Formalizing Statistical Beliefs in Hypothesis Testing Using Program Logic

Yusuke Kawamoto[1,2] , **Tetsuya Sato**[3] , **Kohei Suenaga**[4]

[1]AIST, Japan
[2]PRESTO, JST, Japan
[3]Tokyo Institute of Technology, Japan
[4]Kyoto University, Japan

Abstract

We propose a new approach to formally describing the requirement for statistical inference and checking whether the statistical method is appropriately used in a program. Specifically, we define belief Hoare logic (BHL) for formalizing and reasoning about the statistical beliefs acquired via hypothesis testing. This logic is equipped with axiom schemas for hypothesis tests and rules for multiple tests that can be instantiated to a variety of concrete tests. To the best of our knowledge, this is the first attempt to introduce a program logic with epistemic modal operators that can specify the preconditions for hypothesis tests to be applied appropriately.

1 Introduction

Statistical inference has been widely used to derive and justify scientific knowledge in a variety of academic disciplines, from natural sciences to social sciences. This has significantly increased the importance of statistics, but also brought concerns about the inappropriate procedure and the incorrect interpretation of statistics in scientific research. In fact, previous studies have pointed out that many research articles in biomedical science contain severe errors in the application and interpretation of statistical inference (Lang and Altman 2014). Furthermore, large proportions of these errors have been reported for basic statistical methods, possibly performed by researchers who can use only elementary techniques. For example, the concept of *statistical significance*, evaluated using *p-values*, has been commonly misused and misinterpreted (Wasserstein and Lazar 2016).

One of the main issues behind these human errors is that the logical aspects of statistical inference are described informally or implicitly using natural languages, and handled manually by analysts who may not fully understand the statistical methods. In particular, this makes them overlook some assumptions necessary for statistical methods, hence choosing inappropriate methods. Nevertheless, to our knowledge, no prior work on formal methods has specified the preconditions for statistical inference or verified the choice of statistical techniques.

In this paper, we propose a method for formalizing and reasoning about statistical inference using symbolic logic. Specifically, we introduce *belief Hoare logic* (*BHL*) to formalize the statistical beliefs acquired via hypothesis tests, and to prevent errors in the choice of hypothesis tests by describing their preconditions explicitly. This is the first step to build a framework for formalizing and verifying the validity of empirical science on the basis of formal methods.

Contributions. Our main contributions are as follows:

- We propose a new approach to formalizing and reasoning about statistical inference in a program. In particular, this approach formalizes and checks the requirement for statistical methods to be used appropriately.
- We define an epistemic logic to express statistical beliefs obtained by hypothesis tests on datasets. Specifically, we formalize a statistical belief on a hypothesis φ as the knowledge that either φ holds or the sampled dataset is unluckily far from the population. Then we introduce a Kripke semantics to define the interpretation of the logic.
- Using this epistemic logic, we construct belief Hoare logic (BHL) for formalizing and reasoning about the statistical inference based on hypothesis testing. Specifically, we define axiom schemas and rules for hypothesis tests that can be instantiated to a variety of concrete tests. In particular, BHL does not rely on a specific philosophy of statistics but can deal with both the frequentist and the Bayesian statistics by introducing corresponding axioms.
- We show that BHL is useful to reason about practical issues concerning statistical inference, such as the multiple comparison problem and p-value hacking.
- We provide a whole picture of the justification of statistical belief acquired via hypothesis tests inside and outside BHL. For instance, we discuss the empirical conditions and the epistemic aspects of statistical inference.

To the best of our knowledge, this is the first attempt to introduce a program logic that can specify the preconditions for hypothesis tests to be applied appropriately.

Related Work. The *Hoare logic* (Winskel 1993) is one of the program logic for an imperative programming language. This program logic is then extended and adapted so that it can handle various types of programs and assertions, including heap-manipulating programs (Reynolds 2002), hybrid systems (Suenaga and Hasuo 2011), and probabilistic programs (den Hartog and de Vink 2002). Atkinson and Carbin propose an extension of Hoare logic with epistemic assertions (Atkinson and Carbin 2020). In their work, an epistemic assertion is used to reason about the belief of a program about a partially observable environment, whereas their logic does not deal with a statistical belief arising from statistical tests conducted in a program. To the best of our knowledge, ours is the first program logic that formalizes the concept of statistical beliefs in hypothesis testing.

Epistemic logic (von Wright 1951) is a branch of logic for reasoning about knowledge and belief (Fagin et al. 1995a; Halpern 2003), and is used to specify and verify a variety of knowledge properties in systems, e.g., authentication (Burrows, Abadi, and Needham 1990) and anonymity (Syverson and Stubblebine 1999; Garcia et al. 2005). Many previous works incorporate certain notions of degrees of belief and confidence (Huber and Schmidt-Petri 2008), but not in the sense of the statistical significance in hypothesis testing.

The first attempt to express statistical properties using modal logic is the work on statistical epistemic logic (StatEL) (Kawamoto 2019; Kawamoto 2020). They introduce a belief modality weaker than S5, and interpret it in a Kripke model with an accessibility relation defined in terms of a statistical distance between possible worlds. Unlike this work, however, StatEL cannot describe the procedures of statistical methods or reason about the appropriateness of inference.

From a broader perspective, many studies formalize and reason about programs based on knowledge (Fagin et al. 1995b) and beliefs (Laverny and Lang 2005), including belief updates in programs. For example, the situation calculus is extended to deal with the probabilistic degrees of beliefs in programs with noisy acting and sensing (Belle and Levesque 2015). However, no prior work has studied belief-based programs involving statistical hypothesis testing.

2 Preliminaries

In this section, we introduce notations used in this paper and recall background on statistical hypothesis testing.

Let $\mathbb{N}$, $\mathbb{R}$, $\mathbb{R}_{\geq 0}$ be the sets of non-negative integers, real numbers, and non-negative real numbers, respectively. Let $[0, 1]$ be the set of non-negative real numbers not greater than 1. We denote the dimension of a vector x by $\mathsf{size}(x)$, and the *set of all probability distributions* over a set $\mathcal{S}$ by $\mathbb{D}\mathcal{S}$.

Statistical Hypothesis Testing. *Statistical hypothesis testing* is a method of statistical inference about an unknown *population* x (the collection of items of interest) on the basis of a *dataset* y sampled from x. In a hypothesis test, an *alternative hypothesis* φ_1 is a proposition that we wish to prove about the population x, and a *null hypothesis* φ_0 is a proposition that contradicts φ_1. The goal of the hypothesis test is to determine whether we *accept* the alternative hypothesis φ_1 by *rejecting* the null hypothesis φ_0.

In a hypothesis test, we calculate a *test statistic* $t(y)$ from a dataset y, and see whether the $t(y)$ value contradicts the assumption that the null hypothesis φ_0 is true. Specifically, we calculate the *p-value*, showing the degree of likeliness of obtaining $t(y)$ when the null hypothesis φ_0 is true. If the p-value is smaller than a threshold (e.g., 0.05), we regard the dataset y is unlikely to be sampled from the population satisfying the null hypothesis φ_0, hence we reject φ_0 and accept the alternative hypothesis φ_1.

A hypothesis test is based on a *statistical model* $P(\xi, \theta)$ with unknown true parameters ξ, known parameters θ, and (assumed) probability distributions of the parameters in ξ.

Example 1 (Z-test for two population means)**.** *As an illustrating example, we present the two-tailed Z-test for means of two populations with a known and equal variance. We introduce its statistical model as two normal distributions $N(\mu_{\mathsf{ppl1}}, \sigma^2)$ and $N(\mu_{\mathsf{ppl2}}, \sigma^2)$ with a known variance σ^2 and unknown true means $\mu_{\mathsf{ppl1}}, \mu_{\mathsf{ppl2}}$. Let y_1 and y_2 be two given datasets where each data value was sampled independently from $N(\mu_{\mathsf{ppl1}}, \sigma^2)$ and $N(\mu_{\mathsf{ppl2}}, \sigma^2)$, respectively.*

*In the Z-test, we set the alternative hypothesis $\varphi_1 \stackrel{\text{def}}{=} (\mu_{\mathsf{ppl1}} \neq \mu_{\mathsf{ppl2}})$ and the null hypothesis $\varphi_0 \stackrel{\text{def}}{=} (\mu_{\mathsf{ppl1}} = \mu_{\mathsf{ppl2}})$. We calculate the Z-test statistic $t(y_1, y_2) = \frac{\mathsf{mean}(y_1) - \mathsf{mean}(y_2)}{\sigma\sqrt{1/\mathsf{size}(y_1) + 1/\mathsf{size}(y_2)}}$ where for $b = 1, 2$, $\mathsf{size}(y_b)$ is the sample size of y_b and $\mathsf{mean}(y_b)$ is the mean of all data in y_b. Then the p-*value *is defined by:*

$$\Pr_{(d_1,d_2)\sim N(\mu_{\mathsf{ppl1}},\sigma^2)\times N(\mu_{\mathsf{ppl1}},\sigma^2)}[|t(d_1, d_2)| > |t(y_1, y_2)|]$$

under the null hypothesis φ_0. When the p-value is small enough, the datasets y_1, y_2 are unlikely to be sampled from the same distribution, i.e., the null hypothesis $\mu_{\mathsf{ppl1}} = \mu_{\mathsf{ppl2}}$ is unlikely to hold. Hence, in the Z-test, if the p-value is smaller than a certain threshold (e.g., 0.05), we reject the null hypothesis φ_0 and accept the alternative hypothesis φ_1.

When we have prior knowledge of $\mu_{\mathsf{ppl1}} \geq \mu_{\mathsf{ppl2}}$ (resp. $\mu_{\mathsf{ppl1}} \leq \mu_{\mathsf{ppl2}}$), then we apply the upper-tailed *(resp.* lower-tailed*) Z-test with the alternative hypothesis $\mu_{\mathsf{ppl1}} > \mu_{\mathsf{ppl2}}$ (resp. $\mu_{\mathsf{ppl1}} < \mu_{\mathsf{ppl2}}$), and with the p-value $\Pr[t(d_1, d_2) > t(y_1, y_2)]$ (resp. $\Pr[t(d_1, d_2) < t(y_1, y_2)]$).*

3 Illustrating Example

Throughout the paper, we use the following simple illustrating example to explain the basic ideas on our framework.

