Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning (KR 2020)
Main Track

Towards a Logic of Meta-Analysis

Rafael Peñaloza
University of Milano-Bicocca, Italy
rafael.penaloza@unimib.it

Abstract

We currently have access to a plethora of statistical analyses based on sampling limited parts of a population. Meta-analysis is the task of combining several statistical results to obtain a more precise and reliable picture of the population. By the nature of sampling, all these results are uncertain, and difficult to combine with other knowledge. In this position paper, we propose a first approach for automated reasoning in meta-analyses.

1 Introduction

In the area of (logic-based) knowledge representation, our goal is to be able to express the knowledge of an application domain in a manner that allows for its effective use within intelligent applications. For this purpose, it is well-understood that methods capable of handling domain uncertainty are needed in most practical applications. This need has motivated the study of many uncertainty-representation languages, mostly based on probabilistic logics (Nilsson 1986; Fagin, Halpern, and Megiddo 1990; Halpern 1990; Dubois and Prade 1994).

Abstracting from language and interpretation differences, most probabilistic logics assign a *probability* (a number in $[0, 1]$) to some statements of the knowledge base. A common critique made in these cases is the source of the probabilities; i.e., how are these numbers obtained? A way to slightly answer to this critique is to weaken the probabilistic constraints, providing only some bounds (e.g., the probability is at least 0.5) or using imprecise probabilities (Walley 1991). The reality is that in all these cases, the issue is only being shifted: how do we determine the bounds? In fact, even imprecise probabilities often have precise limits.

Historically, we do have well-understood and robust methods for estimating probabilities, which are based on statistics. In particular, in scientific, dissemination, and other kinds of communications, we often encounter uncertainty represented through a confidence interval (Kiefer 1977). Confidence intervals are obtained by observing only a part of the population and are thus uncertain themselves. In particular, values outside of the interval are still possible (if unlikely). Interestingly, we often encounter several different studies—each providing its own confidence interval—for a single property of interest. Rather than merely intersecting all these intervals, a statistical meta-analysis combines the results of the studies to obtain a more accurate and representative interval. With this information and other statistical tools, it is possible to identify biases or potentially falsified numbers, among other things. A typical example arises in the area of political poling. In them, results are often provided with an associated margin of error. If two polls say that Candidate A will receive, say, $40\% \pm 3\%$ and $42\% \pm 1\%$ of the votes, respectively, we do not immediately conclude that the results are contradictory, or that the actual interval is 41–43%.

Surprisingly, despite its ubiquituous use for population analysis, these kinds of statistical analyses have been largely ignored in knowledge representation. In this position paper, we argue the need for a logic-based representation language capable of handling standard meta-analytical tasks, and more; and present the first steps towards this goal. As we show, dealing with confidence intervals is not a trivial task, as it requires understanding the statistical methods behind them. However, with a few simplifying assumptions and computations, it is possible to obtain robust automated reasoning methods over them.

2 Binomial Confidence Intervals

For the scope of this paper, we are interested in finding out the proportion of a population of interest that satisfies some given properties. As populations are big, and identifying the properties may require costly or intrusive procedures, it is unfeasible to expect to know these proportions precisely; for example, asking every adult for their voting preference is extremely costly. Instead, we rely on statistical techniques that provide approximate knowledge on them. A simple such technique is based on the normal approximation (Wallis 2013), which we briefly describe next.

Suppose that $100p\%$ of the population has property P. We can equivalently express this by saying that the probability of a randomly chosen individual to have property P is p. By the central limit theorem, if n individuals are independently chosen at random, the proportion of elements in this sample that satisfy P (denoted as $\hat{p}$) behaves as a normal distribution with mean p and variance $\frac{p(1-p)}{n}$. If we are only interested in a single value that estimates the true proportion p, we could use $\hat{p}$, which is in fact the maximum likelihood estimator (mle) of p. However, it is more informative to consider a *confidence interval*, which also takes into account the

variability of the sampling process. Although the notion of a confidence interval allows for any confidence degree, the standard scientific and journalistic practices use only a 95% confidence. To simplify the presentation, we consider this restricted setting as well for most of this paper.

