gained through discussion (as individuals).

Notation

N = total number of individuals, indexed by $i = 1{:}N$
N_g = total number of individuals in group g ($g = 1{:}5$ in the SIPS experimental data), indexed by i_g
C = total number of claims, indexed by $c = 1{:}C$
In formulae where we single out c, we will have claims indexed by *j (where $j = 1{:}C$).*

Each claim has outcome o_c, which is 1 if a successful replication takes place and 0 otherwise.

For each claim, c, each individual, i, provides assessments that the claim in question would successfully replicate using the IDEA protocol, which results in estimates for 3 probabilities:

L_c^i = lower bound of individual i for claim c
U_c^i = upper bound of individual i for claim c
B_c^i = best estimate of individual i for claim c

$0 \leq L_c^i \leq\!< B_c^i \leq U_c^i \leq 1$

Note that the IDEA protocol results in both a Round 1 and Round 2 set of probabilities for each claim, but that here we will assume that the final Round 2 responses (after discussion) are being referred to unless otherwise specified.

For each claim, we need to generate a combined *aggregate* probability of replication $\hat{p}_c$(i.e. *Confidence Score*).

Confidence scores

We define $\hat{p}_c$as the aggregate probability of replication to be calculated (i.e. Confidence Score) and $\hat{p}\,^i_c\,(method\ X)$ as the aggregate probability calculations for aggregation method X. For example, the first simple average "confidence score" for claim c and group g will be:

$$\hat{p}\,^g_c(method\ ID)\ =\ \frac{1}{N_g}\sum_{i_g=1}^{N_g} B^i_c$$

As all the "confidence scores" will be per claim (for one large group only) the g superscripts will be simply dropped, hence, for each aggregation method and each claim we will report a $\hat{p}_c(method\ ID)$.

As several of our methods incorporate within-individual variation across claims, we will need to continually update our CSs for those methods, and finalise at the end of the 3000 claims (some people will assess multiple claims throughout the program). We don't anticipate that this will have a large effect on the CSs, and this has been cleared with T&E.

Weights

Given that many of the aggregation methods involve weighted linear combinations, we can define some standard notation to enhance clarity.

We let unnormalized weights be denoted by w_method (with superscripts according to the individual and subscripts according to the claim) and the normalised versions be denoted by $\tilde{w}\_method$. All weights need to be normalised (i.e. to sum to 1), but as the process is the same for all of them, we will set the formulae for the un-normalized weights.

All differential weighted combinations will take the form

$$\hat{p}_c(method\ ID) = \sum_{i=1}^{N} \tilde{w}\_method^i_c\ B^i_c$$

We note that while most weights will be calculated on a per question per individual basis (i.e. the same individual has different weights depending on the claim), three will be calculated only on a per individual basis (i.e. QuizWAgg, VarIndIntWAgg, GranWAgg), and not vary across claims. In these cases, the weighted combination could be simplified to

$$\hat{p}_c(method\ ID) = \sum_{i=1}^{N} \tilde{w}\_method^i\, B_c^i$$

Aggregation Methods

Method IDs correspond to abbreviations in the accompanying summary document.

ArMean: Arithmetic mean of the best estimates

The simplest aggregation of estimates is the unweighted linear average (i.e. simply takes the unweighted $\hat{\mathrm{p}}_c$ average of the best estimates of each individual).

As defined above, the aggregate estimate for claim c is therefore:

$$\hat{\mathrm{p}}_c(\mathrm{ArMean}) = \frac{1}{N}\sum_{i=1}^{\mathrm{N}} B_c^i$$

where N is the number of individuals being aggregated.

Median: Median of the best estimates

Another simple approach is to take the median of the individuals' best estimates.

$$\hat{\mathrm{p}}_c(\mathrm{Median}\ ) = \ Median\{B_c^i\}_{i=1,..,N}$$

LOArMean: Arithmetic mean of the log odds transformed best estimates

The mean of the log odds transformed individual best estimates has outperformed benchmarks in previous analyses.

$$LogOdds_c^i = \frac{1}{N}\sum_{i=1}^{\mathrm{N}} log\left(\frac{B_c^i}{1-\ B_c^i}\right)$$

The mean log odds estimate is then transformed back to probabilities gives a final group estimate.

$$\hat{\mathrm{p}}_c(\mathrm{LOArMean}\,) \;=\; \left(\frac{e^{LogOdds_c^i}}{1+\; e^{LogOdds_c^i}}\right)$$

BetaArMean: A beta-transformed arithmetic mean

This method takes the average of best estimates and pushes it through a beta distribution (which effectively extremizes the mean estimate).

