AGGREGATION OF EXPERTS' JUDGMENTS

In order to aggregate experts' judgments, some methods –linear pooling, Delphi method and random effects meta-analysis- are commonly used. We are going to focus on linear pooling.

In this method, experts' weights are aggregated using linear combinations. The probability distribution of an unknown parameter p(θ) is calculated as shown in below:

$$\sum_i w(e) * pe(\theta)$$

Where w(e) is the expert e's weight and $pe(\theta)$ is the expert e's opinion. Experts can be weighted with different methods. Seed (calibration) questions are commonly used to determine weights.

Various methods are developed in order to determine the weights of experts in the literature. Excalibur is one of the approaches that are presented as software which assesses and combines experts' opinions based on Classical Model. It will be explained through an example in this paper.

Classical Model-Cooke's Method

Using a set of seed (calibration) questions, a group of experts are ranked according to the quality of the answers they provide to these seed questions. Seed questions should be determined considering the aim of study. To give an example, in a study which aims to assess alternative treatments for prostate cancer, the question "Consider the entire population of patients with metastatic prostate cancer starting treatment with a luteinizing hormone-releasing hormone analogue (LHRHa). What proportion of these patients experience clinically significant complications as a result of testosterone 'flare'? " can be used as a seed question.

Weights are calculated by the multiplication of calibration and information scores. It is then possible to aggregate experts' opinions with linear pooling applying the individual weights.

It should be noted that calibration score represents the statistical likelihood that results correspond with expert's assessments. On the other hand, information score represents the degree to which uncertainty distribution of expert is concentrated.

An expert is expected to divide her opinion range into four inter-quantile intervals, p_1=0.05, p_2=0.45, p_3=0.45 and p_4=0.05, for all seed questions. She determines three different values in order to obtain these four intervals, in which the probability that realization is less than or equal to the first value is 0.05, the probability that realization is less than or equal to the second value is 0.5 and the probability that realization is less than or equal to the third value is 0.95.

We will show how to calculate calibration and information scores step by step in an example. Assume that there are 3 experts and you determined 5 seed questions. Answers of experts are provided in Tables 1, 2 and 3 and the realizations are stated in Table 4.

Table 1: Values of expert 1 for seed questions

Seed question\Probability	0.05	0.5	0.95
1	40	45	60
2	75	100	125
3	70	95	110
4	120	130	140
5	130	150	180

Table 2: Values of expert 2 for seed questions

Seed question\Probability	0.05	0.5	0.95
1	30	50	70
2	100	110	120
3	90	100	110
4	100	120	140

5	110	120	130

Table 3: Values of expert 3 for seed questions

Seed question\Probability	0.05	0.5	0.95
1	50	60	70
2	60	70	80
3	70	90	110
4	110	120	130
5	130	150	170

Table 4: Realizations for all seed questions

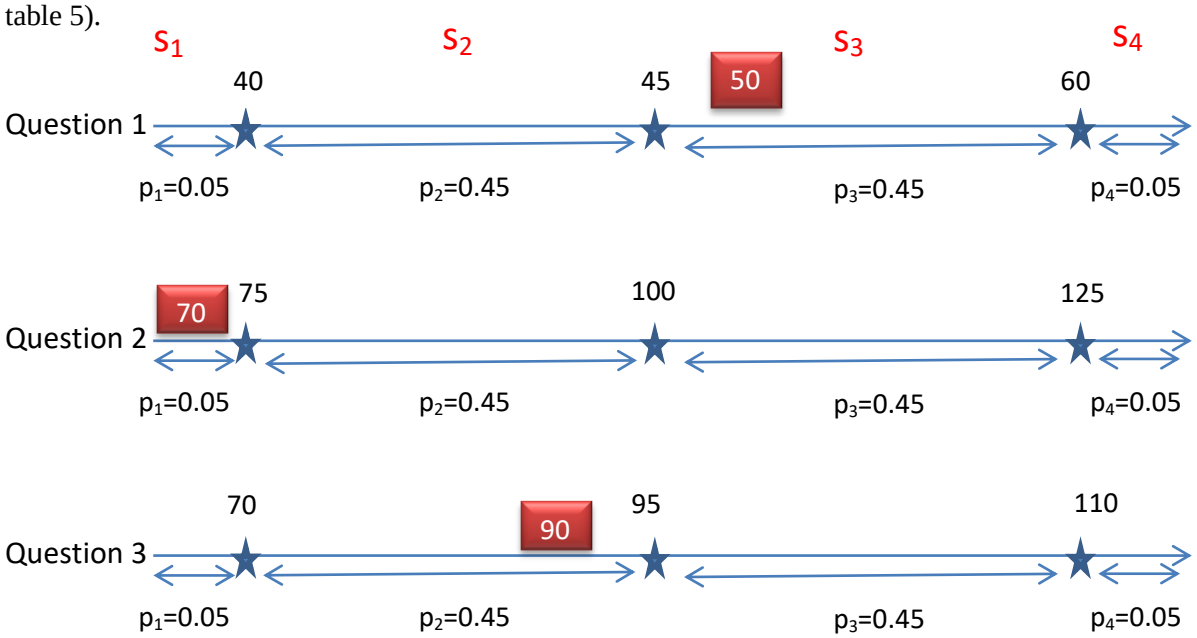
Seed question	Realization
1	50
2	70
3	90
4	100
5	120