In essence, given a sample of size n, with a sample proportion of property P of $\hat{p}$, there is a 95% probability that the *true* proportion p lies within the interval

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right).$$

When reporting these results, the value $\pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is often called the *margin of error*. Note that the constant 1.96 appears only because we are interested in the 95% confidence interval. For other confidence levels, the only difference in the interval would be the use of this constant.

Importantly, a confidence interval expresses a 95% chance that the true value lies in it. Thus, it is totally plausible that p is not in it. Moreover, if several such samples are carried out, the likelihood of at least one of them not containing the true proportion increases. This means that intersecting the intervals is not a good idea, when trying to summarise them into a single result. Still, all the samples are informative.

Note that sampling information is commonly expressed through the central estimator $\hat{p}$ and the margin of error e, but detailed information such as the sample size or the number of successes is often ommited. These can be obtained by recalling that $e = 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, and hence $n \approx 1.96^2 \frac{\hat{p}(1-\hat{p})}{e^2}$. Suppose now that two independently taken samples of sizes n_1 and n_2 yield the mles $\hat{p}_1$ and $\hat{p}_2$, respectively. This can be seen as a unique sample of size $n = n_1 + n_2$ with a proportion of successes equal to $\hat{p} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$. This idea can be generalised in the obvious way to combine the results of multiple independent samples.

While this result is very well understood, it relates a population to one specific property. On the other hand, we often have data about different properties that we would like to relate, or about different populations which we would like to combine for a broader understanding, or to compensate for a lack of data. For example, suppose that we want to find out the proportion of parents intending to vote for Candidate A without having to make a new sample. Or, from samples in different regions, find the global vote intention. This is where KR comes into place to help statistics.

3 CI Representation

We present a simple language, which allows to express subclass relationships and confidence intervals relating properties of individuals. Consider a countable set X of *variables*, which we will often call *properties*. A *subclass statement* (SS) is an expression of the form $\bigwedge_{i=1}^{n} x_i \to y$, where $x_i, y \in X$, $1 \leq i \leq n$. A *confidence interval statement* (CIS) is an expression of the form

$$\left(\bigwedge_{j=1}^{m} y_j \mid \bigwedge_{i=1}^{n} x_i\right) : p \pm e$$

where $x_i, y_j \in X$, $1 \leq i \leq n$, $1 \leq j \leq m$, and $p, e \in (0, 1)$. The variables x_i in an SS or a CIS are called its *body*, while the variable y or the variables y_j are its *head*.

In a nutshell, subclass statements $\bigwedge_{i=1}^{n} x_i \to y$ are just Horn clauses, which express that individuals who have all the properties $x_1, \ldots, x_n$ must also have property y. On the other hand, the CIS $\left(\bigwedge_{j=1}^{m} y_j \mid \bigwedge_{i=1}^{n} x_i\right) : p \pm e$ expresses that $100p\%$ of individuals who belong to $\bigwedge_{i=1}^{n} x_i$ also belong to $\bigwedge_{j=1}^{m} y_j$—or, more formally, that the probability of an individual belonging to the first conjunction to have all the properties of the second conjuction is p—with a margin of error e. A *CI knowledge base* (KB) is a finite set of SSs and CISs. Note that SSs have only one property in the head, while CISs have a conjunction. The reason for this will become clear once we introduce the semantics.

To formalise this semantics, we interpret properties as sets in the usual first-order fashion. An *interpretation* is a pair $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$, where $\Delta^{\mathcal{I}}$ is a non-empty set called the *domain*, and the interpretation function $\cdot^{\mathcal{I}}$ maps every $x \in X$ to a set $x^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$. This function is extended to conjunctions of properties in the obvious manner; that is, $(\bigwedge_{i=1}^{n} x_i)^{\mathcal{I}} = \bigcap_{i=1}^{n} x_i^{\mathcal{I}}$. The interpretation $\mathcal{I}$ *satisfies* the SS $\bigwedge_{i=1}^{n} x_i \to y$ iff $\bigcap_{i=1}^{n} x_i^{\mathcal{I}} \subseteq y^{\mathcal{I}}$.