Example 2 (Comparison tests on drugs)**.** *Let us consider three drugs 1, 2, 3 that may decrease blood pressure. To compare the efficacy of these drugs, we perform experiments and obtain a set y_i of the reduced values of blood pressure after taking drug i. Then we apply hypothesis tests on the dataset $y = (y_1, y_2, y_3)$. Let x_i be the true population from which the data values in y_i are sampled.*

Suppose that drug 1 is composed of drugs 2 and 3, and we want to know whether drug 1 has better efficacy than both *drugs 2 and 3. Then we take the following procedure:*

- *We first compare drugs 1 and 2 concerning the average decreases in blood pressure. We apply a two-tailed Z-test to see whether the means of the true populations x_1 and x_2 are different, i.e., $\mathsf{mean}(x_1) \neq \mathsf{mean}(x_2)$. In this test, the alternative hypothesis is $\varphi_{12} \stackrel{\text{def}}{=} (\mathsf{mean}(x_1) \neq \mathsf{mean}(x_2))$, and the null hypothesis is $\neg\varphi_{12} \equiv (\mathsf{mean}(x_1) = \mathsf{mean}(x_2))$.*
- *Let α_{ij} be the p-value when only comparing drugs i and j.*
- *If $\alpha_{12} \geq 0.05$, the Z-test does not reject the null hypothesis $\neg\varphi_{12}$ and concludes that the efficacy of drugs 1 and 2 may be the same. Then we are not interested in drug 1 any more, and skip the comparison with drug 3.*

- *If* $\alpha_{12} < 0.05$*, the* Z*-test rejects the null hypothesis* $\neg\varphi_{12}$ *and concludes that the alternative hypothesis* φ_{12} *is true. Then we apply another* Z*-test to check whether the alternative hypothesis* $\varphi_{13} \stackrel{\text{def}}{=} (\mathsf{mean}(x_1) \neq \mathsf{mean}(x_3))$ *is true.*
- *Finally, we calculate the* p*-value of the combined tests, with the* conjunctive *alternative hypothesis* $\varphi_{12} \wedge \varphi_{13}$.

Note that these Z*-tests assume that the distribution of each population* x_i *is a normal distribution with a variance* σ^2.

Overview of the Framework. In our framework, we describe a procedure of statistical tests as a program using a programming language (Section 6); in Example 2, we denote the Z-test program comparing drugs i with j by C_{ij}, and the whole procedure by:

$$C_{\text{drug}} \stackrel{\text{def}}{=} C_{12}; \texttt{if } \alpha_{12} < 0.05 \texttt{ then } C_{13} \texttt{ else skip} \quad (1)$$

Then we use an assertion logic (Section 5) to describe the requirement for the hypothesis tests as a *precondition* formula. In Example 2, the precondition is given by:

$$\psi_{\mathsf{pre}} \stackrel{\text{def}}{=} \bigwedge_{i=1,2,3} \psi_i \wedge \bigwedge_{\substack{(i,j)=(1,2),(1,3),\\(2,1),(3,1)}} \mathbf{P}(\mathsf{mean}(x_i) < \mathsf{mean}(x_j))$$

$$\text{where} \quad \psi_i \stackrel{\text{def}}{=} \big(x_i \approx N(\mu_i, \sigma) \wedge y_i \underset{n_i}{\leftsquigarrow} x_i\big).$$

In this formula, ψ_i represents that the true population x_i follows a normal distribution $N(\mu_i, \sigma)$ (with an unknown true mean μ_i), and that a set y_i of n_i data is sampled from x_i. $\mathbf{P}(\mathsf{mean}(x_1) < \mathsf{mean}(x_2))$ and $\mathbf{P}(\mathsf{mean}(x_1) > \mathsf{mean}(x_2))$ represent that both the lower-tail and upper-tail are possible, hence the test C_{12} should be two-tailed. Remark that our assertion logic never deals with quantifiers over variables.

The statistical belief we want to acquire is specified as a *postcondition* formula. In Example 2, the postcondition is:

$$\varphi_{\mathsf{post}} \stackrel{\text{def}}{=} \big(\mathbf{K}_y^{\leq 0.05} \varphi_{12} \rightarrow \mathbf{K}_y^{\leq \min(\alpha_{12}, \alpha_{13})} (\varphi_{12} \wedge \varphi_{13})\big). \quad (2)$$

Intuitively, by testing on a dataset y, when we believe φ_{12} with a p-value $\alpha \leq 0.05$, we believe the combined hypothesis $\varphi_{12} \wedge \varphi_{13}$ with a p-value at most $\min(\alpha_{12}, \alpha_{13})$.

Finally, we combine all the above and describe the whole statistical inference as a *judgment*. In Example 2, we write:

$$\Gamma \vdash \{\psi_{\mathsf{pre}}\}\ C_{\text{drug}}\ \{\varphi_{\mathsf{post}}\} \quad (3)$$

By proving this judgment using rules in BHL (Section 7), we conclude that the statistical inference is appropriate.

We remark that the p-value can be larger for a different purpose of testing. Suppose that in Example 2, drug 1 was a new drug and we wanted to find out it had better efficacy than *at least one* of drugs 2 and 3. Then the procedure is:

$$C_{\text{multi}} \stackrel{\text{def}}{=} C_{12} \,\|\, C_{13}, \quad (4)$$

and the alternative hypothesis is $\varphi_{12} \vee \varphi_{13}$ with a p-value larger than α_{12} and α_{13} (at most $\alpha_{12} + \alpha_{13}$). This is the *multiple comparisons problem* (Bretz, Hothorn, and Westfall 2010), arising when the combined alternative hypothesis is in *disjunctive* form. We explain more details in Section 7.

4 Model

In this section, we introduce a Kripke model for describing statistical properties and formally define hypothesis tests.

4.1 Variables, Data, and Actions

We introduce a set Var of variables comprised of two disjoint sets of *invisible variables* and of *observable variables*: $\mathsf{Var} = \mathsf{Var}_{\mathsf{inv}} \cup \mathsf{Var}_{\mathsf{obs}}$. We can directly observe the values of the latter, but not those of the former. We use x as an invisible variable denoting a population, and y as an observable variable denoting a dataset sampled from the population.

We write $\mathcal{O}$ for the set of all data values we deal with, including the Boolean values, integers, real numbers, and lists of data. A *dataset* is a list of lists of data. In particular, we deal with a list of real vectors as a dataset. Then the vectors range over $\mathcal{X} = \mathbb{R}^l$ for an $l \in \mathbb{N}$, a distribution over the population has type $\mathbb{D}\mathcal{X}$, and a dataset has type $\mathsf{list}\,\mathcal{X}$.

We write $d \sim D^n$ for the sampling of a set d of n data from a population D where all these data are independent and identically distributed (i.i.d.). Let Smpl be a set of i.i.d. samplings of datasets from populations (e.g., $d \sim D^n$), and Cmd be a set of program commands[1] (e.g., $v := e$ and `skip`). Then we define an *action* as a sampling of a dataset or a program command; i.e., $\mathsf{Act} = \mathsf{Smpl} \cup \mathsf{Cmd}$.

4.2 States and Possible Worlds

A *state* is a pair (m, a) consisting of the current assignment $m : \mathsf{Var} \to \mathcal{O}$ of data values to variables, and the action $a \in \mathsf{Act}$ that has been executed in the last transition.

A *possible world* w is a sequence of the states $(w[0], w[1], \ldots, w[k-1])$ where $w[i]$ represents the i-th state. The length k is denoted by $\mathsf{len}(w)$. $w[0]$ and $w[k-1]$ are called the *initial state* and the *current state*, respectively. Since a possible world records all updates of data values, it can be used to model the updates of knowledge and beliefs as with previous works on epistemic logic (Fagin et al. 1995a).

The *observation* of a state $w[i] = (m, a)$ is defined by $obs(w[i]) = (m_{\mathsf{o}}, a)$ with an assignment $m_{\mathsf{o}} : \mathsf{Var}_{\mathsf{obs}} \to \mathcal{O}$ such that $m_{\mathsf{o}}(v) = m(v)$ for all $v \in \mathsf{Var}_{\mathsf{obs}}$. The *observation* of a possible world w is defined by $obs(w) = (obs(w[0]), \ldots, obs(w[k-1]))$. We abuse notations and denote by $w : \mathsf{Var} \to \mathcal{O}$ the assignment of data values to variables in the current state of a possible world w.

4.3 Kripke Model

We introduce a *Kripke model with labeled transitions* where two kinds of relations $\xrightarrow{a}$ and $\mathcal{R}$ may relate possible worlds.

A *transition relation* $w \xrightarrow{a} w'$ represents a transition from a world w to another w' by performing an action a. An *observability relation* $w\mathcal{R}w'$ represents that two possible worlds w and w' have the same observation, i.e., $obs(w) = obs(w')$. In Section 5, this relation is used to model the knowledge in the conventional Hintikka-style.

Definition 1 (Kripke model)**.** We define a *Kripke model* as a tuple $\mathfrak{M} = (\mathcal{W}, (\xrightarrow{a})_{a \in \mathsf{Act}}, \mathcal{R}, (V_w)_{w \in \mathcal{W}})$ consisting of:

[1] In Section 6, we instantiate Cmd with a concrete example of commands used in a simple programming language.

- a non-empty set $\mathcal{W}$ of possible worlds;
- for each $a \in \mathsf{Act}$, a transition relation $\xrightarrow{a} \subseteq \mathcal{W} \times \mathcal{W}$;
- an observability relation $\mathcal{R} = \{(w, w') \in \mathcal{W} \times \mathcal{W} \mid obs(w) = obs(w')\}$;
- for each $w \in \mathcal{W}$, a valuation $V_w : \mathsf{Pred} \to \mathcal{P}(\mathcal{O}^k)$ that maps a k-ary predicate to a set of k-tuples of data;

We assume $w \xrightarrow{a} w'$ implies $w'[\mathsf{len}(w') - 1] = (m, a)$ for some m. We also assume that each world in a model has the same sets $\mathsf{Var}_{\mathsf{inv}}$ and $\mathsf{Var}_{\mathsf{obs}}$ of variables.

4.4 Formulation of Hypothesis Testing

Next, we formalize hypothesis tests. We consider a *basic test type* $s \in \{\mathsf{L}, \mathsf{U}, \mathsf{T}\}$ each representing a *lower-tailed*, *upper-tailed*, and *two-tailed* test. A *hypothesis test* is a tuple $A^{(s)}_{\varphi_0} = (\varphi_0, t, D_{t,\varphi_0}, \preccurlyeq^{(s)}, P(\xi, \theta))$ consisting of:

- φ_0 is an assertion, called a null hypothesis;
- t is a function that maps a dataset $d \in \mathsf{list}\,\mathcal{X}$ to its test statistic $t(d)$, usually with $\mathsf{range}(t) = \mathbb{R}^k$ for a $k \geq 1$;
- $D_{t,\varphi_0} \in \mathbb{D}(\mathsf{range}(t))$ is a probability distribution of the test statistic when the null hypothesis φ_0 is true;
- $\preccurlyeq^{(s)}_t \in \mathsf{range}(t) \times \mathsf{range}(t)$ is a *likeliness relation* where for a test type s and for values d and d' of the test statistic, $d \preccurlyeq^{(s)} d'$ represents that d is at most as likely as d'. For brevity, we often omit t and ${}^{(s)}$ to write $\preccurlyeq^{(s)}$ and $\preccurlyeq$.
- $P(\xi, \theta)$ is a statistical model with unknown parameters ξ and known parameters θ that characterizes the population.

Then we define a *hypothesis testing* over a set Φ of possible hypotheses by $A^{(s)} = (A^{(s)}_{\varphi})_{\varphi \in \Phi}$. We denote by $\mathcal{A}$ a finite set of hypothesis testings we consider. For brevity, we sometimes omit ${}^{(s)}$ to write A_φ and A. We also often omit the statistical model $P(\xi, \theta)$ from the description of $A^{(s)}_{\varphi_0}$.