Calibration Score

If N such quantities (N seed questions) are assessed by the experts, each expert can be seen as a statistical hypothesis, namely that each of the N realization values falls in one of the four inter-quantile intervals with probability vector p={0.05, 0.45, 0.45, 0.05}.

The sample distribution over the expert's inter-quantile intervals can then be formed by summing the number of realizations which fall in each interval, divided by the total number N. Note that the sample distribution depends on expert (e).

That is, in order to calculate calibration score for an expert, first which intervals that realizations fall into are determined for all seed questions (see figure 1). Then, for each expert, the sample distribution over four intervals can be found by dividing the number of realizations which fall in each interval to N (see table 5).



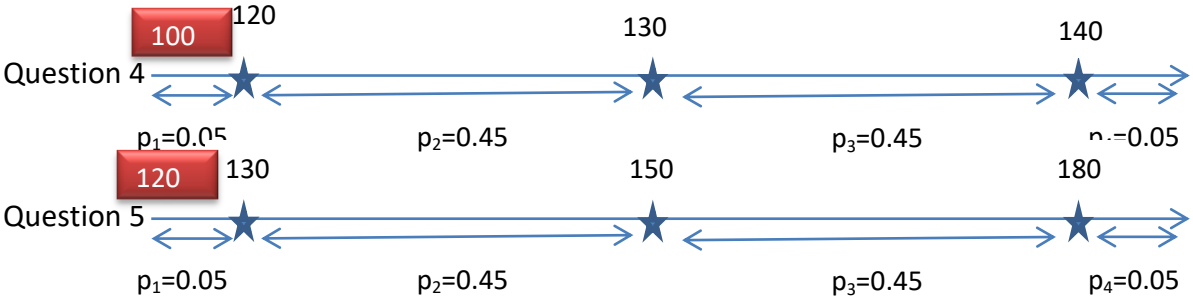


Figure 1: Realizations that fall into an interval for each question for Expert 1

Table 5: Sample distribution over four intervals for Expert 1

s_1	s_2	s_3	s_4
0.6	0.2	0.2	0.0

If the realizations are drawn independently from a distribution with three quantiles, as stated by the expert, then the quantity $2*N*I(s(e)|p)=2*N*\sum_{i=1..4}\{s_{i*}\ln(s_i/p_i)\}$
is asymptotically distributed as a chi-squared variable with 3 degrees of freedom. This is called ***likelihood ratio statistic*** and I(s(e)|p) corresponds to ***relative information*** of distribution *s* with respect to *p* for expert *e*. Relative information can be explained as follows: Let a discrete distribution have a probability function s and let a second distribution have a probability function *p*. Then the relative information of s with respect to p is s*ln(s/p), which is also called "Kullback-Leibler information entropy".
We will score expert *e* as the statistical likelihood of the following hypothesis:
H_e:"the inter-quantile interval containing the true value for each variable is drawn independently from probability vector *p*".
A simple test for this hypothesis uses the information likelihood ratio statistic and the probability value of this hypothesis forms the calibration score:
Calibration score (e): $Prob\{2*N*\ I(s(e)|p)\geq r|\ H_e\}$
This is the probability that information likelihood is equal or larger than *r*, given H_e is true, where *r* is the relevant quantity value from the expert's sample distribution.
Hence calibration score is the probability that under hypothesis He, that a deviation at least as great as r could be observed on N realizations. Although the calibration score uses the langaguage of hypothesis testing, it is NOT used to reject expert hypothesis, rather the terminology is used to measure the degree to which the data supports the hypothesis that the experts probabilities are accurate. High score means that it is likely that expert's probabilities are correct, while low score near zero means that it is unlikely that expert's probabilities are correct.
For example for expert 1:
r=10*(0.6*ln(0.6/0.05)+0.2*ln(0.2/0.45)+ 0.2*ln(0.2/0.45)+0)=
10*(0.6*2.48-0.2*0.81-0.2*0.81)=11.67
P(2*N* I(s(e) | p) ≥11.64)=1- P(2*N* I(s(e)|p) ≤11.67)=1-0.991=0.009.
Likelihood ratio statistics and calibration scores of all experts can be seen in Table 6.
Table 6. Likelihood ratio statistics and calibration scores

