Representing Sampling Distributions In P-$\mathcal{SROIQ}$

Pavel Klinov[1] and Bijan Parsia[2]

[1] University of Arizona, AZ, USA pklinov@email.arizona.edu
[2] The University of Manchester, UK bparsia@cs.man.ac.uk

Abstract. We present a design for a (fragment of) Breast Cancer ontology encoded in the probabilistic description logic P-$\mathcal{SROIQ}$ which supports determining the consistency of distinct statistical experimental results which may be described in diverse ways. The key contribution is a method for approximating sampling distributions such that the inconsistency of the approximation implies the statistical inconsistency of the continuous distributions.

1 Introduction

The current amount of knowledge about breast cancer is overwhelming. For example, a meta-study conducted in 2006 by Key et al. [4] covered 98 *unique* studies focused only on the impact of a single risk factor, alcohol consumption. At the same time there are no common knowledge bases which would combine and formally represent findings produced by the multitude of studies.[3] This makes it difficult to have a global view of breast cancer risk factors and, consequently, develop tools like risk assessment calculators.

The probabilistic description logic P-$\mathcal{SROIQ}$ can be used to represent general knowledge about breast cancer in the form of a probabilistic ontology (the BRC ontology) [5]. However, a general knowledge ontology need not support risk entailments for various combinations of risk factors — that is, compete (poorly) with narrowly specific risk calculators[4] which have a direct implementation of simple equations derived from statistical risk models (such as the Gail model [2]). Instead, its main goal is to formally and unambiguously describe the background theory of breast cancer embracing as many reliable findings as possible and serving as a common knowledge base for more specific tools, such as risk assessment calculators or decision support systems. This sort of task seems to be a better fit for a probabilistic logic.

The set of use cases for the general knowledge ontology is wider than for the BRC ontology. In addition to maintaining a birds-eye view of breast cancer, it may be used for finding and analyzing inconsistencies in outcomes of different

[3] There are some lower level databases, such as ROCK (http://rock.icr.ac.uk/)—a cancer specific functional genomic database. However, they do not explicitly represent case study findings and do not support such services as risk assessment.

[4] Such as http://www.cancer.gov/bcrisktool

studies. It can support studying mechanisms of interactions between risk factors, for example, how alcohol consumption affects estrogen level. Finally, it may play a useful role in planning and coordination of future medical studies by helping to identify the most controversial or insufficiently studied risk factors or exposures.

In this paper, we present a design of general P-$\mathcal{SROIQ}$ ontology about breast cancer (i.e., the BRC ontology) which incorporates a substantial amount of statistical knowledge. While we do not present a fully fleshed out instance of this design, we do tackle a major representational challenge, namely, the representation of the statistical results of experiments. We present a method for approximate representations of different sampling distributions and their use in determining consistency between experimental data.

2 Preliminaries of P-$\mathcal{SROIQ}$

P-$\mathcal{SROIQ}$ [8] is a probabilistic extension of the DL $\mathcal{SROIQ}$ [3]. It provides means for expressing probabilistic relationships between arbitrary $\mathcal{SROIQ}$ concepts and a certain class of probabilistic relationships between classes and individuals. Any $\mathcal{SROIQ}$, and thus OWL 2 DL (as it can be seen as a notational variant of $\mathcal{SROIQ}$), ontology can be used as a basis for a P-$\mathcal{SROIQ}$ ontology, which facilitates transition from classical to probabilistic ontologies. We presume the reader is reasonbly familiar with class/object oriented description logics such as $\mathcal{SROIQ}$, though very little in this paper turns on specific details.

The only syntactic construct in P-$\mathcal{SROIQ}$ (in addition to all of the $\mathcal{SROIQ}$ syntax) is the conditional constraint.

Definition 1 (Conditional Constraint). *A conditional constraint is an expression of the form $(D|C)[l,u]$, where C and D are concept expressions in $\mathcal{SRIQ}$ (i.e., $\mathcal{SROIQ}$ without nominals) called **evidence** and **conclusion**, respectively, and $[l,u] \subseteq [0,1]$ is a closed real-valued interval. In the case where C is $\top$ the constraint is called **unconditional**.*

Ontologies in P-$\mathcal{SROIQ}$ are separated into a classical and a probabilistic part. It is assumed that the set of individual names N_I is partitioned onto two sets: classical individuals N_{CI} and probabilistic individuals N_{PI}.