The semantics of CISs is more subtle. Recall that confidence intervals are built from a partial observation of the population, and are thus surrounded by uncertainty. In fact, the only absolute guarantee that we can make from a non-trivial CI is that some individuals comply with the conditional property while others do not. Formally, the interpretation $\mathcal{I}$ satisfies the CIS $\left(\bigwedge_{j=1}^{m} y_j \mid \bigwedge_{i=1}^{n} x_i\right) : p \pm e$ iff there are elements $\delta, \gamma \in \Delta^{\mathcal{I}}$ such that $\delta \in \bigcap_{i=1}^{n} x_i^{\mathcal{I}} \cap \bigcap_{j=1}^{m} y_j^{\mathcal{I}}$ and $\gamma \in \bigcap_{i=1}^{n} x_i^{\mathcal{I}} \setminus \bigcap_{j=1}^{m} y_j^{\mathcal{I}}$. $\mathcal{I}$ is a *model* of the KB $\mathcal{K}$ iff it satisfies all the SSs and CISs in $\mathcal{K}$. $\mathcal{K}$ *entails* the SS $\bigwedge_{i=1}^{n} x_i \to y$ iff every model of $\mathcal{K}$ satisfies this SS. If that is the case, we denote it by $\mathcal{K} \models \bigwedge_{i=1}^{n} x_i \to y$.

We can readily see at this point that conjunctions on the heads of SSs do not affect the semantics; e.g., $x \to y_1 \wedge y_2$ is equivalent to $x \to y_1$, $x \to y_2$. However, this is not true for CISs; e.g., $(y_1 \wedge y_2 \mid x) : p \pm e$ is not equivalent to the two CISs $(y_1 \mid x) : p \pm e$ and $(y_2 \mid x) : p \pm e$. The differences are more pronounced once we consider the meaning of these confidence intervals as well. Note that, although we cannot instill the requirements of a confidence interval into the notion of satisfiability of a CIs by an interpretation, we will always consider a CIS as a fact obtained through a sound statistical analysis. Moreover, any two CISs in a KB will be considered to be independent; i.e., obtained through an independent sampling method. Intuitively, this means that the sampled individuals are not related throughout the samples.

The question of *consistency*—that is, deciding whether there exists a model of a given KB—can be easily reduced to consistency of a logic program, and hence is decidable in polynomial time. We are more interested in extracting (probabilistic) information from the class of CISs, and using it to make decisions. Two basic meta-analysis inferences are accumulation and rescaling of confidence intervals. These

can be used, together with SSs, to analyse bias and detect suspicious intervals.

3.1 Accumulation

As mentioned earlier, a confidence interval is not guaranteed to contain the proportion of interest, but even in those cases it is informative as it shows the result of a statistical analysis. In its most basic form, a meta-analysis will use several CIs to obtain a more precise approximation of the desired value.

Suppose that the KB contains the CISs $(y \mid x) : p_i \pm e_i$, $1 \leq i \leq k$, and that these are all the CISs of y given x. We want to produce one CIS which summarises all the information of these n CISs. From the k different analyses, we can extract the estimators of the sample sizes (see Section 2) $n_i := 1.96^2 \frac{p_i(1-p_i)}{e_i^2}$ and through the weighted average method get $\hat{n} := \sum_{i=1}^{k} n_i$ and $\hat{p} := \frac{\sum_{i=1}^{k} p_i n_i}{\hat{n}}$. Thus, we can substitute the k CISs with the more accurate $(y \mid x) : \hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{\hat{n}}}$

For example, consider two CISs $(y \mid x) : 0.24 \pm 0.084$ and $(y \mid x) : 0.28 \pm 0.12$, which were obtained by sampling a population with 100 and 50 individuals, respectively—although that information is not given, nor necessary for our approach. These two CISs can be accumulated into the single statement $(y \mid x) : 0.2533 \pm 0.07$, which is a new 95% confidence interval for the same conditional statement. Note that the center of this new interval is not merely the average of the other two centers; in addition, the interval is not the intersection of the other two (which would be $(0.16, 0.324)$), and it is shorter.