Expert	*r*	Calibration Score
Expert 1	1.167	0.008919
Expert 2	1.443	0.002366
Expert 3	1.443	0.002366

Information Scores
Loosely, information in a distribution is the degree to which the distribution is concentrated. Information cannot be measured absolutely but with respect to some background measure. Being concentrated or

spread out is measured relative to some other distribution. We will use the uniform distribution as the background measure.

Measuring information requires associating a probability density with each quantile assessment of each expert. To do this, a unique density function is adopted that complies with the expert's quantiles and minimally informative with respect to the background measure (A distribution that out of all distributions that match the given quantiles, has least information or deviation from the background distribution at other points between the elicited quantiles).

If the background measure is uniform, then the probability density is constant between the assessed quantiles, and is such that the total mass between quantiles agrees with the probability vector p. The background measure is NOT elicited from the experts as it must be the same for all experts. It is chosen by the analyst.

The uniform background measure requires a range. It could be determined by using "k% overshoot rule" that Classical model implements. According to this rule, the lowest value (q_5) and the highest value (q_{95}) among the experts' values are determined for each question as can be seen from the tables 7, 8, 9, 10 and 11.

Table 7: Lowest and highest values for question 1

Expert\Values	q_5	q_{50}	q_{95}
Expert 1	40	45	60
Expert 2	30 ☆	50	70 ☆
Expert 3	50	60	70

Table 8: Lowest and highest values for question 2

Expert\Values	q_5	q_{50}	q_{95}
Expert 1	75	100	125 ☆
Expert 2	100	110	120
Expert 3	60 ☆	70	80

Table 9: Lowest and highest values for question 3

Expert\Values	q_5	q_{50}	q_{95}
Expert 1	70	95	110 ☆
Expert 2	90	100	110
Expert 3	70 ☆	90	110

Table 10: Lowest and highest values for question 4

Expert\Values	q_5	q_{50}	q_{95}
Expert 1	120	130	140 ☆
Expert 2	100 ☆	120	140
Expert 3	110	120	130

Table 11: Lowest and highest values for question 5

Expert\Values	q_5	q_{50}	q_{95}
Expert 1	130	150	180 ☆
Expert 2	110 ☆	120	130
Expert 3	130	150	170

After the intervals are determined, it is extended and a new interval, which is called as ***intrinsic range*** as follows:

$I^*=[q_l,q_h]$

Where:

$q_l=q_5-k*(q_{95}-q_5)/100$

$q_h=q_{95}+ k*(q_{95}-q_5)/100$

The value *k* determined how much the interval will be extended. Larger *k* value makes experts look more informative. In order to obtain 10% overshoot, *k* should be equal to 10. Intrinsic ranges can be seen from Table 12 for all seed questions.

Table 12: Intrinsic ranges for all questions

Seed question (k)	q_{kl}	q_{kh}
1	26	74
2	53.5	131.5
3	66	114
4	96	144
5	103	187

By using intrinsic ranges, information score of expert *e* can be calculated as shown below:
Information score (e): Average relative information wrt Background

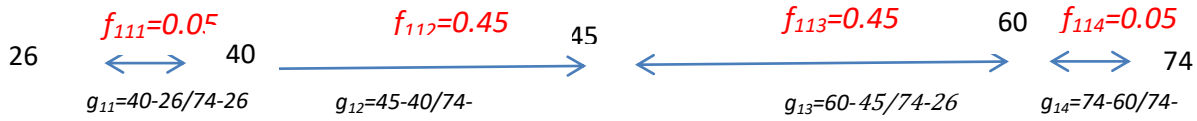
$$= \frac{1}{N}\sum_{k=1}^{N} I(f_{ek} \mid g_k)$$

g_k is the background probability density for variable (seed question variable) k over the extended intrinsic range and f_{ek} is expert e's probability or variable density function for variable k. The relative information over all variables (seed questions) are summed and normalized over N. This normalized sum is an indicator of informativeness of the expert (it is proportional to the relative information of the expert's joint distribution with respect to the background distribution).