Definition 2 (PTBox, PABox, and Probabilistic Knowledge Base). *A **probabilistic TBox** (PTBox) is a pair $PT = (\mathcal{T}, \mathcal{P})$ where $\mathcal{T}$ is a classical (finite) $\mathcal{SROIQ}$ TBox and $\mathcal{P}$ is a finite set of conditional constraints. A **probabilistic ABox** (PABox) is a finite set of conditional constraints associated with a probabilistic individual $o_p \in N_{PI}$. A **probabilistic knowledge base** (or a probabilistic ontology) is a triple $PO = (\mathcal{T}, \mathcal{P}, \{\mathcal{P}_{o_p}\}_{o_p \in N_{PI}})$, where the first two components define a PTBox and the last is a a set of PABoxes.*

Informally, a PTBox constraint $(D|C)[l,u]$ expresses a conditional statement of the form "if a *randomly* chosen individual is an instance of C, the probability of it being an instance of D is in $[l,u]$". A PABox constraint, which we write

as $(D|C)_o[l,u]$ where o is a probabilistic individual, states that "if a *specific* individual (that is, o) is an instance of C, the probability of it being an instance of D is in $[l,u]$". For more details we refer the reader to [8].

3 The Classical Part

The classical part of a P-$\mathcal{SROIQ}$ ontology (or OWL part) provides a medical vocabulary which can be used on its own in a variety of applications or used in the representation of probabilistic knowledge. In this paper we focus on providing an OWL terminology for probabilistic statements. The ontology contains the following main class hierarchies (taxonomies):

Taxonomy of breast cancers Breast cancer is a heterogeneous disease. Some risk factors can be associated with increase in risk of developing one particular type of breast cancer and not the other. Thus it is important to classify types of breast cancer. In particular, our ontology distinguishes breast cancers by hormone receptor status. Estrogen and progesterone positive breast cancers are modeled using concepts `ERPositiveBRC` and `PRPositiveBRC` while their complements are modeled using `ERNegativeBRC` and `PRNegativeBRC` (we use shorthands ER+/- and PR+/- with obvious meaning.). Another important classification is based on histology. The ontology distinguishes between invasive and non-invasive (e.g. in situ) cancers.

Taxonomy of risk factors Dozens of risk factors are known so far. Some are established and strongly associate with increased risks, such as BRCA1(2) gene mutations, while others are controversial. The ontology should provide vocabulary for both to support current and future findings. It includes a taxonomy of concepts rooted at `RiskFactor`. We distinguish between known risk factors (those which can be reported via a questionnaire, such as alcohol intake) and inferred risk factors which require medical examination.

Taxonomy of risks The ontology differentiates absolute and relative risks of developing breast cancer. Absolute risks are further divided into the lifetime risk and the short-term risk. Relative risks are divided into increased and reduced risks. Level of increases is a continuous variable which requires discretization (see below).

The last two taxonomies induce the corresponding classifications of women, i.e., classes of women w.r.t. risk factors and w.r.t. risk. For example, any risk factors `RF` gives rise to a class of women $\texttt{Woman} \sqcap \exists \texttt{hasRiskFactor.RF}$. Women having various combinations of risk factors are modeled as conjunctive concept expressions. Analogously, given a certain kind of risk `R` the expression $\texttt{Woman} \sqcap \exists \texttt{hasRisk.R}$ models those women who are in the risk group `R`, for example, have moderately increased risk of developing ER+ breast cancer. These taxonomies of women may or may not be explicitly present in the ontology. In other words, it is possible, but not essential, to generate a concept name for each interesting class of women since P-$\mathcal{SROIQ}$ (and our reasoner Pronto) allows for complex concept expressions in conditional constraints.

A future, more complete version of the ontology would certainly make use of existing bio-medical ontologies which cover substantial portions of the domain either by direct reuse or by ontology alignment techniques.

4 The Method for Approximating Distributions in the Probabilistic Part

The probabilistic part of the ontology captures statistical background knowledge about breast cancer. We distinguish between knowledge which explicitly associates quantifies specific risk factors and more general statistical relationships which are not necessarily risk related. The distinction could be useful for importing knowledge from other medical ontologies. We begin with the latter.

General statistical knowledge mostly includes relationships between various risk factors. For example, Ashkenazi Jew women are more likely to develop BRCA gene mutations, while early menarche, late first child (or no live births), lack of breastfeeding and alcohol consumption all increase levels of estrogen in blood.[5] Such relationships are important because they can help to infer the presence of some risk factors given the set of known factors. They are typically easy to represent by using conditional constraints of the form `(Woman`$\sqcap\exists$`hasRiskFactor.RFY|Woman`$\sqcap\exists$`hasRiskFactor.RFX)[l,u]` which says that the chances of having risk factor RFY given RFX are between l and u. One possible source of complications is continuous variables, e.g. the level of estrogen, which are discussed below.