The accumulation of CISs can be easily generalised to arbitrary conditional statements containing conjunctions in the body and in the head. In fact, all the computations depend exclusively on the numbers of the CI, and are not influenced by the properties in the conditional.

3.2 Rescaling

Even though the standard for communicating confidence intervals is 95%, this does not mean that other levels of confidence are not of interest. Since we are assuming that all CISs represent confidence intervals at a level of 95%, we know that its margin of error is using the constant factor 1.96. To obtain a different confidence level, it suffices to multiply the error by the constant $\frac{z_{\alpha/2}}{1.96}$, where $z_{\alpha/2}$ is the z-value of the standard normal distribution representing the point after which the probability of observing a value is $\alpha/2$, and the desired confidence level is $100(1-\alpha)$.

We have previously found the CIS $(y \mid x) : 0.2533 \pm 0.07$ which represents a 95% confidence interval for the conditional property $(y \mid x)$. For a 99% confidence interval, we need to know the value of $z_{0.005}$, which is 2.57. Then, $0.2533 \pm 0.07\frac{2.57}{1.96}$ is such an interval for the conditional property $(y \mid x)$. Instead, a 90% CI is $0.2533 \pm 0.07\frac{1.65}{1.96}$.

3.3 Using Background Knowledge

For the previous two inferences, we needed only the properties of confidence intervals and meta-analyses, but it is more interesting to combine them with the background knowledge expressed by the SSs in the KB. First of all, consider the case where we know that $\mathcal{K} \models x \to y$ and $(z \mid y) : p \pm e \in \mathcal{K}$, but we are interested in a confidence interval for $(z \mid x)$. In practical terms, we know the behaviour of a population (y) w.r.t. a property (z), but we are more interested in knowing how a subpopulation (x) behaves. In general, unless x and y are equivalent (i.e., $\mathcal{K} \models y \to x$ as well), we cannot deduce any CI for $(z \mid x)$: it could well be the case that none of the individuals in x satisfy z, or that all of them do, or anything in between. However, assuming that the subpopulation is *not biased* against the property z w.r.t. y, we can approximate the desired CI by simply using the known parameters from the CIS in $\mathcal{K}$. That is, $p \pm e$ is an approximate interval for $(z \mid x)$. Note that even this simple deduction can aid in complex inferences. In fact, it can be used as a preprocessing step to obtain more CISs to be accumulated when trying to get a better understanding of a conditional property, or dually, could be used after accumulation to obtain some new information.

Let us consider now the case where we have a confidence interval for a conditional property—perhaps obtained through accumulation and deduction—which we trust, and are given a new interval, for which we want to evaluate how likely it is, given our previous knowledge. Recall that a 95% confidence interval states that the true parameter is contained in this interval with probability 0.95. If two confidence intervals intersect (even with a very short intersection), then they are coherent with each other, and we cannot refute any of them under the 95% confidence being used. If they do not intersect, then assuming that one of them is correct means that the other has a likelihood of less than 5% of being a real confidence interval. But the question is whether we can provide a more precise estimate for this likelihood. Clearly, the farther away the two intervals are, the easier it should be to refute one of them (as their mutual likelihood decreases).

Suppose that we know the CIS $(y \mid x) : p \pm e$ and want to find the likelihood of an interval $(y \mid x) : p' \pm e'$ where $p \pm e$ and $p' \pm e'$ do not intersect. Suppose w.l.o.g. that $p+e < p'-e'$; the opposite case can be treated analogously. Recall that we can rescale the confidence interval by simply multiplying a factor to the error bound. We are interested in finding the highest confidence interval which still excludes $p'-e'$; that is, we want a constant f such that $p+f = p'-e'$, which is easily computed as $f = p' - e' - p$. To rescale, we want a $z_{\alpha/2}$ such that $e\frac{z_{\alpha/2}}{1.96} = f$; that is, $z_{\alpha/2} = 1.96f/e$. By using the inverse z relation, we get the value α such that the $100(1-\alpha)\%$ confidence interval still excludes $p' \pm e'$.