The justification for equal parameters ($\alpha = \beta$) are outlined in Satopää et al (2014) and the references therein (note that the method outlined in that paper is called a beta-transformed linear opinion pool). We used two datasets to find the optimal value of alpha. For the ACE IDEA dataset, the optimal value is 3, whereas for the SIPS data set, the optimal value is 7. These values are calculated based on the full datasets. In the full datasets each claim had more than 5 assessments (ACE IDEA has 6-48 assessments per claim and SIPS known-outcome has 25 per claim). When the same analysis is performed for random subsets of 5 assessments per claim, which we expect to have for most of the claims assessed by repliCATS, the optimal values for the 2 datasets remain mostly unchanged (3 for ACE and 6 for SIPS - both 6 and 7 give very similar average Brier scores, i.e. 0.1105 vs 0.1104). This suggests that the optimal value depends on the dataset rather than number of assessments per claim. Assuming this is true, we should use $\alpha = \beta = 6$.

$$\hat{p}_c(BetaArMean) = H_{\alpha,\beta}\left(\frac{1}{N}\sum_{i=1}^{\mathrm{N}} B_c^i\right)$$

where $H_{\alpha,\beta}$ is the cumulative distribution function of the Beta distribution with parameters α, β.

IntWAgg: Weighted by interval width

The interval width (i.e. precision) of the bounds provided by individuals may be an indicator of knowledge, and hence accuracy. There are many different ways to use interval width to weight the best estimates, with one possible approach being to weight according to the interval width across individuals for that claim, defined as follows:

$$w\_Interval_c^i = \frac{1}{(U_c^i \,-\, L_c^i)}$$

$$\hat{p}_c(IntWAgg\,) = \;\sum_{i=1}^{\mathrm{N}} \widetilde{w}\_Interval_c^i\, B_c^i$$

IndIntWAgg:

Given that people seem to vary in how wide they are willing to set their widest intervals it can make sense to rescale the interval width across all claims for that individual, which results in a rescaled interval width weight ($w\_nIndivInterval_c^i$), for individual *i* for claim c, relative to the widest interval provided by that individual across all claims C:

$$w\_nIndivInterval_c^i = \frac{1}{\frac{(U_c^i - L_c^i)}{\max_{j=1,..,C}\{(U_j^i - L_j^i)\}}}$$

where $\{(U_j^i \; - \; L_j^i)\}$ are individual *i*'s judgements for all claims. Then

$$\hat{p}_c(IndIntWAgg) = \sum_{i=1}^{\mathrm{N}} \widetilde{w}\_nIndivInterval_c^i \, B_c^i$$

VarIndIntWAgg: Weighted by variation in individuals' interval widths

A related issue is that people differ in how much they vary in their interval widths. A higher variance may indicate a higher responsiveness to the evidential situation. We define:

$w\_varIndivInterval^i = \; var\{(U_c^i \; - \; L_c^i)\}$ calculated across all claims $c = 1, \ldots, C$, for individual *i*. Then

$$\hat{p}_c(VarIndIntWAgg) = \sum_{i=1}^{\mathrm{N}} \widetilde{w}\_varIndivInterval_c^i \qquad B_c^i$$

AsymWAgg: Weighted by asymmetry of intervals

Just as the width of an interval may be an indicator of knowledge or informativeness, the asymmetry of an interval relative to the corresponding best estimate may likewise be an indicator. One simple way to define asymmetry is:

If $B_c^i \geq (U_c^i - L_c^i)/2$:

$$w\_asym_c^i = 1 \; - \; 2(U_c^i - B_c^i)/(U_c^i - L_c^i)$$

else:

$$w\_asym_c^i = 1 \; - \; 2(B_c^i - L_c^i)/(U_c^i - L_c^i)$$

$$\hat{p}_c(AsymAgg) = \sum_{i=1}^{\mathrm{N}} \widetilde{w}\_asym_c^i B_c^i$$

IndIntAsymWAgg: Weighted by individuals' interval widths and asymmetry

This method involves a weighted aggregation that combines the weights calculated in the AsymWAgg and IndIntWAgg methods.