Example 3 (The likeliness relation for Z-test)**.** *The two-tailed Z-test for two populations (Example 1) is denoted by $A_{\varphi_0} = (\varphi_0, t, N(0,1), \preccurlyeq^{(\mathsf{T})}, N(\mu_{\mathsf{ppl1}}, \sigma^2) \times N(\mu_{\mathsf{ppl2}}, \sigma^2))$. The likeliness relation $d \preccurlyeq^{(\mathsf{T})} d'$ expresses $|d| \geq |d'|$. When the null hypothesis φ_0 is true, the test statistic $t(y_1, y_2)$ follows the standard normal distribution $N(0,1)$, hence*

$$\Pr[t(y_1, y_2) \preccurlyeq^{(\mathsf{T})} 1.96] = \Pr[|t(y_1, y_2)| \geq 1.96] = 0.05.$$

In contrast, for the upper-tailed (lower-tailed) test, with alternative hypothesis $\varphi_{\mathsf{U}} \stackrel{\text{def}}{=} (\mu_{\mathsf{ppl1}} > \mu_{\mathsf{ppl2}})$ (resp. $\varphi_{\mathsf{L}} \stackrel{\text{def}}{=} (\mu_{\mathsf{ppl1}} < \mu_{\mathsf{ppl2}})$), the likeliness relation $d \preccurlyeq^{(\mathsf{U})} d'$ (resp. $d \preccurlyeq^{(\mathsf{L})} d'$) is defined by $d \geq d'$ (resp. $d \leq d'$).

Next we define the *disjunctive combination* of two hypothesis tests $A_{\varphi_b} = (\varphi_b, t_b, D_{t_b,\varphi_b}, \preccurlyeq^{(s_b)}_{t_b}, P_b)$ for $b = 1, 2$ by $A_{\varphi_1 \vee \varphi_2} = (\varphi_1 \vee \varphi_2, t, D_{t,(\varphi_1,\varphi_2)}, \preccurlyeq^{(s_1,s_2)}_t, P)$ where $t(y_1, y_2) = (t_1(y_1), t_2(y_2))$, $D_{t,(\varphi_1,\varphi_2)} = D_{t_1,\varphi_1} \times D_{t_2,\varphi_2}$, $(d_1, d_2) \preccurlyeq^{(s_1,s_2)}_t (d'_1, d'_2)$ iff either $d_1 \preccurlyeq^{(s_1)}_{t_1} d'_1$ or $d_2 \preccurlyeq^{(s_2)}_{t_2} d'_2$, and $P = P_1 \times P_2$. Similarly, we define the *conjunctive combination* by $A_{\varphi_1 \wedge \varphi_2} = (\varphi_1 \wedge \varphi_2, t, D_{t,(\varphi_1,\varphi_2)}, \preccurlyeq^{(s_1,s_2)}_t, P)$ where $(d_1, d_2) \preccurlyeq^{(s_1,s_2)}_t (d'_1, d'_2)$ iff $d_1 \preccurlyeq^{(s_1)}_{t_1} d'_1$ and $d_2 \preccurlyeq^{(s_2)}_{t_2} d'_2$.

5 Assertion Language

Next we define an *assertion logic* that can express epistemic properties including knowledge and statistical beliefs.

5.1 Syntax of the Assertion Logic

We introduce two kinds of epistemic modality $\mathbf{K}$ and $\mathbf{K}^{<\epsilon}_{y,A}$. Intuitively, a *knowledge* $\mathbf{K}\varphi$ expresses that we know φ, and this has been studied in a lot of previous work on epistemic logic. In contrast, a *statistical belief* $\mathbf{K}^{<\epsilon}_{y,A}\,\varphi$ expresses that we believe a hypothesis φ based on a statistical test A on an observed dataset y with a certain error level (p-value) at most ϵ. We formalize this as the knowledge that either the hypothesis φ holds or the observed dataset y is unluckily far from the population (from which y is sampled).

Formally, for a set Var of variables and a set Pred of predicates, the set Fml of formulas are defined by:

$$\varphi ::= \eta(x_1, \ldots, x_n) \mid \neg\varphi \mid \varphi \vee \varphi \mid \mathbf{K}\varphi$$

where $\eta \in \mathsf{Pred}$ and $x_1, \ldots, x_n \in \mathsf{Var}$. The formulas have no quantifiers over variables. We denote the set of all observable (resp. invisible) variables occurring in a formula φ by $\mathsf{fv}^{\mathsf{obs}}(\varphi)$ (resp. $\mathsf{fv}^{\mathsf{inv}}(\varphi)$). Let $\mathsf{fv}(\varphi) = \mathsf{fv}^{\mathsf{obs}}(\varphi) \cup \mathsf{fv}^{\mathsf{inv}}(\varphi)$.

As syntax sugar, we use *conjunction* $\wedge$, *implication* $\rightarrow$, and *epistemic possibility* $\mathbf{P}$, defined as usual by: $\varphi_0 \wedge \varphi_1 \stackrel{\text{def}}{=} \neg(\neg\varphi_0 \vee \neg\varphi_1)$, $\varphi_0 \rightarrow \varphi_1 \stackrel{\text{def}}{=} \neg\varphi_0 \vee \varphi_1$, and $\mathbf{P}\varphi \stackrel{\text{def}}{=} \neg\mathbf{K}\neg\varphi$.

We introduce three predicates for statistical inference:

- $x \approx P$ represents that a population x follows a probability distribution that we assume in a statistical model P.
- For $x = (x_1, \ldots, x_k)$ and $n = (n_1, \ldots, n_k)$, $y \overset{}{\underset{n}{\leftsquigarrow}} x$ represents that for each $i = 1, 2, \ldots, k$, a dataset y_i is obtained by sampling n_i data from the population x_i.
- For $\bowtie \in \{=, \leq, \geq, <, >\}$ and $\epsilon \in [0, 1]$, $\tau^{\bowtie\epsilon}_A(y)$ represents that the observation of a dataset y is unlikely to occur (with exception $\bowtie \epsilon$) according to a hypothesis test A, and the dataset y is used in no other hypothesis tests.

We introduce a set $\mathsf{Pred}_{\mathsf{G}}$ of *global predicate*, whose interpretations are identical in every possible worlds. We assume $\approx \in \mathsf{Pred}_{\mathsf{G}}$ and $\mathsf{Pred} = \mathsf{Pred}_{\mathsf{G}} \cup \{\tau^{\bowtie\epsilon}_A, \leftsquigarrow\}$.

As syntax sugar, we introduce the *statistical belief modality* $\mathbf{K}^{\bowtie\epsilon}_{y,A}$ such that for a formula φ representing a hypothesis,

$$\mathbf{K}^{\bowtie\epsilon}_{y,A}\,\varphi \stackrel{\text{def}}{=} \mathbf{K}(\varphi \vee \tau^{\bowtie\epsilon}_{A_{\neg\varphi}}(y))$$

where $A_{\neg\varphi}$ is a test with a null hypothesis $\neg\varphi$. Then we define the *statistical possibility* $\mathbf{P}^{\bowtie\epsilon}_{y,A}$ by $\mathbf{P}^{\bowtie\epsilon}_{y,A}\,\varphi \stackrel{\text{def}}{=} \neg\,\mathbf{K}^{\bowtie\epsilon}_{y,A}\,\neg\varphi$. For brevity, we write $\mathbf{K}^{\epsilon}_{y,A}$ instead of $\mathbf{K}^{=\epsilon}_{y,A}$. For a finite set $\mathcal{A}$ of hypothesis testings, we write $\mathbf{K}^{\bowtie\epsilon}_{y}\,\varphi \stackrel{\text{def}}{=} \bigvee_{A \in \mathcal{A}} \mathbf{K}^{\bowtie\epsilon}_{y,A}\,\varphi$ and $\mathbf{P}^{\bowtie\epsilon}_{y}\,\varphi \stackrel{\text{def}}{=} \bigvee_{A \in \mathcal{A}} \mathbf{P}^{\bowtie\epsilon}_{y,A}\,\varphi$. A formula ψ is $\tau^{\bowtie}$*-free* if $\tau^{\bowtie\epsilon}_A$, $\mathbf{K}^{\bowtie\epsilon}_{y,A}$, $\mathbf{P}^{\bowtie\epsilon}_{y,A}$, $\mathbf{K}^{\bowtie\epsilon}_{y}$, $\mathbf{P}^{\bowtie\epsilon}_{y}$ do not occur in ψ.

5.2 Semantics of the Assertion Logic

In this section, we define semantics for the assertion logic.

We define the interpretation of formulas in a world w in a Kripke model $\mathfrak{M}$ (Definition 1) by:

$$\begin{aligned}
\mathfrak{M}, w \models \eta(x_1, \ldots, x_k) & \text{ iff } (w(x_1), \ldots, w(x_k)) \in V_w(\eta) \\
\mathfrak{M}, w \models \neg\varphi & \text{ iff } \mathfrak{M}, w \not\models \varphi \\
\mathfrak{M}, w \models \varphi \vee \varphi' & \text{ iff either } \mathfrak{M}, w \models \varphi \text{ or } \mathfrak{M}, w \models \varphi' \\
\mathfrak{M}, w \models \mathbf{K}\varphi & \text{ iff for all } w' \in \mathcal{W},\ (w, w') \in \mathcal{R} \\
& \quad \text{implies } \mathfrak{M}, w' \models \varphi.
\end{aligned}$$

$\mathfrak{M}$ is sometimes omitted when it is clear from the context.

Next we define the interpretation of predicates. Each global predicate has the same interpretation in all worlds; i.e., for any $\eta \in \mathsf{Pred}_\mathsf{G}$ and $w, w' \in \mathcal{W}$, $V_w(\eta) = V_{w'}(\eta)$. Let $A_\varphi = (\varphi, t, D_{t,\varphi}, \preccurlyeq^{(s)})$ be a hypothesis test. Recall that the distribution over the population has type $\mathbb{D}\mathcal{X}$, and that $\preccurlyeq^{(s)}$ is the likeliness relation (Section 4.4). In a world w, we interpret the predicates by:

$$V_w(\approx) = \{(X, D) \in \mathbb{D}\mathcal{X} \times \mathbb{D}\mathcal{X} \mid X = D\}$$

$$V_w(\leftsquigarrow) = \left\{(d, D, n) \in (\mathsf{list}\,\mathcal{X}) \times \mathbb{D}\mathcal{X} \times \mathbb{N} \,\middle|\, \begin{array}{l}\text{There is an } i \in \mathbb{N} \\ \text{s.t. } w[i] \xrightarrow{d \sim D^n} w[i+1]\end{array}\right\}$$

$$V_w(\tau^{<\epsilon}_{A_\varphi}) = \left\{o \in \mathcal{O} \,\middle|\, \begin{array}{l}\Pr_{d \sim D_{t,\varphi}}[\, d \preccurlyeq^{(s)} t(o)\,] < \epsilon \text{ and} \\ w \text{ has no transition where } o \text{ is used} \\ \text{in other hypothesis tests than } A_\varphi.\end{array}\right\}.$$

Intuitively, $V_w(\tau^{<\epsilon}_{A_\varphi})$[2] is the set of all dataset that reject a null hypothesis φ. More specifically, the *p-value* $\Pr_{d \sim D_{t,\varphi}}[\, d \preccurlyeq^{(s)} t(o)\,]$ is the probability that a data d is at most as likely as the test statistic $t(o)$ when it is sampled from the distribution $D_{t,\varphi}$ in the world w; e.g., $\Pr_{d \sim N(0,1)}[\, d \preccurlyeq^{(\mathsf{T})} 1.96\,] = \Pr_{d \sim N(0,1)}[\,|d| \geq 1.96\,] = 0.05$. By definition, we have:

$$\mathfrak{M}, w \models \tau^{<\epsilon}_{A_\varphi}(y) \text{ iff } \Pr_{d \sim D_{t,\varphi}}[\, d \preccurlyeq^{(s)} t(w(y))\,] < \epsilon \text{ and } w(y) \text{ is not used in other tests than } A_\varphi.$$

This represents that in the possible world w, the observation of a dataset y is unlikely to occur (except with probability ϵ) according to the hypothesis test A_φ where the test statistic follows the distribution $D_{t,\varphi}$ in the world w.