For example, consider the CIS $(y \mid x) : 0.2533 \pm 0.07$ which we have computed before, and suppose that we are told that, through a new sample, the interval 0.4 ± 0.06 was found. Then, we obtain $f = 0.4 - 0.06 - 0.2533 = 0.0867$ and $z_{\alpha/2} = 2.83$, which means that $\alpha/2 = 0.0023$. That is, the likelihood of the interval 0.4 ± 0.06 to be a result from an analysis of the same population is less than 0.5%. Thus, we can safely refute this interval, as not being trustworthy.

As a final inference, recall that at the beginning of this section we described a method for approximating a confidence interval for a subpopulation, under the assumption that the subpopulation is unbiased w.r.t. the desired prop-

erty. Using this approximation, and the method for refuting unlikely intervals just described, it is also possible to check whether the population is biased or not. Suppose that from the KB we can deduce $x \to y$, $(z \mid y) : p \pm e$ and $(z \mid x) : p' \pm e'$. If the subpopulation x was unbiased, then $p \pm e$ and $p' \pm e'$ should be compatible confidence intervals for $(z \mid x)$. If these intervals do not intersect, we can compute the likelihood of obtaining them together, giving us a likelihood of bias.

4 Combining Intervals

Going beyond the usual tasks of meta-analysis, one may want to combine the known confidence intervals for different properties to obtain new ones, which have not been studied yet. We study two such approaches next.

First, consider two CISs $(y_i \mid x) : p_i \pm e_i$ with $i = 1, 2$. Once again, it is in general impossible to compute an interval for $(y_1 \wedge y_2 \mid x)$ without further knowledge, as we do not know how y_1 and y_2 are correlated. In cases where we can assume independence of y_1 and y_2, we are trying to compute an estimate for the product of two proportions (Buehler 1957). We can achieve this following the well-known *Delta method* (Doob 1935). In essence, we need to approximate the mean and the variance of the distribution of the product parameter. The center estimator is, as expected, just the product of the two estimators; e.g., $\hat{p} = p_1 p_2$; the more complex part is to compute the variance. This can be approximated by the function

$$Var(\hat{p}) = p_1^2 \left(\frac{e_2}{1.96}\right)^2 + p_2^2 \left(\frac{e_1}{1.96}\right)^2,$$

which tells us that a 95% confidence interval for $(y_1 \wedge y_2 \mid x)$ is

$$p_1 p_2 \pm 1.96 \sqrt{p_1^2 \left(\frac{e_2}{1.96}\right)^2 + p_2^2 \left(\frac{e_1}{1.96}\right)^2}.$$

If rather than trying to conjoin the variables in the head of CISs we are interested in chaining conditionals, we can follow a similar approach thanks to Bayes' rule. Given the two CISs $(z \mid y) : p \pm e$ and $(y \mid x) : p' \pm e'$, it is impossible to deduce an interval for $(z \mid x)$, mainly because we have no information about the proportion of elements that are in x but not in y who still have the property z. However, we can compute one for $(y \wedge z \mid x)$. Indeed, by Bayes rule, we know that $P(y \wedge z \mid x) = P(z \mid y \wedge x) P(y \mid x)$. Formally, we do not know the first factor of this product, but under the covering rule, when $\mathcal{K} \models y \to x$ then the equality reduces to $P(y \wedge z \mid x) = P(z \mid y) P(y \mid x)$. This means that to estimate the proportion of elements of x that have both properties y, z, we need to make the product of two estimators as done in the previous paragraph.