$$w\_nIndIntW\_asym_c^i = \widetilde{w}\_nIndivInterval_c^i * \widetilde{w}\_asym_c^i$$

$$\hat{p}_c(IndIntWAsymAgg) = \sum_{i=1}^{\mathrm{N}} \widetilde{w}\_nIndInt\_asym_c^i B_c^i$$

KitchSinkWAgg: Weighted by everything but the kitchensink

This is a rather ad-hoc method that combines factors from the methods above.

$$w\_kitchSink_c^i = \widetilde{w}\_asym_c^i * \widetilde{w}\_nIndivInterval_c^i * \widetilde{w}\_varIndivInterval^i$$

$$\hat{p}_c(KitchSinkWAgg\ ) = \sum_{i=1}^{\mathrm{N}} \widetilde{w}\_kitchSink_c^i B_c^i$$

OutWAgg: Downweighting outliers

NB: This is not one of our favoured methods, as it did not perform well in preliminary analyses on a similar dataset. Also, outliers may be the individuals who know something that the others don't know. Downweighting them may not be wise, however, it may be more justifiable after discussion and feedback. As we have already formulated the method, we will include it in the full list of candidate methods for now.

This method down-weights outliers by using the differences from the central tendency of their best estimates:

$$d_c^i = \left(median\{B_c^i\}_{i=1,..,N} - B_c^i\right)^2$$

$$w\_outlier_c^i B_c^i = 1 - \frac{d_c^i}{\max_{j=1,..,N}\{d_c^j\}}$$

$$\hat{p}_c(OutWAgg) = \sum_{i=1}^{N} \tilde{w}\_outlier_c^i B_c^i$$

DistLimitWAgg: Weighted by the distance of the best estimate from the closest certainty limit

Preliminary analysis of the SIPS data reveals a positive correlation of 0.33 between Brier score and distance of the best estimate from nearest certainty limit (0 or 1) at the claim level. This indicates that giving greater weight to best estimates that are closer to certainty limits may be beneficial.

$$w\_distLimit_c^i = max\{B_c^i, \quad 1 - B_c^i\}$$

$$\hat{p}_c(DistLimitWAgg\ ) = \sum_{i=1}^{N} \tilde{w}\_distLimit_c^i B_c^i$$

ShiftWAgg: Weighted by judgments that shifted the most after discussion

Previous analyses on the ACE IDEA dataset indicate that when participants change their round 2 judgement, they become more accurate (Hanea et al 2017). Therefore, weighting individuals' best estimates by the change in best estimates from Round 1 to Round 2 (after discussion) on a given claim may be beneficial. Greater shifts will be given greater weight, such that

$$w\_shift_c^i = \left|B1_c^i - B_c^i\right|$$

$$\hat{p}_c(ShiftWAgg) = \sum_{i=1}^{N} \tilde{w}\_shift_c^i\ B_c^i$$

where $B1_c^i$ is the Round 1 estimate prior to discussion and B_c^i is the individual's Round 2 estimate after discussion and revision.

GranWAgg: Weighted by the granularity of best estimates

Probability scales can be broken down into segments, with the level of segmentation reflecting *granularity* (i.e. 40-50, versus 40-44, 45-49, and so on) (Yaniv & Foster, 1995). More skilled forecasters might be expected to have a finer grained appreciation of uncertainty and thus make more granular forecasts. Accordingly, it may be sensible to give greater weight to participants who more frequently use "granular" numbers like 63 or 67 instead of rounded ones like 65 or 70.

This is supported by the findings of Mellers et al. (2015) who found that Super forecasters (i.e very high performing forecasters) made more granular forecasts than other forecasters, being more likely to make forecasts divisible by 1% and only 1% (e.g., 17%, 28%, and 83%, and excluding all multiples of 5% and 10%). See also Friedman et al (2018).

Weighting by granularity could be done within claims or based on individual differences in forecasters across multiple claims. Based on the approach and finding of Mellers et al. (2015) we have chosen to focus on implementing the latter approach, where the proportion of granular forecasts made by a forecaster is used to generate a per-individual score, which is then used to weight that individuals responses.

Specifically, individuals' received a score of 1 for each claim that their best estimate was specified at a more granular level than 5% (i.e. not a multiple of 5%), and a zero otherwise. The scores per claims are summed to form a weight per individual, such that

$$w\_gran^{i} = \frac{1}{C}\sum_{j=1}^{\mathrm{C}}\left\lceil\frac{B_j^i}{0.05} - \left\lfloor\frac{B_j^i}{0.05}\right\rfloor\right\rceil$$

$$\hat{p}_c(GranWAgg\ ) = \sum_{i=1}^{\mathrm{N}}\widetilde{w}\_gran_c^iB_c^i$$

where $\lfloor x \rfloor$ and $\lceil x \rceil$ are the mathematical floor and ceiling operators respectively.