Then the statistical belief modality $\mathbf{K}^{<\epsilon}_{y,A}$ is interpreted as:

$$\begin{aligned}
& \mathfrak{M}, w \models \mathbf{K}^{<\epsilon}_{y,A}\,\varphi \\
& \text{iff } \mathfrak{M}, w \models \mathbf{K}(\varphi \vee \tau^{<\epsilon}_{A_{\neg\varphi}}(y)) \\
& \text{iff for all } w',\ (w, w') \in \mathcal{R} \text{ implies } \mathfrak{M}, w' \models \neg\varphi \to \tau^{<\epsilon}_{A_{\neg\varphi}}(y).
\end{aligned}$$

Intuitively, $\mathbf{K}^{<\epsilon}_{y,A}\,\varphi$ represents a belief that an alternative hypothesis φ on the population is true. More specifically, $w' \models \neg\varphi \to \tau^{<\epsilon}_{A_{\neg\varphi}}(y)$ means that if we consider a possible world w' where the null hypothesis $\neg\varphi$ is true, then a hypothesis test $A_{\neg\varphi}$ would conclude that the observation of a dataset y is unlikely to occur (with exceptions at most ϵ), i.e., $\tau^{<\epsilon}_{A_{\neg\varphi}}(y)$ holds in w'. We discuss the implication of our formalization of $\mathbf{K}^{<\epsilon}_{y,A}$ in Section 5.3.

Although the modality $\mathbf{K}$ expresses the knowledge in terms of S5, the syntax sugar $\mathbf{K}^{<\epsilon}_{y,A}\,\varphi$ represents belief instead of knowledge. This is because φ can be false when $\tau^{<\epsilon}_{A_{\neg\varphi}}(y)$ holds (i.e., we may have a false belief on φ when the sampled dataset y is unluckily far from the population).

[2]We also define the interpretation of $\tau^{\bowtie\epsilon}_{A_\varphi}(y)$ analogously.

Example 4 (Statistical belief in Z-tests). *Recall again the two-tailed Z-test for two population means in Example 1. The alternative hypothesis is $\varphi_1 \stackrel{\text{def}}{=} (\mu_{\mathsf{ppl1}} \neq \mu_{\mathsf{ppl2}})$, and the null hypothesis is $\varphi_0 \stackrel{\text{def}}{=} (\mu_{\mathsf{ppl1}} = \mu_{\mathsf{ppl2}})$. We denote this Z-test by $A_{\varphi_0} = (\varphi_0, t, N(0,1), \preccurlyeq^{(\mathsf{T})})$.*

Suppose that in a world w, we sample two datasets $w(y_1)$ and $w(y_2)$ respectively from two populations $w(x_1)$ and $w(x_2)$, and calculate the Z-test statistic $t(w(y_1), w(y_2))$ defined in Example 1. When the null hypothesis φ_0 is true, $t(w(y_1), w(y_2))$ follows the distribution $N(0,1)$.

If $t(w(y_1), w(y_2)) = 3$, $\Pr_{d \sim N(0,1)}[d \preccurlyeq^{(\mathsf{T})} t(w(y_1), w(y_2))] < 0.05$. Then the null hypothesis φ_0 is rejected, and we obtain the statistical belief that the alternative hypothesis φ_1 is true with the significance level 0.05, i.e., $w \models \mathbf{K}^{<0.05}_{y,A}\,\varphi_1$. In contrast, if $t(w(y_1), w(y_2)) = 1.8$, then $w \models \neg\,\mathbf{K}^{<0.05}_{y,A}\,\varphi_1$ because $\Pr_{d \sim N(0,1)}[\, d \preccurlyeq^{(\mathsf{T})} t(w(y_1), w(y_2))\,] > 0.05$.

5.3 Remarks on the Formalization

Implication The universe $\mathcal{W}$ of the model $\mathfrak{M}$ is assumed to include all possible worlds we can imagine. If there exists no possible world satisfying the null hypothesis $\neg\varphi$ in the model $\mathfrak{M}$, then φ is satisfied in all worlds in $\mathfrak{M}$, hence so are $\mathbf{K}\varphi$ and $\mathbf{K}^{<\epsilon}_{y,A}\,\varphi$. This reflects the fact that if we cannot imagine a possible world where $\neg\varphi$ is true, then we already know that φ is true without executing hypothesis tests.

Type II error The *type II error rate* is the probability that the hypothesis test A does not reject the null hypothesis φ_{null} when φ_{null} is false. Assume that the true population satisfies a hypothesis ξ in the world w. Let y' be a dataset such that the p-value (type I error rate) α of the test A is 0.05; i.e., $w \models \mathbf{K}^{0.05}_{y',A}\,\neg\varphi_{\mathsf{null}}$. Then the type II error rate β when $\alpha = 0.05$ is given by $w \models \mathbf{K}^{\beta}_{y',A}\,\neg\xi$.

5.4 Properties of Statistical Beliefs

The statistical possibility $\mathbf{P}^{<\epsilon}_{y,A}\,\varphi$ means that we think a null hypothesis φ may be true after a hypothesis test A did not reject φ with a significance level ϵ. Formally, we have:

$$\begin{aligned}
& \mathfrak{M}, w \models \mathbf{P}^{<\epsilon}_{y,A}\,\varphi \\
& \text{iff there is a } w' \text{ s.t. } (w, w') \in \mathcal{R} \text{ and } \mathfrak{M}, w' \not\models \neg\varphi \vee \tau^{<\epsilon}_{A_\varphi}(y) \\
& \text{iff } \mathfrak{M}, w \models \mathbf{P}(\varphi \wedge \neg\tau^{<\epsilon}_{A_\varphi}(y)).
\end{aligned}$$

Now we show basic properties of statistical beliefs.

Proposition 1 (Properties of $\mathbf{K}^{<\epsilon}_{y,A}$). *Let φ be a formula, A be a hypothesis test, $y \in \mathsf{Var}_{\mathsf{obs}}$, and $\varepsilon \in \mathbb{R}_{\geq 0}$.*

1. *Knowledge is also regarded as belief: $\models \mathbf{K}\varphi \to \mathbf{K}^{<\epsilon}_{y,A}\,\varphi$.*
2. *If we believe φ based on a test A, then we know this statistical belief; i.e., $\models \mathbf{K}^{<\epsilon}_{y,A}\,\varphi \to \mathbf{K}\,\mathbf{K}^{<\epsilon}_{y,A}\,\varphi$.*
3. *If we failed to reject φ and think it possible, then we know this possibility; i.e., $\models \mathbf{P}^{<\epsilon}_{y,A}\,\varphi \to \mathbf{K}\,\mathbf{P}^{<\epsilon}_{y,A}\,\varphi$.*
4. *If $\epsilon \leq \epsilon'$, $\models \mathbf{K}^{<\epsilon}_{y,A}\,\varphi \to \mathbf{K}^{<\epsilon'}_{y,A}\,\varphi$ and $\models \mathbf{P}^{<\epsilon'}_{y,A}\,\varphi \to \mathbf{P}^{<\epsilon}_{y,A}\,\varphi$.*

5. $\mathbf{K}^{<\epsilon}_{y,A}\,\varphi$ *may represent a false belief. The alternative hypothesis φ we believe may be false, i.e., the rejected null hypothesis $\neg\varphi$ may be true: $\epsilon > 0$ iff $\not\models \mathbf{K}^{<\epsilon}_{y,A}\,\varphi \to \varphi$.*
6. $\models \mathbf{K}^{<\epsilon}_{y,A}\,\varphi \to \mathbf{K}^{<\epsilon}_{y}\,\varphi$.

The proofs are straightforward from definitions. See Appendix B.2 for the proofs.

6 A Simple Programming Language

We introduce an imperative programming language Prog.

6.1 Syntax of Prog

We define the syntax of Prog by the following BNF:

$$
\begin{aligned}
T &::= \texttt{bool} \mid \texttt{int} \mid \texttt{real} \mid T \times T \mid \texttt{list}(T) && \text{(Types)}\\
e &::= v \mid f(e_1, \dots, e_k) && \text{(Terms)}\\
c &::= \texttt{skip} \mid v := e && \text{(Commands)}\\
C &::= c \mid C; C \mid C \| C \mid \texttt{if}\ e\ \texttt{then}\ C\ \texttt{else}\ C \mid \texttt{loop}\ e\ \texttt{do}\ C && \text{(Programs)}
\end{aligned}
$$

where v is an element of $\mathsf{Var}_{\mathsf{obs}}$, f is a (built-in) function symbol, and constants are dealt as functions with arity 0. Notice that a program can handle only observable variables.

T represents *types*. A type is either `bool` for Boolean values, `int` for integers, `real` for real numbers, $T_1 \times T_2$ for pairs consisting of a value of type T_1 and a value of type T_2, or $\texttt{list}(T)$ for lists of values of type T. e represents *expressions* that evaluate to values. An expression is either a variable v or a function call $f(e_1, \dots, e_k)$; the latter is typically a call to a function that computes a test statistic. c and C represent *commands* and *programs* respectively. We give their intuitive explanation as follows.

- `skip` does nothing.
- $v := e$ updates v with the result of an evaluation of e.
- $C_1; C_2$ executes C_1 and then C_2.
- $C_1 \| C_2$ executes C_1 and C_2 in parallel that may share some data.
- $\texttt{if}\ e\ \texttt{then}\ C_1\ \texttt{else}\ C_2$ executes C_1 if e evaluates to `true`; executes C_2 if e evaluates to `false`.
- $\texttt{loop}\ e\ \texttt{do}\ C$ iteratively executes C as long as e evaluates to `true`.

For instance, the example program in Section 7.5 conforms to the programming language Prog.

Hereafter we assume that all programs are well-typed although we do not explicitly mention the types. Checking this condition for our language can be done by adapting a standard type-checking algorithm to our setting.

We write $\mathsf{upd}(C)$ for the set of variables that may be updated by executing C: $\mathsf{upd}(\texttt{skip}) = \emptyset$, $\mathsf{upd}(v{:=}e) = \{v\}$, $\mathsf{upd}(C_1; C_2) = \mathsf{upd}(C_1 \| C_2) = \mathsf{upd}(\texttt{if}\ e\ \texttt{then}\ C_1\ \texttt{else}\ C_2) = \mathsf{upd}(C_1) \cup \mathsf{upd}(C_2)$, and $\mathsf{upd}(\texttt{loop}\ e\ \texttt{do}\ C) = \mathsf{upd}(C)$.