Most of statistical findings available in medical literature quantitatively describe risk increase for categories of women with specific risk factors. Such findings are presented by giving estimated parameters of a probability distribution where the random variable represents the relative risk of a random woman in the population. Such parameters include the estimated mean value and the estimated variance. Table 1 presents an example of the reported association between alcohol intake and the risk increase among postmenopausal women taken from [10]. There are two main difficulties with representing this kind of data in P-$\mathcal{SROIQ}$. First, the risk increase is a continuous random variable so it needs to be discretized. Second, the available language supports only conditional constraints so a straightforward encoding of probability distributions is not possible.

Table 1. Example of a reported association between alcohol intake and the risk of hormone receptor-specific breast cancer (excerpt from [10])

Alcohol (g)	**ER+**	**ER-**	**PR+**	**PR-**
	RR (95% CI)	**RR (95% CI)**	**RR (95% CI)**	**RR (95% CI)**
0	1.00 (ref)	1.00 (ref)	1.00 (ref)	1.00 (ref)
$\leq$4	1.06 (0.91 - 1.22)	1.40 (1.00 - 1.96)	1.04 (0.89 - 1.23)	1.24 (0.95 - 1.62)
$\geq$4	1.07 (0.90 - 1.26)	1.64 (1.14 - 2.35)	1.12 (0.93 - 1.34)	1.28 (0.96 - 1.71)

[5] See `http://tinyurl.com/4jpsdvk`

Discretization of a continuous variable is technically straightforward. We introduce a set of disjoint concept names each of which models women in the corresponding group of risk. Specifically, we define concepts `WomenAtWeakRisk`, `WomenAtModerateRisk` and `WomenAtHighRisk` with the obvious meanings described using OWL 2 datatype support to describe the exact boundaries. We have chosen ranges (1, 1.5], (1.5, 3.0] and (3.0, $+\inf$) respectively.[6]

The inability to represent distributions is a more severe limitation. It leaves the modeler with the only option of *approximating* the continuous distribution using a finite set of points. In other words, each distribution, for example, risk increase for women consuming a certain amount of alcohol, can be approximated by specifying the probability that a *randomly* taken woman with the given exposure belongs to a specific group of risk, i.e. `WomenAtWeakRisk`, `WomenAtModerateRisk` or `WomenAtHighRisk`. This *is* the semantics of P-$\mathcal{SROIQ}$ conditional constraints.

Assuming that the random variable is real-valued, a standard way of approximating a continuous distribution is to take each interval and compute the probability that the variable takes on a value in that interval. Then the approximation of a distribution $Pr(x)$ w.r.t. a finite set of intervals U is simply a function $\hat{Pr}$ such that $\hat{Pr}(U_i) = \int_{U_i} Pr(x)\mathrm{d}x$.

Unfortunately, this approximation of results of statistical experiments is unsatisfactory because it maps every interval to a single point. The problem is that *any* arbitrarily small difference between two or more sampling distributions will results in conflicting probabilistic statements for every interval (because the point-valued probabilities will be different) even though the results can confirm each other from a purely statistical point of view. Consequently this approach does not support working with results reported by multiple studies.

Our goal is to approximate sampling distributions in P-$\mathcal{SROIQ}$ in a *statistically coherent* way. Informally it means that satisfiability of probabilistic formulas representing two or more sampling distributions must agree with their mutual statistical consistency, i.e., whether they support a common statistical hypothesis. The hypothesis, in this case, is that there exists a distribution (not necessarily a unique one) over G with parameters μ, σ such that it is supported by all sampling distributions with the required level of confidence.

We assume a (finite) population G of size N_G and a random variable X which is normally distributed across G. We also make the realistic assumption that G is large enough so that evaluating X for all members of G is not feasible. A common approach is to take one or more random samples from G, evaluate X for them and estimate the actual distribution over G based on the sampling distributions. We use μ, σ to denote the mean and the variance of the actual distribution and $\overline{X^{(i)}}, S^{(i)}$ for the mean and the variance of the sample $X^{(i)}$. For simplicity we finally assume that the population distribution is normal.