ReasonWAgg: Weighted by the breadth of reasoning provided to support the individuals' estimate

When individuals provide multiple unique reasons in support of their judgment, this may indicate a breadth of thinking, understanding and knowledge about the claim and its context, and may also reflect engagement[1]. Giving greater weight to best estimates that are accompanied by higher reasoning scores may be beneficial.

$$w\_reason_c^i = r_{c,}^i$$

$$\hat{p}_c(ReasonWAgg) = \sum_{i=1}^{\mathrm{N}}\widetilde{w}\_reason_c^iB_c^i$$

[1] Invoking different reasons across multiple claim assessments may be a further marker for this. We may develop this method further to incorporates two components of reasoning: 1) The number of unique reasons given by an individual compared to others in the group for that claim, and 2) the diversity of reasons given by an individual over multiple claims (i.e. individuals are assigned more weight on a given claim for providing reasons that they have not provided previously).

Where we let $r_{c,}^{i}$ be the individual's reasoning score, calculated based on the number of unique reasons provided by thatindividual in support of their estimate for claim c.

Qualitative statements made by individuals as they evaluate claims / studies will be coded by the repliCATS Reasoning team, according to a coding manual. Reasoning scores will be based on 25 categories derived from this manual. Individuals will receive one point for each of these reasoning categories they draw on over the course of their evaluation (i.e. Round 1 and Round 2 statements).

These categories are:

1. Plausibility of claim
2. Reliability of effect
3. Availability of data and/or materials
4. Inconsistencies in the reported methods / analysis
5. Detail / clarity / transparency of the reported methods / analysis
6. Population or subject characteristics
7. Blinding
8. Confounding variables
9. Effect size
10. Effect type (interaction or main effect)
11. p-value
12. Sample size / power of study (e.g. high/low)
13. Power analysis
14. Appropriateness of statistical analysis (including models)
15. Questionable Research Practices (QRPs) implied
16. Questionable Research Practices (QRPs) specified
17. Constructs / operationalisation / instruments / measurement
18. Domain expertise / familiarity with existing literature
19. Private knowledge
20. Date of publication
21. Discipline reputation
22. Journal reputation
23. Author / Institutional reputation
24. Design of study / experimental design
25. References to the group discussion / revision statements *(NB: unlike the other categories, these are not discrete reasons, but are elements of reasoning)*

As these 25 categories are a subset of a larger codebook, they will be iteratively developed to ensure that each code is used reliably by multiple analysts (i.e., coders), as measured by Krippendorf's alpha. For the purposes of the aggregation method, only those categories that meet a minimum cut-off value of α = 0.667 (point estimate) will be included in the reasoning weight analysis[2]. This is intentionally generous, to allow as many codes as possible

[2] Krippendorff (2019, p.357) tentatively provides some guidelines for establishing sufficient reliability: "Do not accept data with reliabilities whose confidence interval reaches below the smallest acceptable reliability α_{min}, for example, of 0.800, but no less than 0.667." The minimum that

be considered and not pre-emptively excluded from consideration. Poor reliability codes canbe collapsed with related codes (where applicable) and have their reliability re-calculated.Final alpha calculations for a code will be based on an analysis of at least 250 units (responses) by 2-3 coders.

NB: a proportion of the claims will be hand-coded in NVIVO 12. Once our human coders reach 'saturation' (no new coding categories are created), the program will be able to code the other claims automatically. The majority of claims will be coded automatically by the program, but instances that are hand-coded will be flagged as needed. The coding manual and other materials will be publicly available at the end of the project.

QuizWAgg: Weighted by performance on the quiz

As part of this project, individuals are asked to take a quiz before commencing the main task of evaluating research claims. The quiz is encouraged, but not compulsory. Performance on the quiz may demonstrate understanding and knowledge relevant to assessing claims, and choosing to take the quiz at all may also reflect engagement. Giving greater weight to individuals with higher quiz scores, and to those who commence the quiz, may be beneficial.

$$w\_quiz^i = \boldsymbol{Q}\,\mathbf{v}$$

$$\hat{p}_c(QuizWAgg) = \sum_{i=1}^{\mathrm{N}} \widetilde{w}\_gran_c^i B_c^i$$

where w_quiz^i is the weight of the score of individual i on the pre-workshop quiz, as defined below.