Then we impose the following restriction to every occurrence of $C_1 \| C_2$: $\mathsf{upd}(C_1) \cap \mathsf{Var}(C_2) = \mathsf{upd}(C_2) \cap \mathsf{Var}(C_1) = \emptyset$. This restriction is to ensure that an execution of C_1 does not interfere with that of C_2.

6.2 Semantics of Prog

We define the semantics of Prog over a Kripke model $\mathfrak{M}$ with labeled transitions given in Section 4.3. The semantics is based on the standard structural operational semantics (e.g. (Nielson and Nielson 2007)).

For a possible world $w \in \mathcal{W}$ and $n = \mathsf{len}(w)$, we write

$$w = w[0], w[1], \dots, w[n-2], (m, a)$$

where (m, a) is the current state $w[n-1]$ with an assignment $m\colon \mathsf{Var} \to \mathcal{O}$ and an action a in the last transition in $\mathfrak{M}$.

For the assignment m of the current state of w, we define the evaluation $[\![e]\!]_m$ of a term e inductively by $[\![v]\!]_m = m(v)$ and $[\![f(e_1, \dots, e_k)]\!] = [\![f]\!]([\![e_1]\!]_m, \dots, [\![e_k]\!])$.

As in Figure 1, we define a binary relation

$$\longrightarrow \subseteq (\mathsf{Prog} \times \mathcal{W}) \times ((\mathsf{Prog} \times \mathcal{W}) \cup \mathcal{W})$$

that relates a pair $\langle C, w\rangle$ consisting of a program C and a possible world w to its next step of execution. If C is terminated, the next step will be a possible world w', otherwise the execution continues to the $\langle C', w'\rangle$.

$$\langle \texttt{skip}, w\rangle \longrightarrow w; (m, \texttt{skip}),$$
$$\langle v := e, w\rangle \longrightarrow w; (m[v \mapsto [\![e]\!]_m], v := e),$$
$$\frac{\langle C_1, w\rangle \longrightarrow w'}{\langle C_1; C_2, w\rangle \longrightarrow \langle C_2, w'\rangle}$$
$$\frac{\langle C_1, w\rangle \longrightarrow \langle C_1', w'\rangle}{\langle C_1; C_2, w\rangle \longrightarrow \langle C_1'; C_2, w'\rangle}$$
$$\langle \texttt{if}\ e\ \texttt{then}\ C_1\ \texttt{else}\ C_2, w\rangle \longrightarrow \begin{cases} \langle C_1, w\rangle & [\![e]\!]_m = \top \\ \langle C_2, w\rangle & [\![e]\!]_m = \bot \end{cases}$$
$$\langle \texttt{loop}\ e\ \texttt{do}\ C, w\rangle \longrightarrow \begin{cases} \langle C; \texttt{loop}\ e\ \texttt{do}\ C, w\rangle & [\![e]\!]_m = \top \\ w & [\![e]\!]_m = \bot \end{cases}$$
$$\frac{\langle C_1, w\rangle \longrightarrow \langle C_1', w'\rangle}{\langle C_1 \| C_2, w\rangle \longrightarrow \langle C_1' \| C_2, w'\rangle}$$
$$\frac{\langle C_1, w\rangle \longrightarrow w'}{\langle C_1 \| C_2, w\rangle \longrightarrow \langle C_2, w'\rangle}$$
$$\frac{\langle C_2, w\rangle \longrightarrow \langle C_2', w'\rangle}{\langle C_1 \| C_2, w\rangle \longrightarrow \langle C_1 \| C_2', w'\rangle}$$
$$\frac{\langle C_2, w\rangle \longrightarrow w'}{\langle C_1 \| C_2, w\rangle \longrightarrow \langle C_1, w'\rangle}$$

Figure 1: Rules of execution of programs.

Remark that the semantics of a program contains the trace of commands executed in it. Hence, even if programs finally have the same result, their semantics may be different.

$$\langle v := v + 1, ([v{\mapsto}1], a)\rangle \longrightarrow ([v{\mapsto}1], a), ([v{\mapsto}2], v := 1 + 1)$$
$$\langle v := 2 * v, ([v{\mapsto}1], a)\rangle \longrightarrow ([v{\mapsto}1], a), ([v{\mapsto}2], v := 2 * v)$$

We define the semantic relation $[\![C]\!] \subseteq \mathcal{W} \times \mathcal{W}$ by

$$[\![C]\!](w) = \{w' \mid \langle C, w\rangle \longrightarrow^* w'\}$$

where $\longrightarrow^*$ is the transitive closure of $\longrightarrow$.

Remark on Parallel Compositions Since parallel compositions are nondeterministic, $w' \in [\![C]\!](w)$ may not be unique. However, the resulting world w' are essentially the same, because for parallel composition $C_1 \| C_2$, the world $w' \in [\![C_1 \| C_2]\!](w)$ is convertible to a pair of $w_1 \in [\![C_1]\!](w)$ and $w_2 \in [\![C_1]\!](w)$ and vise versa.

If we have $\langle C_b, w\rangle \longrightarrow^* w; u_b$ for $b = 1, 2$, then by $\mathsf{upd}(C_b) \cap \mathsf{Var}(C_{3-b}) = \emptyset$, we obtain a sequence u' such that $\langle C_1 \| C_2, w\rangle \longrightarrow^* w; u'$ by combining u_1 and u_2.

Conversely, if $\langle C_1 \| C_2, w\rangle \longrightarrow^* w; u'$, we can decompose u' into u_1 and u_2 such that $\langle C_b, w\rangle \longrightarrow^* w; u_b$ for $b = 1, 2$ (for detail, see Appendix B.2).

Then, for any pair of $\tau^{\bowtie}$-free assertions φ_1 and φ_2 satisfying $\mathsf{upd}(C_b) \cap \mathsf{fv}(\varphi_{3-b}) = \emptyset$ for $b = 1, 2$, for any possible world $w' \in [\![C_1 \| C_2]\!](w)$, we can write $w' = w; u'$, and decompose u' into u_1 and u_2 as above. We then obtain $w' \models \varphi_1 \wedge \varphi_2$ iff $w; u_1 \models \varphi_1$ and $w; u_2 \models \varphi_2$.

Procedures of Hypothesis Testing Finally, we present the interpretation of a program $f_{A^{(s)}_{\varphi_0}}$ for a hypothesis test $A^{(s)}_{\varphi_0} = (\varphi_0, t, D_{t,\varphi_0}, \preccurlyeq^{(s)})$ with a null hypothesis φ_0, a test statistic t, and a test type s. For a dataset y and an assignment m, $[\![f_{A^{(s)}_{\varphi_0}}(y)]\!]_m$ represents the p-value:

$$[\![f_{A^{(s)}_{\varphi_0}}(y)]\!]_m = \Pr_{d \sim D_{t,\varphi_0}} [\, d \preccurlyeq^{(s)} t(m(y)) \,], \tag{5}$$

which is the probability that a data d is at most as likely as the test statistic $t(m(y))$ when it is sampled from D_{t,φ_0} in the world where the null hypothesis φ_0 is true.

7 Belief Hoare Logic for Hypothesis Testing

We introduce *belief Hoare logic* (BHL) for formalizing and reasoning about statistical inference using hypothesis tests. Then we describe the reasoning about the *multiple comparison problem* and *p-value hacking* using BHL.

7.1 Hoare Triples

An *environment* is defined as a pair $\Gamma = (\Gamma^{\mathsf{inv}}, \Gamma^{\mathsf{obs}})$ consisting of an *invisible environment* Γ^{inv} and an *observable environment* Γ^{obs} that assign types to invisible variables and to observable variables, respectively. We denote by $\mathsf{Var}(\Gamma)$ the set of all variables occurring in an environment Γ, and by Env the set of all possible environments.

A *judgment* is of the form $\Gamma \vdash \{\psi\}\ C\ \{\varphi\}$ where $\Gamma \in \mathsf{Env}$, $\psi, \varphi \in \mathsf{Fml}$, and $C \in \mathsf{Prog}$. Intuitively, this represents that when the *precondition* ψ is satisfied, executing the program C results in satisfying the *postcondition* φ.

We say that a judgment $\Gamma \vdash \{\psi\}\ C\ \{\varphi\}$ is *valid* iff for any model $\mathfrak{M}$ and any possible world w, if $\mathfrak{M}, w \models \psi$, then $\mathfrak{M}, w' \models \varphi$ for all $w' \in [\![C]\!](w)$.

We write $\Gamma \models \varphi$ if $\mathfrak{M}, w \models \varphi$ for any model $\mathfrak{M}$ and any world w that respects the type information in Γ (i.e., the type of $w(v)$ being $\Gamma(v)$ for any variable $v \in \mathsf{Var}$).

7.2 Inference Rules

Next, we present the inference rules for belief Hoare logic (BHL). The rules are classified into those for basic command constructs (Figure 2) and for hypothesis testing constructs (Figure 3). The latter includes axiom schemas that can be instantiated to a variety of concrete hypothesis test methods; we present such instantiation in Section 7.3.

$$\Gamma \vdash \{\psi\}\ \mathtt{skip}\ \{\psi\} \qquad \text{(SKIP)}$$

$$\frac{\Gamma(x) = \Gamma(y)}{\Gamma \vdash \{\varphi[x \mapsto y]\}\ x := y\ \{\varphi\}} \qquad \text{(UPDVAR)}$$

$$\frac{\Gamma \vdash \{\psi\}\ C_1\ \{\psi'\} \qquad \Gamma \vdash \{\psi'\}\ C_2\ \{\varphi\}}{\Gamma \vdash \{\psi\}\ C_1; C_2\ \{\varphi\}} \qquad \text{(SEQ)}$$

$$\frac{\Gamma \vdash \{\psi\}\ C\ \{\varphi\}}{\Gamma \vdash \{\psi\}\ \mathtt{skip} \parallel C\ \{\varphi\}} \qquad \text{(PAR-SKIPADDL)}$$

$$\frac{\Gamma \vdash \{\psi\}\ C\ \{\varphi\}}{\Gamma \vdash \{\psi\}\ C \parallel \mathtt{skip}\ \{\varphi\}} \qquad \text{(PAR-SKIPADDR)}$$

$$\frac{\Gamma \vdash \{\psi\}\ \mathtt{skip} \parallel C\ \{\varphi\}}{\Gamma \vdash \{\psi\}\ C\ \{\varphi\}} \qquad \text{(PAR-SKIPRML)}$$

$$\frac{\Gamma \vdash \{\psi\}\ C \parallel \mathtt{skip}\ \{\varphi\}}{\Gamma \vdash \{\psi\}\ C\ \{\varphi\}} \qquad \text{(PAR-SKIPRMR)}$$

$$\frac{\Gamma \vdash \{\psi \wedge e\}\ C_1\ \{\varphi\} \qquad \Gamma \vdash \{\psi \wedge \neg e\}\ C_2\ \{\varphi\}}{\Gamma \vdash \{\psi\}\ \mathtt{if}\ e\ \mathtt{then}\ C_1\ \mathtt{else}\ C_2\ \{\varphi\}} \qquad \text{(IF)}$$

$$\frac{\Gamma \vdash \{\psi \wedge e\}\ C\ \{\psi\}}{\Gamma \vdash \{\psi\}\ \mathtt{loop}\ e\ \mathtt{do}\ C\ \{\psi \wedge \neg e\}} \qquad \text{(LOOP)}$$

$$\frac{\Gamma \models \psi \to \psi' \quad \Gamma \vdash \{\psi'\} C \{\varphi'\} \quad \Gamma \models \varphi' \to \varphi}{\Gamma \vdash \{\psi\} C \{\varphi\}} \qquad \text{(CONSEQ)}$$

Figure 2: Rules for basic constructs for commands.