$$I(f_{ek} \mid g_{ek}) = \sum_{i=1}^{4} f_{eki} \ln(f_{eki} / g_{eki})$$

For an expert *e*:
$g_{k1}=q_{k5}-q_{kl}/q_{kh}-q_{kl}$
$g_{k2}=q_{k50}-q_{k5}/q_{kh}-q_{kl}$
$g_{k3}=q_{k95}-q_{k50}/q_{kh}-q_{kl}$
$g_{k4}=q_{kh}-q_{k95}/q_{kh}-q_{kl}$
and f_{eki} values are 0.05, 0.45, 0.45 and 0.05 correspondingly for i=1,2,3 and 4 for all experts E and for all questions (*k*).

For Expert 1, Question 1



For example, for expert (simplified notation)

Question 1: g_1=40-26/74-26=0.292 g_2=45-40/74-26=0.104 g_3=60-45/74-26=0.313 g_4=74-60/74-26=0.292	Question 2: g_1=75-53.5/131.5-53.5=0.276 g_2=100-75/131.5-53.5=0.321 g_3=125-100/131.5-53.5=0.321 g_4=131.5-125/131.5-53.5=0.083
Question 3: g_1=70-66/114-66=0.083 g_2=95-70/114-66=0.521 g_3=110-95/114-66=0.313 g_4=114-110/114-66=0.083	Question 4: g_1=120-96/144-96=0.5 g_2=130-120/144-96=0.208 g_3=140-130/144-96=0.208 g_4=144-140/144-96=0.083

Question 5: g_1=130-103/187-103=0.321 g_2=150-130/187-103=0.238 g_3=180-150/187-103=0.357 g_4=187-180/187-103=0.083	

1/5*[0.05*ln(0.05/0.292)+0.45*ln(0.45/0.104)+0.45*ln(0.45/0.313)+0.05*ln(0.05/0.292)+
0.05*ln(0.05/0.276)+0.45*ln(0.45/0.321)+0.45*ln(0.45/0.321)+0.05*ln(0.05/0.083)+
0.05*ln(0.05/0.083)+0.45*ln(0.45/0.521)+0.45*ln(0.45/0.313)+0.05*ln(0.05/0.083)+
0.05*ln(0.05/0.5)+0.45*ln(0.45/0.208)+0.45*ln(0.45/0.208)+0.05*ln(0.05/0.083)+0.05*ln(0.05/0.321)+0.
45*ln(0.45/0.238)+0.45*ln(0.45/0.357)+0.05*ln(0.05/0.083)]=0.342

Information scores for all experts can be seen in Table 12.

Expert	Information Score
Expert 1	0.342
Expert 2	0.516
Expert 3	0.718

The information score does not depend on the realizations. An expert can give her/himself a high information score by choosing his quantiles very close together. The information score of e depends on the intrinsic range and on the assessments of the other experts. Hence, information scores cannot be compared across studies.

Weights of Experts

Weights of experts can be calculated by the multiplication of calibration and information scores. The analyst determines a threshold level so that the experts' opinions whose calibration score cannot exceed this threshold have zero weight. This achieved by using an indicator function. That is, the priority is on accuracy (calibration) rather than information.

$w_a(e)=Ind_a(\text{calibration score}(e))*\text{calibration score}(e)*\text{information score}(e)$
where Ind_a denotes an indicator function with $Ind_a(x)=0$ if $x<a$ and 1 otherwise.

The value of *a* is not predetermined and it depends on how much calibration expected from the experts to take their opinions into consideration is. It also can be determined so as to maximize the final decisions when the weights of experts are linearly pooled.
After the weights of experts are calculated, they are normalized to be linearly pooled with the formulation below. Normalized weight of expert e:
Normalized $w(e)=w(e)/\sum_e w(e)$
Normalized weights of all experts can be seen in Table 13.

Expert	Normalized weight
Expert 1	0.510521
Expert 2	0.204677
Expert 3	0.284802

References

W. Aspinal,
http://dutiosc.twi.tudelft.nl/~risk/extrafiles/EJcourse/Sheets/Aspinall%20Briefing%20Notes.pdf.