The quiz contains $n_{quiz} = 22$ questions. Individuals provide answers for each question, resulting in a $N \times n_{quiz}$ matrix $\boldsymbol{Q}$**,** where each element q_{ij} is 1 if the individual i answered question j correctly, and 0 otherwise. For each question answered correctly, the individual receives points, with the number of points received for a correct answer each of the 22 questions specified in the points vector

$$\mathbf{v} = \begin{cases} v_k = 0.5\,, & \text{for } k \in \{1{:}\,10{,}16{,}17\} \\ v_k = 1\,, & \text{for } k \in \{11{:}\,15, 18{:}\,22\} \end{cases}$$

we nominate is based on the point estimate (not the lower CI bound as Krippendorff suggested), as we anticipate that many of our codes (the 25 categories) will be either quite rare or quite commonly used, which require larger sample sizes than a code with a 50/50 present/absent split in order to achieve equivalently narrow confidence intervals. The rarity of some codes/reasons presents valuable information, but can complicate the reliability calculations. So, some balance needs to be found between reliability as measured by this quantitative index, and the ability to actually make use of the data from the qualitative analysis.

This results in a quiz score that ranges from 0 to 16, with higher scores equating to better performance.

Missing data: Individuals are assigned zero points for any questions missed on the quiz. Individuals who did not take the quiz at all will receive zero weight (and non-zero weight for those who responded to at least one item in the quiz). If only one person assessing a given claim took the quiz, the *QuizWAgg* Confidence Score for that claim will be based solely on their judgment. If quiz data is missing for a claim entirely (i.e. no group members took the quiz), we will replace it with the log odds transformed mean of best estimates provided by the group members (i.e. in the *QuizWAgg* column of our CS output by claim. Instances of this will be flagged and provided as needed. The quiz and scoring materials will be made publicly available at the end of the project.

DistribArMean: Arithmetic mean of the non-parametric distributions

This method assumes that the lower bound of the individual per claim corresponds to the 5% percentile q_5^i, the best estimate corresponds to the median q_{50}^i, and the upper bound corresponds to the 95% percentile, q_{95}^i. With these 3 percentiles, we can build the minimally informative non-parametric distribution that basically spreads the mass uniformly between the 3 percentiles.

$$F_i(x) = \begin{cases} 0, & \text{for } x < 0 \\ \frac{0.05}{q_5^i} \cdot x, & \text{for } 0 \le x < q_5^i \\ \frac{0.45}{q_{50}^i - q_5^i} \cdot \left(x - q_5^i\right) + 0.05, & \text{for} q_5^i \le x < q_{50}^i \\ \frac{0.45}{q_{95}^i - q_{50}^i} \cdot \left(x - q_{50}^i\right) + 0.5, & \text{for} q_{50}^i \le x < q_{95}^i \\ \frac{0.05}{1 - q_{95}^i} \cdot \left(x - q_{95}^i\right) + 0.95, & \text{for} q_{95}^i \le x < 1 \\ 1, & \text{for } x \ge 1. \end{cases}$$

Then it averages all such fitted distributions of participants (for claim c)

$$AvDistribution = \frac{1}{N}\sum_{i=1}^{N} F_i(x)$$

And the aggregation is the median of the average distribution

$$\hat{p}_c(LPArMean) = AvDistribution^{-1}(0.5)$$

BayTriVar: Bayesian Triple-Variability Method

This model assumes three kinds of variability around best estimates: 1. generic claim variability; 2. generic participant variability; 3. the claim*person specific uncertainty, operationalised by bounds. The model takes the log odds transformed individual estimates as input (data), uses a normal likelihood function and derives a posterior distribution for the probability of replication.