The mainstream approach for comparing two or more sampling distributions is based on statistical hypothesis tests. For example, given two normal distribu-

[6] The choice of intervals is obviously ambiguous but this issue is orthogonal to the approximation method presented in this paper.

tions $\overline{X^{(1)}}, S^{(1)}, \overline{X^{(2)}}, S^{(1)}$ it is common to take $\overline{X^{(1)}} - \overline{X^{(2)}}$, which is a normally distributed random variable, and perform a *z-test* (or a Student's t-test depending on the sample sizes) to see if the difference can be taken as 0 with the required level of confidence. It amounts to calculating standard errors of the mean (SE) for both distributions and then computing the difference *in units of SE*. If the probability of observing such difference given the null hypothesis,[7] which can be found in standard tables, is low enough, e.g., ≤ 0.05, a statistician would accept the hypothesis that both distributions are consistent.

Our approach is slightly different from the outlined above. It is not based on tests but on *confidence regions* for sampling distributions. The approach, which generalizes confidence intervals and dates back to Mood [9], is to estimate a region $\mathcal{R}_\gamma$ in the parameter space for (μ, σ^2) such that it will contain the μ, σ^2 pair of the actual distribution $100(1-\gamma)\%$ times as the number of estimations goes to infinity. More formally, a $100(1-\gamma)\%$ confidence region $\mathcal{R}_\gamma$ is a *random set* for parameters (μ, σ^2) based on a group of independent normally distributed variables X (i.e., a sample) such that [1]:[8]

$$P((\mu, \sigma^2) \in \mathcal{R}_\gamma) = 1 - \gamma, \text{ for all } (\mu, \sigma^2) \tag{1}$$

Informally, the confidence region specifies how far sampling distributions can deviate from the population distribution while supporting it with $100(1-\gamma)\%$ confidence. Following Mood [9] we will show that for the normal distribution the region is a convex set and, therefore can be represented by boundary values of (μ, σ^2) such that *any* sampling distribution inside the boundary will be consistent with the current distribution.

Consider the sample $X_1, \ldots, X_n$ where all X_i are independent random variables with the normal distribution $(N(\mu, \sigma^2))$. Then $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, i.e., the sample mean and the sample variance, are random variables. It is well known that $\overline{X}$ has the normal distribution $N(\mu, \frac{\sigma^2}{n})$ (or, equivalently, $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$) while $(n-1)S^2/\sigma^2$ has the chi-square distribution with $n-1$ degrees of freedom [9].

The standard tables for $N(0,1)$ and χ^2_{n-1} provide numbers a, b, c such that for fixed p_1, p_2 the following equalities hold [1]:

$$P(-a < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < a) = p_1,$$
$$P(b < (n-1)S^2/\sigma^2 < c) = p_2$$

[7] The null hypothesis is a default position which, in this case, could be that the population mean is different from at least one of $\overline{X^{(1)}}, \overline{X^{(2)}}$.

[8] We deliberately leave out a precise definition of random set. For the purposes of this paper it is sufficient to think of a random set as of a random variable which takes on subsets of some space.

The crucial fact is that the two random variables are independent (see [9] for a proof) which implies that:

$$p_1 p_2 = P(-a < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < a, b < \frac{(n-1)S^2}{\sigma^2} < c) =$$
$$P(\overline{X} - a\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + a\frac{\sigma}{\sqrt{n}}, \frac{(n-1)S^2}{c} < \sigma^2 < \frac{(n-1)S^2}{b})$$

Thus, the $100(p_1)(p_2)\%$ confidence region for (μ, σ^2) takes the following form:

$$\mathcal{R}_{p_1,p_2}(\overline{X}, S) = \left\{(\mu, \sigma^2) : \ \overline{X} - \alpha\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + \alpha\frac{\sigma}{\sqrt{n}}, \right. \tag{2}$$
$$\left. \frac{(n-1)S^2}{\gamma} < \sigma^2 < \frac{(n-1)S^2}{\beta} \right\}$$

Figure 1 shows the joint confidence region $\mathcal{R}$ in the parameter space (μ, σ^2). Note that it is possible, although technically messy, to generalize the definition (2) to the case of several independent sampling distributions. The simultaneous confidence region for k samples $X^{(1)}, \ldots, X^{(k)}$ will be a region in the $2k$-dimensional parameter space which projections on each plane $(\mu^{(i)}, (\sigma^{(i)})^2)$ will look as (2). Then the notion of consistency of sampling distributions can be defined as follows (we limit the attention to two samples for clarity):