Finally, suppose that we know that two properties y_1 and y_2 partition the population x and, in contrast with the assumptions made so far in this paper, we know the *exact* proportion of elements of x that satisfy y_1; say that this proportion is ℓ. Then, from the CISs $(z \mid y_i) : p_i \pm e_i$, $i = 1, 2$, we can compute a CI for $(z \mid x)$ using the mixture distribution of the two known conditionals (Seidel 2011; Frühwirth-Schnatter 2006). In this case, the center of the interval is the weighted sum of the two centers p_1, p_2 based on the proportion ℓ which we have assumed to know. That is, $\hat{p} = p_1 \ell + p_2 (1 - \ell)$. For the variance of the estimator, the construction is slightly more complex, but it results in $\hat{v} := \ell \left(\frac{e_1}{1.96}\right)^2 + (1 - \ell) \left(\frac{e_2}{1.96}\right)^2 + \ell(1 - \ell)(p_1 - p_2)^2$. Thus, we can conclude that the 95% confidence interval for $(z \mid x)$ is

$$p_1 \ell + p_2 (1 - \ell) \pm 1.96 \sqrt{\hat{v}}.$$

5 Conclusions

We have considered the first steps towards a logic for performing automated meta-analysis based on a finite class of confidence intervals and subset relationships as background knowledge. For this purpose, we assume that all confidence intervals are expressed with the *de-facto* standard of a 95% confidence, and obtained through independent sampling. Moreover, we consider that the intervals for the binomials—which are the only kind of distribution that we treat in this paper—are all built using the normal approximation based on the central limit theorem.

In this paper, we showed how to make basic inferences for combining the confidence intervals akin to the tasks commonly required in a meta-analysis, and went beyond with a few novel derivations. One of the goals of this paper is to show that even at this basic level, dealing with confidence intervals is far from trivial, and requires computations not commonly available in logic. This is perhaps why, despite widely available statistical theory and techniques, this kind of logic has been widely disregarded in the past.

Obviously, the formalism presented here has several limitations, which need to be handled in future work. First and foremost, the language considered is very inexpressive. From the logical side, it cannot express any negations, nor any complex relationship between properties. From the probabilistic perspective, it is limited to a class of confidence intervals for the parameter of a binomial. Moreover, throughout the paper we assume that all CIs are independent, which stops being true when further derivations are based in implicit consequences computed before. Our main goal is to make this formalism robust, allowing for different statistical analyses, and a more expressive language for the background knowledge which allows for a tighter integration between logic and statistics.

The main goal of this paper is to serve as a starting point for a more ambitious agenda, in which knowledge representation and statistical methods coexist.

References

Buehler, R. J. 1957. Confidence intervals for the product of two binomial parameters. *Journal of the American Statistical Association* 52(280):482–493.

Doob, J. L. 1935. The limiting distributions of certain statistics. *The Annals of Mathematical Statistics* 6(3):160–169.

Dubois, D., and Prade, H. 1994. Non-standard theories of uncertainty in knowledge representation and reasoning. In Doyle, J.; Sandewall, E.; and Torasso, P., eds., *Principles of Knowledge Representation and Reasoning*, The Morgan Kaufmann Series in Representation and Reasoning. Morgan Kaufmann. 634 – 645.

Fagin, R.; Halpern, J. Y.; and Megiddo, N. 1990. A logic for reasoning about probabilities. *Information and Computation* 87(1):78 – 128. Special Issue: Selections from 1988 IEEE Symposium on Logic in Computer Science.

Frühwirth-Schnatter, S. 2006. *Finite mixture and Markov switching models*. Berlin: Springer, 1st edition.

Halpern, J. Y. 1990. An analysis of first-order logics of probability. *Artif. Intell.* 46(3):311–350.

Kiefer, J. 1977. Conditional confidence statements and confidence estimators. *Journal of the American Statistical Association* 72(360a):789–808.

Nilsson, N. J. 1986. Probabilistic logic. *Artif. Intell.* 28(1):71–88.

Seidel, W. 2011. Mixture models. In Lovric, M., ed., *International Encyclopedia of Statistical Science*. Springer. 827–829.

Walley, P. 1991. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall.

Wallis, S. 2013. Binomial confidence intervals and contingency tests: Mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics* 20(3):178–208.