The Bayesian method requires specification of a likelihood function of the data (in this case, lower L_c^i, best B_c^i and upper estimate U_c^i). Best estimate B_c^i is logit transformed as follows:

$$logit\left(B_c^i\right) = \log\left(\frac{B_c^i}{\left(1 - B_c^i\right)}\right)$$

The ***likelihood*** function for $logit\left(B_c^i\right)$ is:

$$logit\left(B_c^i\right) \sim N(\mu_c, \sigma_{i,c}^2)$$

with *N*(,) denoting the Normal distribution with mean and variance respectively, $\mu_c$denoting the mean estimated probability of replication for claim c, and $\sigma_{i,c}^2$denoting the variance of the estimated probability of replication for claim $c$and individual i. Parameter $\sigma_{i,c}$ is defined as:

$$\sigma_{i,c} = \left(U_c^i - L_c^i + 0.01\right) \times \sqrt{{\sigma_i}^2 + {\sigma_c}^2}$$

with σ_i denoting the standard deviation of estimated probabilities of replication for individual i and σ_c denoting the standard deviation of the estimated probability of replication for claim c. To complete the specification of the Bayesian model, priors need to be given for μ_c, σ_i, and σ_c:

Priors

$$\mu_c \sim N(0, 3^2)$$
$$\sigma_i \sim U(0,10)$$
$$\sigma_c \sim U(0,10)$$

with *U*(,) denoting the Uniform distribution with lower and upper bound, respectively. The quantity of interest is the median of the posterior distribution of the estimated probability of replication for each claim c.

$$\text{Posterior} \propto Likelihood \times Prior$$

An MCMC algorithm is used to generate samples from this posterior distribution, after the parameters were back transformed to original values.

$$\hat{\mathrm{p}}_c(\mathrm{BayTriVar}) = Posterior^{-1}(0.5)$$

Method of Parameter Estimation

The model will be fitted with the R package *R2jags* (Su & Majima 2019) or *greta* (Golding 2019).

Model Checking Tests

Model checking will include posterior predictive checks on the Claims Assessment data. Cross-validation will be undertaken on the Claims Assessment data for which there are known outcomes.

BayPRIORsAgg: Prior derived from predictive models, updated with best estimates

This *BayPRIORsAgg* method uses Bayesian updating to update a prior probability of replication estimated from a predictive model with an aggregate of the experts' best estimates for any given claim.

Predictive Model generating Prior Estimates

We developed a Multilevel Logistic regression model to produce our prior estimates for each claim. We used a database of replications conducted as part of recent large scale replication projects (e.g., the Reproducibility Project: Psychology; Open Science Collaboration, 2015) to fit a multilevel logistic regression model that predicts the probability of replication using attributes of the original study:

$$\begin{aligned} Logit(replication\,odds) &= \beta_0 + \beta_1 n\_authors + \beta_2 interaction\ + \beta_3 ln(sample\_size) \\ &+\ \beta_4 n\_statcheck\_errors\ +\ u_{article}\ + u_{project}\ +\ e_{id} \end{aligned}$$

Included as predictors in the model are
1. The number of authors included in the original article (*n_authors*)
2. Whether the effect of interest is an interaction or not (*interaction*)
3. The natural log of the sample size in the original article (*ln(sample_size)*)
4. The number of statcheck errors detected in the original article (*n_statcheck_errors*)

In the model we also include random effects for the original article ($u_{article}$), as there are multiple replications of effects from the same original article, and a random effect for the replication project that each of the original replication projects was included in ($u_{project}$).

For each claim, these elements are extracted from the CoS TA1 CVD, except for *n_statcheck_errors*. Regular expressions are used to match and extract *n_authors* and *interaction*. For *sample_size* we use a combination of the R package words_to_numbers and regular expressions to extract the sample size from the CVD. To obtain *n_statcheck_errors*we pull the PDFs or XML files for each article from the OSF and run them through statcheck (Nuijten, Hartgerink, van Assen, Epskamp, &Wicherts, 2015).

For each claim c in the CVD we use the regression model to generate predictions about the probability of successful replication $\widehat{p_c}$, with the standard error of prediction representing uncertainty about that estimate, σ_c .

Bayesian Updating Model

We use the Bayesian Triple-Variability Method, *BayTriVar,* to perform Bayesian updating on our priors. The model is summarised in the Kruschke style diagram below. As *per BayTriVar*, we first logit transform each experts' best estimate B_c^i, where:

$$logit(B_c^i) = log\left(\frac{B_c^i}{(1 - B_c^i)}\right)$$

The likelihood function for the logit transformed best estimates is:

$$logit(B_c^i) \sim N\,(\mu_c, \sigma_{i,c}^2)$$

where the logit transformed best estimates are normally distributed, having a mean of $\mu_c$and variance of $\sigma_{i,c}^2$. $\mu_c$denotes the mean estimated probability of replication for claim c, and $\sigma_{i,c}^2$denotes the variance of the estimated probability of replication for claim c by individual i. The standard deviation, $\sigma_{i,c}$, is defined as:

$$\sigma_{i,c}{=}(U_c^i - L_c^i\ + 0.01)\ \times\ \sqrt{\sigma_c^2 + \sigma_i^2}$$

where $\sigma_i$represents the standard deviation of the estimated probabilities of replication for individual i, and $\sigma_c$represents the standard deviation of the estimated probability of replication for claim c. Both these standard deviations are on the logit scale.