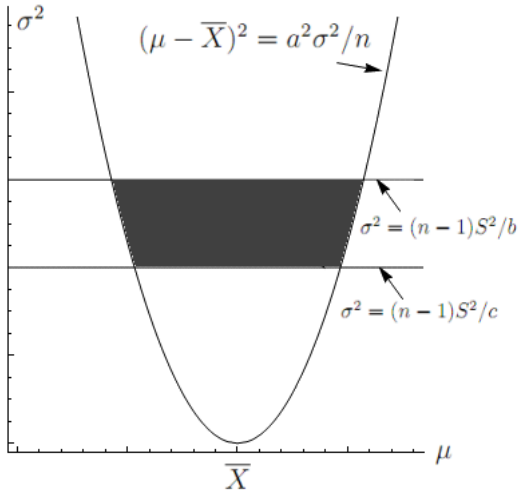

Fig. 1. Joint confidence region for (μ, σ^2)

Definition 3. *Let $Pr(X^{(1)}), Pr(X^{(2)})$ be distributions on two samples $X^{(1)}, X^{(2)}$ drawn independently from a population G. They are said to be consistent with confidence $100p\%$ if there exists a point (μ, σ^2) which belongs to both $\mathcal{R}_p(\overline{X^{(1)}}, S^{(1)})$ and $\mathcal{R}_p(\overline{X^{(2)}}, S^{(2)})$.*

Now we can return to the issue of approximating a continuous sampling distribution by a discrete set of points. Assume that the domain E of a continuous real-valued random variable X is a disjoint union of a finite number of intervals $U = \{(-\infty, r_1], (r_1, r_2], \ldots, (r_{l-1}, r_l], (r_l, +\infty)\}$. Then the *approximation* of the sampling distribution $Pr(X)$ with mean and variance $(\overline{X}, S^2)$ is the function $\hat{Pr}$ which maps each interval U_i to the following real-valued set:

$$\hat{Pr}(U_i; \overline{X}, S) = \{g(\mu, \sigma^2) | (\mu, \sigma^2) \in \mathcal{R}_{p_1,p_2}(\overline{X}, S)\} \tag{3}$$
$$g(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{U_i} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x$$

Now we are ready to define the notion of approximate consistency of sampling distribution with respect to a set of intervals U:

Definition 4. *Two sampling distributions $Pr(X^{(1)}), Pr(X^{(2)})$ are approximately consistent given a finite set of intervals U if $\hat{Pr}(U_i; \overline{X^{(1)}}, S^{(1)}) \cap \hat{Pr}(U_i; \overline{X^{(2)}}, S^{(2)})$ is non-empty for all $U_i \in U$.*

As with any approximation, the utility of approximations of sampling distributions depends on what conclusions they help to draw about the distributions themselves. Given that we are interested in the matter of consistency, it is important to understand the relationships between the notions of consistency and approximate consistency of sampling distributions. Fortunately, consistency implies approximate consistency regardless of partitioning of the real line:

Theorem 1. *If two sampling distributions $Pr(X^{(1)}), Pr(X^{(2)})$ are consistent, then they are approximately consistent for any choice of real-valued intervals.*

Proof. For the distribution $Pr(X^{(1)})$ a confidence region $\mathcal{R}_{p_1,p_2}(\overline{X^{(1)}}, S^{(1)})$ is connected (see Definition 2). The function $g(\mu, \sigma^2)$ (Definition 3) is continuous on it which implies that for any U_i, the set $\hat{Pr}(U_i; \overline{X^{(1)}}, S^{(1)})$ is a real-valued interval (l_1, u_1). Now consider a point $\mu_0, \sigma_0^2 \in \mathcal{R}_{p_1,p_2}(\overline{X^{(1)}}, S^{(1)}) \cap \mathcal{R}_{p_1,p_2}(\overline{X^{(2)}}, S^{(2)})$ which exists since the distributions are consistent. It follows that $l_1 < g(\mu_0, \sigma_0^2) < u_1$ (and analogously $l_2 < g(\mu_0, \sigma_0^2) < u_2$ for $\hat{Pr}(U_i; \overline{X^{(2)}}, S^{(2)})$), so $g(\mu_0, \sigma_0^2)$ is a common point for both approximations on U_i. As such the distributions are approximately consistent.

The following corollary from the above theorem is at heart of our method. As we demonstrate below, the inconsistency of approximations can be proved by logical reasoning in P-$\mathcal{SROIQ}$ (i.e., by solving the probabilistic satisfiability problem), which means that the result enables approximate reasoning about

sampling distributions in a purely logical way. Even though the power of such reasoning is currently limited to consistency checking, its integration with OWL/DL reasoning and the ability to use common, formally defined terminology for representation of statistical experiments is promising.