The rules for basic constructs in Figure 2 are standard; the readers are referred to a standard textbook on Hoare logic (Winskel 1993) for details. We remark the features:

- In the rules IF and LOOP, the guard condition e is a Boolean expression implicitly used as a logical predicate in the preconditions and the postconditions as usual.
- The rule CONSEQ is used to weaken the precondition and strengthen the postcondition of a triple. The relation $\Gamma \models \varphi$ defined above is used in this rule.

Schemas for Single Hypothesis Testing In Figure 3 the axiom schemas TWO-T, LOW-T, and UP-T correspond to two-tailed, lower-tailed, and upper-tailed tests, respectively.

In these schemas, a dataset y is sampled from a population x, which follows a statistical model $P(\xi; \theta)$ with unknown true parameters $\xi \in \Phi$ and known parameters $\theta \in \Theta$.

To reason about the unknown parameter ξ, we perform a hypothesis test A_{φ_0} with a null hypothesis φ_0. Let φ_{L}, φ_{U}, and $\varphi_{\mathsf{T}} (\stackrel{\text{def}}{=} \varphi_{\mathsf{L}} \vee \varphi_{\mathsf{U}})$ be the alternative hypotheses for the lower-tailed, upper-tailed, and two-tailed tests, respectively.

When we consider both the lower-tail φ_{L} and upper-tail φ_{U} are possible before performing the test, we express them by $\mathbf{P}\varphi_{\mathsf{L}}$ and by $\mathbf{P}\varphi_{\mathsf{U}}$ in the precondition. Then we apply the schema TWO-T. When the two-tailed test $f_{A^{(\mathsf{T})}_{\varphi_0}}(y)$ returns the p-value $\alpha \in [0, 1]$, we obtain a statistical belief on the alternative hypothesis φ_{T}, namely, $\mathbf{K}^{\alpha}_{y,A}\, \varphi_{\mathsf{T}}$. When we consider only the lower-tail φ_{L} (resp. upper tail φ_{U}) possible, we apply LOW-T (resp. UP-T) and obtain a statistical belief on φ_{L} (resp. φ_{U}), namely, $\mathbf{K}^{\alpha}_{y,A}\, \varphi_{\mathsf{L}}$ (resp. $\mathbf{K}^{\alpha}_{y,A}\, \varphi_{\mathsf{U}}$).

$$\frac{\begin{array}{c}\Gamma^{\mathsf{inv}} = \{\xi : \Xi,\ x : \mathbb{D}\mathcal{X}\} \cup \mathsf{Var}_{\mathsf{inv}}(\{\psi', \varphi_{\mathsf{L}}, \varphi_{\mathsf{U}}\}),\ \ \Gamma^{\mathsf{obs}} = \{\theta : \Theta,\ n : \mathbb{N},\ y : \mathsf{list}\,\mathcal{X},\ \alpha : [0,1]\} \cup \mathsf{Var}_{\mathsf{obs}}(\{\psi', \varphi_{\mathsf{L}}, \varphi_{\mathsf{U}}\}),\\ \alpha \notin \mathsf{fv}(\{\varphi_{\mathsf{L}}, \varphi_{\mathsf{U}}\}),\ \ \psi' : \tau^{\bowtie}\text{-free},\ \ \psi \stackrel{\text{def}}{=} (x \approx P(\xi,\theta) \wedge y \underset{n}{\leftsquigarrow} x \wedge \psi')\end{array}}{(\Gamma^{\mathsf{inv}}, \Gamma^{\mathsf{obs}}) \vdash \{\psi[\alpha \mapsto f_{A^{(\mathsf{T})}_{\varphi_0}}(y)] \wedge \mathbf{P}\varphi_{\mathsf{L}} \wedge \mathbf{P}\varphi_{\mathsf{U}}\}\ \alpha := f_{A^{(\mathsf{T})}_{\varphi_0}}(y)\ \{\psi \wedge \mathbf{K}^{\alpha}_{y,A}\,\varphi_{\mathsf{T}}\}}\ (\text{TWO-T})$$

$$\frac{\begin{array}{c}\Gamma^{\mathsf{inv}} = \{\xi : \Xi,\ x : \mathbb{D}\mathcal{X}\} \cup \mathsf{Var}_{\mathsf{inv}}(\{\psi', \varphi_{\mathsf{L}}, \varphi_{\mathsf{U}}\}),\ \ \Gamma^{\mathsf{obs}} = \{\theta : \Theta,\ n : \mathbb{N},\ y : \mathsf{list}\,\mathcal{X},\ \alpha : [0,1]\} \cup \mathsf{Var}_{\mathsf{obs}}(\{\psi', \varphi_{\mathsf{L}}, \varphi_{\mathsf{U}}\}),\\ \alpha \notin \mathsf{fv}(\{\varphi_{\mathsf{L}}, \varphi_{\mathsf{U}}\}),\ \ \psi' : \tau^{\bowtie}\text{-free},\ \ \psi \stackrel{\text{def}}{=} (x \approx P(\xi,\theta) \wedge y \underset{n}{\leftsquigarrow} x \wedge \psi')\end{array}}{(\Gamma^{\mathsf{inv}}, \Gamma^{\mathsf{obs}}) \vdash \{\psi[\alpha \mapsto f_{A^{(\mathsf{L})}_{\varphi_0}}(y)] \wedge \mathbf{P}\varphi_{\mathsf{L}} \wedge \neg\mathbf{P}\varphi_{\mathsf{U}}\}\ \alpha := f_{A^{(\mathsf{L})}_{\varphi_0}}(y)\ \{\psi \wedge \mathbf{K}^{\alpha}_{y,A}\,\varphi_{\mathsf{L}}\}}\ (\text{LOW-T})$$

$$\frac{\begin{array}{c}\Gamma^{\mathsf{inv}} = \{\xi : \Xi,\ x : \mathbb{D}\mathcal{X}\} \cup \mathsf{Var}_{\mathsf{inv}}(\{\psi', \varphi_{\mathsf{L}}, \varphi_{\mathsf{U}}\}),\ \ \Gamma^{\mathsf{obs}} = \{\theta : \Theta,\ n : \mathbb{N},\ y : \mathsf{list}\,\mathcal{X},\ \alpha : [0,1]\} \cup \mathsf{Var}_{\mathsf{obs}}(\{\psi', \varphi_{\mathsf{L}}, \varphi_{\mathsf{U}}\}),\\ \alpha \notin \mathsf{fv}(\{\varphi_{\mathsf{L}}, \varphi_{\mathsf{U}}\}),\ \ \psi' : \tau^{\bowtie}\text{-free},\ \ \psi \stackrel{\text{def}}{=} (x \approx P(\xi,\theta) \wedge y \underset{n}{\leftsquigarrow} x \wedge \psi')\end{array}}{(\Gamma^{\mathsf{inv}}, \Gamma^{\mathsf{obs}}) \vdash \{\psi[\alpha \mapsto f_{A^{(\mathsf{U})}_{\varphi_0}}(y)] \wedge \neg\mathbf{P}\varphi_{\mathsf{L}} \wedge \mathbf{P}\varphi_{\mathsf{U}}\}\ \alpha := f_{A^{(\mathsf{U})}_{\varphi_0}}(y)\ \{\psi \wedge \mathbf{K}^{\alpha}_{y,A}\,\varphi_{\mathsf{U}}\}}\ (\text{UP-T})$$

$$\frac{\begin{array}{c}\text{For } b = 1,2,\ \Gamma^{\mathsf{inv}}_b = \mathsf{fv}^{\mathsf{inv}}(\{\psi_b, \psi'_b, \varphi_b\}),\ \ \Gamma^{\mathsf{obs}}_b = \{y_b : \mathsf{list}\,\mathcal{X},\ \alpha_b : [0,1]\} \cup \mathsf{fv}^{\mathsf{obs}}(\{\psi_b, \psi'_b, \varphi_b\}) \cup \mathsf{upd}(C_b),\\ \mathsf{upd}(C_b) \cap (\mathsf{fv}(\{\psi'_{3-b}, \varphi_{3-b}\}) \cup \{y_1, y_2\}) = \emptyset,\ \ \psi'_b : \tau^{\bowtie}\text{-free},\ \ \alpha_1, \alpha_2 \notin \mathsf{fv}(\{\varphi_1, \varphi_2\})\\ (\Gamma^{\mathsf{inv}}_1, \Gamma^{\mathsf{obs}}_1) \vdash \{\psi_1\}\ C_1\ \{\psi'_1 \wedge \mathbf{K}^{\alpha_1}_{y_1}\varphi_1\} \qquad (\Gamma^{\mathsf{inv}}_2, \Gamma^{\mathsf{obs}}_2) \vdash \{\psi_2\}\ C_2\ \{\psi'_2 \wedge \mathbf{K}^{\alpha_2}_{y_2}\varphi_2\}\end{array}}{(\Gamma^{\mathsf{inv}}_1 \cup \Gamma^{\mathsf{inv}}_2,\ \Gamma^{\mathsf{obs}}_1 \cup \Gamma^{\mathsf{obs}}_2) \vdash \{\psi_1 \wedge \psi_2\}\ C_1 \parallel C_2\ \{\psi'_1 \wedge \psi'_2 \wedge \mathbf{K}^{\leq \alpha_1 + \alpha_2}_{(y_1,y_2)}(\varphi_1 \vee \varphi_2)\}}\ (\text{MULT-}\vee)$$

$$\frac{\begin{array}{c}\text{For } b = 1,2,\ \Gamma^{\mathsf{inv}}_b = \mathsf{fv}^{\mathsf{inv}}(\{\psi_b, \psi'_b, \varphi_b\}),\ \ \Gamma^{\mathsf{obs}}_b = \{y_b : \mathsf{list}\,\mathcal{X},\ \alpha_b : [0,1]\} \cup \mathsf{fv}^{\mathsf{obs}}(\{\psi_b, \psi'_b, \varphi_b\}) \cup \mathsf{upd}(C_b),\\ \mathsf{upd}(C_b) \cap (\mathsf{fv}(\{\psi'_{3-b}, \varphi_{3-b}\}) \cup \{y_1, y_2\}) = \emptyset,\ \ \psi'_b : \tau^{\bowtie}\text{-free},\ \ \alpha_1, \alpha_2 \notin \mathsf{fv}(\{\varphi_1, \varphi_2\})\\ (\Gamma^{\mathsf{inv}}_1, \Gamma^{\mathsf{obs}}_1) \vdash \{\psi_1\}\ C_1\ \{\psi'_1 \wedge \mathbf{K}^{\alpha_1}_{y_1}\varphi_1\} \qquad (\Gamma^{\mathsf{inv}}_2, \Gamma^{\mathsf{obs}}_2) \vdash \{\psi_2\}\ C_2\ \{\psi'_2 \wedge \mathbf{K}^{\alpha_2}_{y_2}\varphi_2\}\end{array}}{(\Gamma^{\mathsf{inv}}_1 \cup \Gamma^{\mathsf{inv}}_2,\ \Gamma^{\mathsf{obs}}_2 \cup \Gamma^{\mathsf{obs}}_2) \vdash \{\psi_1 \wedge \psi_2\}\ C_1 \parallel C_2\ \{\psi'_1 \wedge \psi'_2 \wedge \mathbf{K}^{\leq \min(\alpha_1, \alpha_2)}_{(y_1,y_2)}(\varphi_1 \wedge \varphi_2)\}}\ (\text{MULT-}\wedge)$$

Figure 3: Axiom schemas and rules for hypothesis tests. TWO-T, LOW-T, and UP-T are schemas for two-tailed, lower-tailed, and upper-tailed tests, respectively. MULT-$\vee$ is the rule for the Bonferroni's method, and MULT-$\wedge$ is for the simultaneous tests without correction of α.