Priors

Priors for the parameters μ_c, $\sigma_i$and $\sigma_c$are given by:

$$\mu_c \sim\ N(\widehat{p_c}, \sigma_c^2)$$
$$\sigma_i \sim\ U(0{,}10)$$
$$\sigma_c \sim\ U(0{,}10)$$

where the mean and standard deviation of the distribution of $\mu_c$are derived from the predictive model, and claim-level standard deviation, σ_c, and standard deviation of the estimated probability of replication for individual i , σ_i, are specified as flat priors with uniform distributions.

Computing the Posterior Distribution

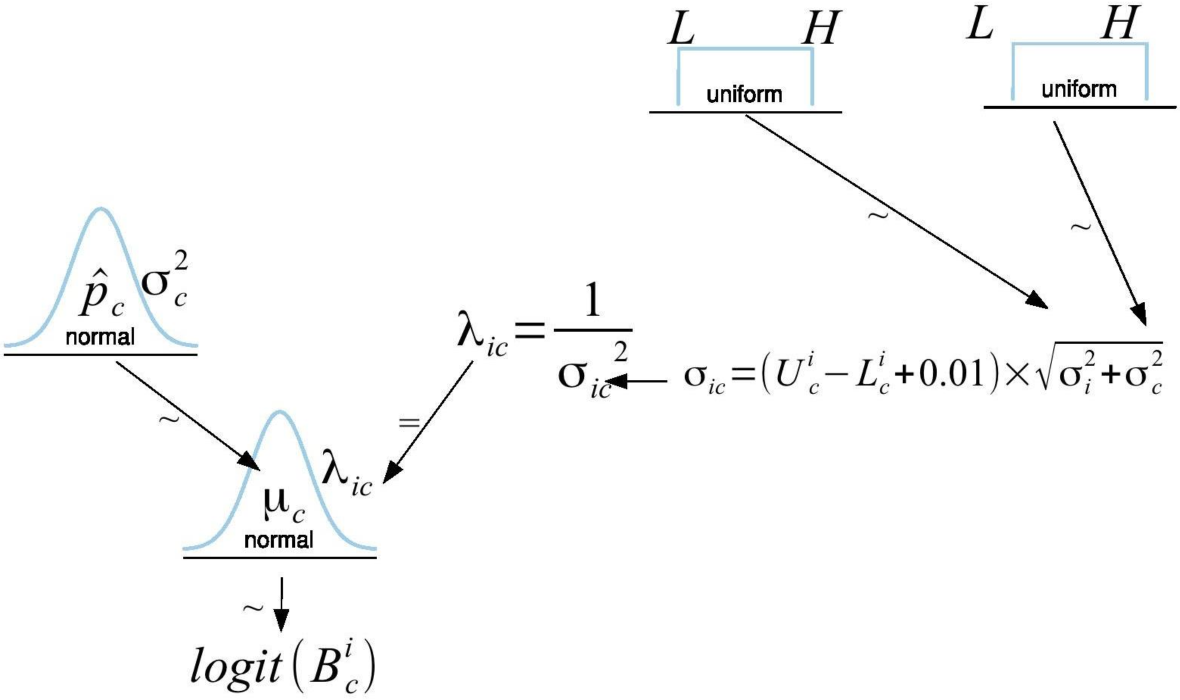
We are interested in the mean of the posterior distribution of the estimated probability of replication for each claim c, which is computed as:

$$\text{Posterior} \propto\ Likelihood\ \times\ Prior$$

An MCMC algorithm is used to generate samples from this posterior distribution, after the parameters were back transformed to original values.

$$\hat{p}_c(BayPRIORsAgg)\ =\ Posterior^{-1}(0.5)$$

Please note that this method may be adjusted. We will keep T&E regularly informed of changes.