Corollary 1. *If sampling distributions* $Pr(X^{(1)}), Pr(X^{(2)})$ *are approximately inconsistent for some choice of real-valued intervals, then they are inconsistent.*

5 Example of Approximate Modelling

Now we present an example of approximate representation of sampling distributions in P-$\mathcal{SROIQ}$. The task is to take two results of statistical experiments aimed at investigating associations between alcohol consumption and the increased risk of breast cancer among postmenopausal women. Unfortunately it is common for medical papers to not explicitly present all parameters that characterize results of their statistical analyses. Typically, only the estimated mean and the confidence interval are presented while, for example, the kind of distribution is left to the reader to infer from other information. Due to that fact and because the approach above has only been developed for normal distributions, we illustrate it on an artificial example. The information given in the example is analogous to that given in medical literature, e.g. [10, 11], but is complete in the sense that all parameters and the type of sampling distributions are known.

Example 1. Consider two hypothetical papers which report results of independent studies of associations between alcohol consumption among postmenopausal women and their relative risk of developing breast cancer. According to study A the mean relative risk (RR) of ER+ breast cancer for women drinking $\geq$ 4g of ethanol a day is 1.8 and has variance of 0.5. Study B has reported that the mean RR of ER+ breast cancer for the same level of drinking is 2.2 (variance 0.7). The number of cases in the studies was 230 and 150 respectively.

We propose the following four step procedure for an approximate representation of statistical results, similar to those in the example above, in P-$\mathcal{SROIQ}$:

1. Preparing concepts The first step is to define the concepts/roles used to describe the distribution. In our case evidence concepts should describe categories of women with respect to specific risk factors, e.g. alcohol intake, while conclusion concepts describe groups of women stratified by risk increase. For instance, the concept expression $\mathtt{C} \equiv \mathtt{Woman} \sqcap \exists \mathtt{hasRiskFactor}.(\mathtt{Postmenopause} \sqcap \mathtt{ModerateConsumption})$ is used to model postmenopausal women with moderate level of alcohol intake.[9] On the other hand the expression:

$$D \equiv \mathtt{Woman} \sqcap \exists \mathtt{hasRisk}.(\mathtt{ModeratelyIncreasedRisk} \sqcap \exists \mathtt{riskOf}.\mathtt{ERPositiveBRC})$$

[9] The level of intake is a continuous variable which we also split onto categories `LimitedConsumption`, `ModerateConsumption` and `HeavyConsumption` which correspond to ≤ 4, $4 - 9.9$ and ≥ 10g of ethanol per day.

models women who are at moderately increased risk of developing ER-positive breast cancer. Using these expressions the modeler can specify the probability than a random women the class `C` also belong the risk group `D` as `(D|C)[l,u]`.

2. Determining parameters of sampling distributions (if required) Sometimes parameters of sampling distributions can be determined from other information. For example, knowing the kind of distribution, sample mean, sample size, confidence interval and the methodology of its estimation, one can calculate the sample variance.[10] In our case it is not needed as the distributions are normal and the parameters are known.

3. Choosing intervals Choice of intervals for an approximation of a continuous random variable is driven by balancing the quality of the approximation (i.e., how closely it models the continuous distribution) and the number of statements required. The latter has a direct impact on performance. For Example 1 we use three concepts `WomenAtWeakRisk`, `WomenAtModerateRisk` and `WomenAtHighRisk` which correspond to relative risk intervals of `(1,1.5]`, `(1.5, 3.0]` and `(3.0,`$+\infty$`)` respectively.

4. Computing the approximation The final (and the central) step is to compute probability intervals for the statements that approximate the continuous distribution. Each statement specifies the lower and upper probabilities that the continuous random variable X will fall into an interval U_i given that parameters of the distribution can vary within the confidence region (2). More formally, given the interval U_i, e.g. `(1,1.5]` for `WomenAtWeakRisk`, and the sampling distribution $(\overline{X}, S^2)$ the interval $[l_i, u_i]$ can be computed by solving the following non-linear optimization problem ():

$$\begin{aligned} &l_i \text{ (resp. } u_i) = \ min \text{ (resp. } max) \ g(\mu, \sigma^2) \text{ s.t.} \\ &(\mu, \sigma^2) \in \mathcal{R}_{p_1,p_2}(\overline{X}, S) \\ &g(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int\limits_{U_i} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x \end{aligned} \tag{4}$$

In other words, $[l_i, u_i]$=$[inf\ \hat{Pr}(U_i; \overline{X}, S), sup\ \hat{Pr}(U_i; \overline{X}, S)]$.