Rules for Multiple Tests The rule MULT-$\vee$ corresponds to the multiple tests by the *Bonferroni's method*. As illustrated in Section 3, a typical example is to test whether a drug has better efficacy than *at least one* of multiple drugs.

In the rule MULT-$\vee$, we have two datasets y_1 and y_2 respectively obtained by sampling from two populations x_1 and x_2 (that may have statistical relevance). Then we apply two separate hypothesis tests on y_1 in C_1 and on y_2 in C_2 to derive the *disjunctive* alternative hypothesis $\varphi_1 \vee \varphi_2$. We denote by α_1 and α_2 the p-values of these two tests when performed separately; i.e., $\mathbf{K}^{\alpha_1}_{y_1}\,\varphi_1$ and $\mathbf{K}^{\alpha_2}_{y_2}\,\varphi_2$ are satisfied.

However, the p-value when performing both the tests simultaneously (in $C_1 \| C_2$) is larger than α_1 and α_2. This is the so-called *multiple comparison problem*. By applying the Bonferroni's method, the p-value in total is bounded above by $\alpha_1 + \alpha_2$, namely, $\mathbf{K}^{\leq \alpha_1+\alpha_2}_{(y_1,y_2)}(\varphi_1 \vee \varphi_2)$ is satisfied. This reflects that BHL does *not* derive elementary mistakes (e.g., $\mathbf{K}^{\alpha_1}_{(y_1,y_2)}\,\varphi_1$) where the reported p-value α_1 is lower than the actual p-value in multiple comparison.

In contrast, the rule MULTI-$\wedge$ formalizes the tests for the *conjunctive* alternative hypothesis $\varphi_1 \wedge \varphi_2$, e.g., the program C_{drug} in Example 2, which tests whether a drug has better efficacy than *both* drugs. According to statistics, this does not make the p-value higher, i.e., the p-value is at most $\min(\alpha_1, \alpha_2)$. See Appendix for the validity.

Finally, we obtain the soundness of BHL from the validity of the rules. We show the proof in Appendix.

Theorem 1 (Soundness). *Every derivable judgment is valid.*

7.3 Instantiation to Concrete Test Methods

The axiom schemas for hypothesis tests in Figure 3 are instantiated with concrete examples of tests as follows. We first show the case of the two-tailed Z-test (Example 1).

Example 5 (Z-test)**.** The axiom for the Z-test comparing means of two populations with datasets y_1, y_2 of sample sizes $\mathsf{size}(y_1), \mathsf{size}(y_2)$ and a known variance σ^2 is given by instantiating TWO-T with the following parameters:

$$\begin{aligned} P_b(\mu_{\mathsf{ppl}\,b}, \sigma^2) &\stackrel{\text{def}}{=} N(\mu_{\mathsf{ppl}\,b}, \sigma^2) \quad \text{ for } b = 1,2\\ \varphi_0 &\stackrel{\text{def}}{=} (\mu_{\mathsf{ppl}1} {=} \mu_{\mathsf{ppl}2}) \quad \varphi_{\mathsf{T}} \stackrel{\text{def}}{=} (\mu_{\mathsf{ppl}1} {\neq} \mu_{\mathsf{ppl}2})\\ \preccurlyeq^{(\mathsf{T})} &= \{(r_1, r_2) \in \mathbb{R} \times \mathbb{R} \mid |r_1| > |r_2|\}\\ t(y_1, y_2) &= \frac{\mathsf{mean}(y_1) - \mathsf{mean}(y_2)}{\sigma\sqrt{1/\mathsf{size}(y_1) + 1/\mathsf{size}(y_2)}}\\ D_{t_\theta, \varphi_0} &= N(0,1) \quad \text{(the standard normal distribution).}\end{aligned}$$

Next we show the instantiation to the classical likelihood ratio test with a simple null hypothesis $\xi = \xi_0$ and a simple alternative hypothesis $\xi = \xi_1$, namely, in the setting of the Neyman-Pearson lemma.

Example 6 (Likelihood ratio test)**.** The goal of (the simplest version of) the likelihood ratio test is to determine which of two candidate distributions $D_p, D_q \in \mathbb{DR}$ is better to fit a dataset $y = (y_1, \ldots, y_n)$ of sample size n. The test can be reformulated with the statistical model $x \approx P(\xi)$ defined by

$$P(\xi) = \begin{cases} D_q & \text{if } \xi = \xi_0\\ D_p & \text{if } \xi = \xi_1\end{cases}$$

and the following null and alternative hypotheses:

$$\varphi_0 \overset{\text{def}}{=} (\xi = \xi_0) \qquad \varphi_{\mathsf{L}} \overset{\text{def}}{=} (\xi = \xi_1).$$

The the likelihood function L for this test is given by

$$L(y|\xi_0) = \prod_{i=1}^{n} q(y_i) \quad L(y|\xi_1) = \prod_{i=1}^{n} p(y_i).$$

Then, the likelihood ratio $t(y)$ is given by

$$t(y) = \frac{L(y|\xi_0)}{L(y|\xi_1)} = \frac{\prod_{i=1}^{n} q(y_i)}{\prod_{i=1}^{n} p(y_i)}$$

where p and q are the density functions of D_p and D_q respectively. In the likelihood ratio test, for a given α and a threshold k such that $\Pr_{d_1,\ldots,d_n\sim D_q}[t((d_1,\ldots,d_n)) \leq k] \leq \alpha$, if we have $t(y) \leq k$, the likelihood $L(y|\xi_0)$ is too small to accept the distribution D_q. We then conclude that the other candidate D_p is better to fit y (thus this test is lower-tailed).

Conversely, the p-value of this test is defined by

$$\Pr_{d_1,\ldots,d_n\sim D_q}[t((d_1,\ldots,d_n)) \leq t(y)]. \tag{6}$$

The axiom for the likelihood ratio test is given by instantiating LOW-T with the above $P(\xi)$, φ_0, φ_{L}, $t(y)$ and

$$\preccurlyeq^{(\mathsf{L})} = \{(r_1, r_2) \in \mathbb{R} \times \mathbb{R} \mid r_1 \leq r_2\}$$
$$D_{t_\theta,\varphi_0} = \frac{\prod_{i=1}^{n} q(D_q)}{\prod_{i=1}^{n} p(D_q)}$$

where $p(D_q)$ and $q(D_q)$ are the probability distributions respectively defined by $p(D_q)(A) \overset{\text{def}}{=} D_q(p^{-1}(A))$ and $q(D_q)(A) \overset{\text{def}}{=} D_q(q^{-1}(A))$ for any measurable subset $A \subseteq \mathbb{R}$. Intuitively, $p(D_q)$ and $q(D_q)$ represent the probability distributions of $p(y)$ and $q(y)$ when y is sampled from D_q. By instantiating the p-value $[\![f_{A^{(s)}_{\varphi_0}}(y)]\!]$ in (5), we obtain (6).

Bayesian hypothesis test is given in an analogous way.

Example 7 (Bayesian hypothesis test)**.** Consider the Bayesian likelihood ratio test with a dataset y of sample size n, prior distributions $D_{p'}, D_{q'} \in \mathbb{DR}$ with density functions p' and q', and posterior distributions $D_{p(z)}, D_{q(z)} \in \mathbb{DR}$ with density functions $p(-|z)$ and $q(-|z)$.

The goal of this test is to determine whether the dataset y is sampled from $D_{q(z)}$ when z follows $D_{q'}$. The null hypothesis is that y is sampled from $D_{p(z)}$ when z follows $D_{p'}$.

A statistical model $x \approx P(\xi)$ of this test is defined as follows: First, we introduce the statistical models $z \approx P_0(\xi)$ and $x \approx P_1(\xi, z)$ of prior and posterior distributions by

$$P_0(\xi) = \begin{cases} D_{q'} & \text{if } \xi = \xi_0 \\ D_{p'} & \text{if } \xi = \xi_1 \end{cases} \quad P_1(\xi, z) = \begin{cases} D_{q(z)} & \text{if } \xi = \xi_0 \\ D_{p(z)} & \text{if } \xi = \xi_1 \end{cases}.$$

Next, for each $\xi = \xi_0, \xi_1$, we define the probability measure $P(\xi) \in \mathbb{DR}$ by for any measurable subset $A \subseteq \mathbb{R}$,

$$(P(\xi))(A) \overset{\text{def}}{=} \int_{\mathbb{R}} h_{\xi,A}\, d\mu_\xi$$

where $\mu_\xi = P_0(\xi)$ and $h_{\xi,A} \colon \mathbb{R} \to \mathbb{R}$ is a measurable function defined by $h_{\xi,A}(z) = (P_1(\xi, z))(A)$.

The axiom for this test is given by instantiating LOW-T with the above model P and the following parameters:

$$\varphi_0 \overset{\text{def}}{=} (\xi = \xi_0) \quad \varphi_{\mathsf{L}} \overset{\text{def}}{=} (\xi = \xi_1)$$
$$\preccurlyeq^{(\mathsf{L})} = \{(r_1, r_2) \in \mathbb{R} \times \mathbb{R} \mid r_1 \leq r_2\}$$
$$t(y) = \frac{\int q'(z) \prod_{i=1}^{n} q(y_i|z)dz}{\int p'(z) \prod_{i=1}^{n} p(y_i|z)dz}$$
$$D_{t_\theta,\varphi_0} = \frac{\int q'(z) \prod_{i=1}^{n} q(D_{q(z)}|z)dz}{\int p'(z) \prod_{i=1}^{n} p(D_{q(z)}|z)dz}.$$

Unlike the (classical) likelihood ratio test, the *Bayes factor* $t(y)$ is the ratio of the following marginal likelihoods:

$$L(y|\xi_0) = \int q'(z) \prod_{i=1}^{n} q(y_i|z)dz$$
$$L(y|\xi_1) = \int p'(z) \prod_{i=1}^{n} p(y_i|z)dz.$$

7.4 Reasoning About Multiple Comparison

We illustrate how BHL reasons about the multiple comparison in Example 2, where the derivation of (3) guarantees that the hypothesis tests are applied appropriately in C_{drug} in (1). More details can be found in Appendix.