L
H
uniform
L
H
uniform
~
~
$\hat{p}_c \; \sigma_c^2$
normal
$\lambda_{ic} = \frac{1}{\sigma_{ic}^2}$
$\sigma_{ic} = (U_c^i - L_c^i + 0.01) \times \sqrt{\sigma_i^2 + \sigma_c^2}$
~
=
λ_{ic}
μ_c
normal
~
$logit(B_c^i)$

Method of Parameter Estimation

Predictive model: This model will be fitted to the PRIORS dataset that contains actual replication outcomes and correlates of replication success coded by Team Reproducibility. We will use a generalised linear mixed model fitting package in R to implement parameter estimation. The structure of this model was determined using a combination of a priori decision-making based on the nesting structure of the data, as well as exploratory data analysis, and is detailed in Singleton Thorn et al. (in prep.).

Bayesian model: The model will be fitted with the R package *R2jags* (Su & Majima 2019) or *greta* (Golding, 2019). Priors for each claim will be parsed from the Predictive model.

Model Checking & Evaluation

Predictive model: Model assumptions will be checked for violation, and the model will undergo calibration and validation testing against the PRIORS dataset containing known replication outcomes.

Bayesian model: Bayesian model checking will include posterior predictive checks on the Claims Assessment data. Cross-validation will be undertaken on the Claims Assessment data for which there are known outcomes.

REFERENCES

Clemen, R.T., and Robert L. Winkler, R.L. (1999) Combining Probability Distributions From Experts in Risk Analysis, Risk Analysis, 19(2), 187-203

Cooke, R.M. (1991) *Experts in uncertainty: Opinion and subjective probability in science*. Oxford University Press, New York.

Cooke, R.M., Marti, D. and Mazzuchi, T. (in prep). Expert Judgment and the Inevitability of Validation. In preparation for John Evans Festschrift.

Friedman J., Baker J., Mellers B., Tetlock P. and Zeckhauser R. (2018) The Value of Precision in Probability Assessment: Evidence from a Large-Scale Geopolitical Forecasting Tournament. *International Studies Quarterly***62**, 410–422.

Golding, N. (2019). greta: Simple and Scalable Statistical Modelling in R. R package version 0.3.1. https://CRAN.R-project.org/package=greta

Hanea A. M., McBride M., Burgman M. A., Wintle B. C., Fidler F., Flander L., Twardy C. R., Manning B. and Mascaro S. (2017) Investigate Discuss Estimate Aggregate for structured expert judgement. *International Journal of Forecasting***33**, 267-279.

Hemming, V., Burgman, M.A., Hanea, A.M., McBride, M.F. & Wintle, B.C. (2018) A practical guide to structured expert elicitation using the IDEA protocol. *Methods in Ecology and Evolution,* 9, 169-181.

Krippendorff, K. (2019) Content Analysis: An Introduction to Its Methodology. 4th Edition. SAGE Publications: Los Angeles.

Lyon A., Wintle B. C. and Burgman M. (2015) Collective wisdom: Methods of confidence interval aggregation. Journal of Business Research 68, 1759-1767.

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S.E., Ungar, L., Bishop, M.M., Horowitz, M., Merkle, E. &Tetlock, P. (2015) The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of experimental psychology: applied,***21,** 1-14.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., … Tetlock, P. (2015). Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions. *Perspectives on Psychological Science*, *10*(3), 267–281. https://doi.org/10.1177/1745691615577794

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., &Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). Behavior Research Methods, 1-22. doi:10.3758/s13428-015-0664-2

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251). doi: 10.1126/science.aac4716

R Core Team (2019) R: A Language and Environment for Statistical Computing. Vienna, Austria. https://www.R-project.org

Satopää, V. A., Jensen, S. T., Pemantle, R., and Ungar, L. H. (2017) “Partial Information Framework: Aggregating Estimates from Diverse Information Sources.” The Electronic Journal of Statistics 11: 3781-3814.

Satopää V. A., Baron J., Foster D. P., Mellers B. A., Tetlock P. E. and Ungar L. H. (2014) Combining multiple probability predictions using a simple logit model. International Journal of Forecasting 30, 344-356.

Singleton Thorn, F., Gould, E., Fraser, H., Vesk, P. (in prep) Identifying Predictors of Replication.

Su, Y. and Yajima, M. (2015) R2jags: Using R to Run 'JAGS'. https://CRAN.R-project.org/package=R2jags

Yaniv, I. (1997). Weighting and trimming: Heuristics for aggregating judgments under uncertainty. Organizational Behavior and Human Decision Processes, 69(3), 237–249.

Yaniv, I., & Foster, D. P. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. Journal of Experimental Psychology: General, 124(4), 424-432. http://dx.doi.org/10.1037/0096-3445.124.4.424