The last preparatory step is to calculate confidence regions according to (2). The 95% confidence regions for distributions $(\overline{X^{(1)}}, S^{(1)}), (\overline{X^{(2)}}, S^{(2)})$ in Example 1 (abbreviated as $R^{(1)}_{0.95}$ and $R^{(2)}_{0.95}$) are defined by the following inequalities:

$$R^{(1)}_{0.95} = \left\{ (\mu, \sigma^2) : \ 1.8 - \frac{2.241\sigma}{\sqrt{230}} < \mu < 1.8 + \frac{2.241\sigma}{\sqrt{230}}, 0.409 < \sigma^2 < 0.623 \right\}$$
$$R^{(2)}_{0.95} = \left\{ (\mu, \sigma^2) : \ 2.2 - \frac{2.241\sigma}{\sqrt{150}} < \mu < 2.2 + \frac{2.241\sigma}{\sqrt{150}}, 0.548 < \sigma^2 < 0.923 \right\}$$

[10] The variable $T = (\overline{X} - \mu)/(S/\sqrt{n})$ has the t-dustribution with $n-1$ degrees of freedom. Confidence interval is standardly computed as $[\overline{X} - a, \overline{X} + a]$ where $a = t_{\frac{1-\alpha}{2},n-1}\frac{S}{\sqrt{n}}$ ($t_{\frac{1-\alpha}{2},n-1}$ is the $\alpha-$percentile of the Student distribution). If the confidence interval and α are known, then S can be calculated.

Now the optimization problem (4) can be solved numerically[11] to obtain the following approximations for both sampling distributions:

$$inf\ \hat{Pr}((1,1.5];\overline{X^{(1)}},S^{(1)}) = 0.219 \qquad sup\ \hat{Pr}((1,1.5];\overline{X^{(1)}},S^{(1)}) = 0.298$$
$$inf\ \hat{Pr}((1.5,3.0];\overline{X^{(1)}},S^{(1)}) = 0.655 \qquad sup\ \hat{Pr}((1.5,3.0];\overline{X^{(1)}},S^{(1)}) = 0.878$$
$$inf\ \hat{Pr}((3.0,+\infty);\overline{X^{(1)}},S^{(1)}) = 0.239 \qquad sup\ \hat{Pr}((3.0,+\infty);\overline{X^{(1)}},S^{(1)}) = 0.586$$
$$inf\ \hat{Pr}((1,1.5];\overline{X^{(2)}},S^{(2)}) = 0.116 \qquad sup\ \hat{Pr}((1,1.5];\overline{X^{(2)}},S^{(2)}) = 0.224$$
$$inf\ \hat{Pr}((1.5,3.0];\overline{X^{(2)}},S^{(2)}) = 0.562 \qquad sup\ \hat{Pr}((1.5,3.0];\overline{X^{(2)}},S^{(2)}) = 0.769$$
$$inf\ \hat{Pr}((3.0,+\infty);\overline{X^{(2)}},S^{(2)}) = 0.189 \qquad sup\ \hat{Pr}((3.0,+\infty);\overline{X^{(2)}},S^{(2)}) = 0.568$$

So, for this example, the sampling distributions are approximately represented in P-$\mathcal{SROIQ}$ using the following conditional constraints.

$\{(\mathtt{W} \sqcap \exists \mathtt{hR}.(\mathtt{WeaklyIncreasedRisk} \sqcap \exists \mathtt{riskOf}.\mathtt{ERPositiveBRC})|C)[0.219, 0.298],$
$(\mathtt{W} \sqcap \exists \mathtt{hR}.(\mathtt{ModeratelyIncreasedRisk} \sqcap \exists \mathtt{riskOf}.\mathtt{ERPositiveBRC})|C)[0.655, 0.878],$
$(\mathtt{W} \sqcap \exists \mathtt{hR}.(\mathtt{StronglyIncreasedRisk} \sqcap \exists \mathtt{riskOf}.\mathtt{ERPositiveBRC})|C)[0.239, 0.586]\}$
and
$\{(\mathtt{W} \sqcap \exists \mathtt{hR}.(\mathtt{WeaklyIncreasedRisk} \sqcap \exists \mathtt{riskOf}.\mathtt{ERPositiveBRC})|C)[0.116, 0.224],$
$(\mathtt{W} \sqcap \exists \mathtt{hR}.(\mathtt{ModeratelyIncreasedRisk} \sqcap \exists \mathtt{riskOf}.\mathtt{ERPositiveBRC})|C)[0.562, 0.769],$
$(\mathtt{W} \sqcap \exists \mathtt{hR}.(\mathtt{StronglyIncreasedRisk} \sqcap \exists \mathtt{riskOf}.\mathtt{ERPositiveBRC})|C)[0.189, 0.568]\}$
where W and hR abbreviate Woman and hasRisk, respectively and
$\mathtt{C} \equiv \mathtt{W} \sqcap \exists \mathtt{hasRiskFactor}.(\mathtt{Postmenopause} \sqcap \mathtt{ModerateConsumption})$

Probabilistic consistency of the above set of statements can be proved by solving the probabilistic satisfiability problem (PSAT). Modern algorithms can decide PSAT for over a thousand of P-$\mathcal{SROIQ}$ statements (in addition to thousands of OWL axioms), so the method could be computationally practical [6].

6 Conclusion

Checking consistency of sampling distributions in P-$\mathcal{SROIQ}$ may well appear cumbersome and pointless given that the same task can be done in a much simpler way and without any logical reasoning, e.g. via testing or by analyzing confidence regions. However, our aim is *not* to reduce statistical testing to logical reasoning (that aim is indeed pointless). Our aim is to represent results of statistical experiments using *common, unambiguously defined logical vocabulary* and be able to reason about them. Even though probabilistic reasoning about statistical results is currently limited to approximate consistency checking, the

[11] We use Wolfram Mathematica for this purpose.

potential benefits are in combining it with reasoning about the classical knowledge. For example, the BCRA ontology contains a little taxonomy of breast cancers by hormone receptor status. This enables us to combine results of the studies which are of different levels of granularity. For instance, Sellers et al. [10] report associations between alcohol intake and ER(+/-) breast cancer risk, while Suzuki et al. [11] divide it further to ER(+/-)PR(+/-) risks. In that simple case non-logical reasoning about the reported results becomes much less straightforward, while studies can also distinguish histologic types of breast cancer (see [7]). In such complex situations reasoning about findings does involve reasoning about background knowledge, e.g. the taxonomy of breast cancers, so a combination of OWL and probabilistic reasoning is potentially beneficial.

References

1. Arnold, B.C., Shavelle, R.M.: Joint confidence sets for the mean and variance of a normal distribution. The American Statistician 52(2), 133–140 (1998)
2. Gail, M.H., Brinton, L.A., Byar, D.P., Corle, D.K., Green, S.B., Schairer, C., Mulvihill, J.J.: Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. Journal of the National Cancer Institute 81(25), 1879–1886 (1989)
3. Horrocks, I., Kutz, O., Sattler, U.: The even more irresistible $\mathcal{SROIQ}$. In: Knowledge Representation and Reasoning. pp. 57–67 (2006)
4. Key, J., Hodgson, S., Omar, R.Z., Jensen, T.K., Thompson, S.G., Boobis, A.R., Davies, D.S., Elliott, P.: Meta-analysis of studies of alcohol and breast cancer with consideration of the methodological issues. Cancer Causes Control 17, 759–770 (2006)
5. Klinov, P., Parsia, B.: Probabilistic modeling and OWL: A user oriented introduction into P-$\mathcal{SHIQ}$(D). In: OWL: Experiences and Directions (2008), `http://www.webont.org/owled/2008/papers/owled2008eu_submission_32.pdf`
6. Klinov, P., Parsia, B.: A hybrid method for probabilistic satisfiability. In: CADE. pp. 354–368 (2011)
7. Lew, J.Q., Freedman, N.D., Leitzmann, M.F., Brinton, L.A., Hoover, R.N., Hollenbeck, A.R., Schatzkin, A., Park, Y.: Alcohol and risk of breast cancer by histologic type and hormone receptor status in postmenopausal women the nih-aarp diet and health study. American Journal of Epidemiology 170(3), 308–317 (2009)
8. Lukasiewicz, T.: Expressive probabilistic description logics. Artificial Intelligence 172(6-7), 852–883 (2008)
9. Mood, A.M.: Introduction to the Theory of Statistics. McGraw-Hill (1950)
10. Sellers, T.A., Vierkant, R.A., Cerhan, J.R., Gapstur, S.M., Vachon, C.M., Olson, J.E., Pankratz, V.S., Kushi, L.H.: Interaction of dietary folate intake, alcohol, and risk of hormone receptor-defined breast cancer in a prospective study of postmenopausal women. Cancer Epidemiology, Biomarkers and Prevention 11, 1104–1107 (2002)
11. Suzuki, R., Ye, W., Rylander-Rudqvist, T., Saji, S., Colditz, G.A., Wolk, A.: Alcohol and postmenopausal breast cancer risk defined by estrogen and progesterone receptor status: A prospective cohort study. Journal of the National Cancer Institute 97(21), 1601–1608 (2005)