In the derivation, we obtain the following judgments:

$$\Gamma \vdash \{\psi_{\mathsf{pre}}\}\; C_{12}\; \{\psi_{\mathsf{pre}} \wedge \mathbf{K}_y^{\alpha_{12}}\, \varphi_{12}\}$$
$$\Gamma \vdash \{\psi_{\mathsf{pre}} \wedge \mathbf{K}_y^{\alpha_{12}}\, \varphi_{12} \wedge \alpha_{12} \leq 0.05\}\; C_{13}\; \{\varphi_{12}^{\mathsf{bel}} \to \varphi^{\mathsf{bel}}\}$$
$$\Gamma \vdash \{\psi_{\mathsf{pre}} \wedge \mathbf{K}_y^{\alpha_{12}}\, \varphi_{12} \wedge \alpha_{12} > 0.05\}\; \texttt{skip}\; \{\varphi_{12}^{\mathsf{bel}} \to \varphi^{\mathsf{bel}}\}$$

where $\varphi_{12}^{\mathsf{bel}} \overset{\text{def}}{=} \mathbf{K}_y^{\leq 0.05}\, \varphi_{12}$ and $\varphi^{\mathsf{bel}} \overset{\text{def}}{=} \mathbf{K}_y^{\leq \alpha}(\varphi_{12} \wedge \varphi_{13})$. The second judgment is derived by the two-tailed test axiom, MULT-$\wedge$, and CONSEQ. The last judgment is derived from $\Gamma \models (\psi_{\mathsf{pre}} \wedge \mathbf{K}_y^{\alpha_{12}}\, \varphi_{12} \wedge \alpha_{12} > 0.05) \to \neg\varphi_{12}^{\mathsf{bel}}$ and $\Gamma \models \neg\varphi_{12}^{\mathsf{bel}} \to (\varphi_{12}^{\mathsf{bel}} \to \varphi^{\mathsf{bel}})$ and CONSEQ. Applying IF to the last two judgments, we have $\Gamma \vdash \{\psi_{\mathsf{pre}} \wedge \mathbf{K}_y^{\alpha_{12}}\, \varphi_{12}\}$ $\texttt{if}\ \alpha_{12} \leq 0.05\ \texttt{then}\ C_{13}\ \texttt{else skip}$ $\{\varphi_{12}^{\mathsf{bel}} \to \varphi^{\mathsf{bel}}\}$; composing it with the first judgment by applying SEQ, we have the judgment in (3).

In contrast, the program $C_{12} \parallel C_{13}$ in (4) shows the multiple comparison problem. Since the alternative hypothesis $\varphi_{12} \vee \varphi_{13}$ is disjunctive, we apply the rule MULT-$\vee$ to drive the belief $\mathbf{K}_{(y,y)}^{\leq \alpha_{12}+\alpha_{13}}(\varphi_{12} \vee \varphi_{13})$, with a p-value (larger than α_{12} and α_{13}) at most $\alpha_{12} + \alpha_{13}$.

7.5 Reasoning About p-Value Hacking

We informally describe how our framework can be applied to reason about a program for *p-value hacking*, i.e., a scientifically malignant technique to obtain a low p-value. The following program c_{pHack} is an example of p-value hacking that conducts a hypothesis test on different datasets y_1 and y_2, and ignores the experiment showing a higher p-value to report only a lower p-value:

$$(\alpha_1 := f_{A^{(\mathsf{T})}_{\varphi_0}}(y_1) \| \alpha_2 := f_{A^{(\mathsf{T})}_{\varphi_0}}(y_2));$$
$$\texttt{if}\ \alpha_1 < \alpha_2\ \texttt{then}\ \alpha := \alpha_1\ \texttt{else}\ \alpha := \alpha_2$$

We write φ_{alt} for the alternative hypothesis of this test.

For the reported p-value α to be an actual p-value, $\mathbf{K}_{(y_1,y_2)}^{\leq \alpha}\varphi_{\mathsf{alt}}$ needs to hold as a postcondition of c_{pHack}. Then

$(\alpha_1 < \alpha_2 \rightarrow \mathbf{K}^{\leq \alpha_1}_{(y_1,y_2)}\, \varphi_{\mathsf{alt}}) \wedge (\alpha_1 \geq \alpha_2 \rightarrow \mathbf{K}^{\leq \alpha_2}_{(y_1,y_2)}\, \varphi_{\mathsf{alt}})$ must hold at the end of the first line of c_{pHack} due to the rules UPDVAR and IF. However, this is not implied by the post-conditions of MULT-$\vee$ or MULT-$\wedge$. Since $\mathbf{K}^{<\alpha_1+\alpha_2}_{(y_1,y_2)}(\varphi_1 \vee \varphi_2)$ is a postcondition of c_{pHack}, the total p-value $\alpha_1 + \alpha_2$ should be reported without ignoring any experiments.

8 Discussion

In this section, we provide a whole picture of the justification of statistical belief inside and outside BHL.

A statistical belief derived in a program relies on the following three issues: (i) the validity of hypothesis test methods themselves, (ii) the satisfaction of the empirical conditions required for the hypothesis tests, and (iii) the appropriate usage of hypothesis tests in the program. In our framework, these are respectively addressed by (a) the validity of axioms and rules, (b) the (manual) confirmation of the preconditions in a judgment, and (c) the proof for the judgment.

8.1 Validity of Hypothesis Test Methods

The validity of hypothesis test methods is not ensured by mathematics alone. The philosophy of statistics has a long history of argument on the proper interpretation of hypothesis testing. One of the most notable examples is the argument between the frequentist and the Bayesian statistics, which still has many issues to be discussed (Sober 2008).

We also remark that statistical methods occasionally involve some approximation of numerical values. However, we may not always confirm the validity of approximation rigorously, i.e., whether the approximation is valid to the specific situation we apply the statistical methods.

For these reasons, we do not attempt to formalize the "justification" for hypothesis test methods within BHL, and left them for future work. Instead, we define axiom schemas for hypothesis tests that are commonly used in practice and explained in textbooks, e.g., (Kanji 2006). Then we focus on the logical aspects of the appropriate usage of hypothesis tests, which has been a long-standing, practical concern but has not been formalized using a symbolic logic before.

One of the advantages of this approach is that we do not adhere to a specific philosophy of statistics, but can model both the frequentist and the Bayesian statistics by introducing an axiom/rule corresponding to each hypothesis test.

8.2 Clarification of Empirical Conditions

The hypothesis test methods usually assume some empirical conditions on the unknown population from which the dataset is sampled. Typically, many parametric tests require that the population follows a normal distribution. For instance, the Z-test in Example 5 assumes that the population follows a normal distribution with known variance, but this cannot be rigorously confirmed or justified in general.

In some cases, such conditions on the unknown population may be confirmed approximately or partially by some exploratory observations on the sampled data and by prior knowledge of some properties on the population (outside the statistical inference). As far as we know, there has been no general method for justifying such empirical conditions rigorously. Thus, the formal justification of those conditions themselves would require further research in statistics.

In the present paper, the empirical conditions on the unknown population remain to be assumptions from the viewpoint of formal logic. Hence, we describe empirical conditions as the preconditions of a judgment in BHL.

Explicit specification of the preconditions would be useful for non-experts to prevent errors in the choice of statistical methods. Furthermore, when we formalize empirical science in future work, it would be crucial to clarify the empirical conditions that justify scientific conclusions.

8.3 Epistemic Aspects of Statistical Inference

One of our contributions is to show that epistemic logic is useful to formalize statistical inference. Although the outcome of a hypothesis test is a *knowledge* determined by the test action, it may form a false *belief*; i.e., a rejected null hypothesis may be true, and a retained one may be false. Hence the formalization of statistical inference deals with both truth and beliefs, for which epistemic logic is suitable.

The key to formalizing statistical beliefs is to introduce a Kripke semantics with a possible world w_0 where a null hypothesis is true (Section 5.2). This possible world w_0 may not be the real world where we actually apply the hypothesis test on an observed dataset.

In the Kripke model, a transition between states is used to model the update of statistical beliefs by a hypothesis test. Since the world records the executions of all tests, BHL does not allow for hiding some tests to manipulate the statistics (e.g., p-value in multiple comparison and p-value hacking).

Furthermore, the choice of two-tailed or one-tailed tests requires describing a prior belief using the possibility modality $\mathbf{P}$. Without this modality, we cannot express the belief that both lower-tail and upper-tail are possible before applying the test, since this belief may not be true.

Without this Kripke semantics, the formalization of hypothesis testing would deal with only the purely mathematical propositions satisfied in the possible world w_0 where a null hypothesis is true, hence could not reason about the appropriate usage of hypothesis tests in the real world.

9 Conclusion

In this work, we proposed a new approach to formalizing and reasoning about statistical inference in programs. Specifically, we introduced belief Hoare logic (BHL) for formalizing and checking the requirement for hypothesis tests to be employed appropriately. Then we showed that BHL is useful to reason about practical issues in statistics. We also discussed a whole picture of the justification of statistical inference. We emphasize that this is the first attempt to introduce a program logic for the appropriate application of hypothesis tests.

In ongoing and future work, we are extending our framework to other kinds of statistical methods. We plan to investigate the relative completeness of BHL and to develop a verification tool based on this framework. Another possible research would be to study justification logic for statistics.

Acknowledgments

The authors are supported by ERATO HASUO Metamathematics for Systems Design Project (No. JPMJER1603), JST. In particular, we thank Ichiro Hasuo for providing the opportunity for us to meet and collaborate in that project. Yusuke Kawamoto is supported by JST, PRESTO Grant Number JPMJPR2022, Japan, and by JSPS KAKENHI Grant Number 21K12028, Japan. Tetsuya Sato is supported by JSPS KAKENHI Grant Number 20K19775, Japan. Kohei Suenaga is supported by JST CREST Grant Number JPMJCR2012, Japan.

References

Atkinson, E., and Carbin, M. 2020. Programming and reasoning with partial observability. *Proc. ACM Program. Lang.* 4(OOPSLA):200:1–200:28.

Belle, V., and Levesque, H. J. 2015. ALLEGRO: belief-based programming in stochastic dynamical domains. In *Proc. IJCAI 2015*, 2762–2769. AAAI Press.

Bretz, F.; Hothorn, T.; and Westfall, P. 2010. *Multiple Comparisons Using R*. Chapman and Hall/CRC.

Burrows, M.; Abadi, M.; and Needham, R. M. 1990. A logic of authentication. *ACM Trans. Comput. Syst.* 8(1):18–36.

den Hartog, J., and de Vink, E. P. 2002. Verifying probabilistic programs using a Hoare like logic. *Int. J. Found. Comput. Sci.* 13(3):315–340.

Fagin, R.; Halpern, J.; Moses, Y.; and Vardi, M. 1995a. *Reasoning about Knowledge*. The MIT Press.

Fagin, R.; Halpern, J. Y.; Moses, Y.; and Vardi, M. Y. 1995b. Knowledge-based programs. In *Proc. PODC 1995*, 153–163. ACM.

Garcia, F. D.; Hasuo, I.; Pieters, W.; and van Rossum, P. 2005. Provable anonymity. In *Proc. of FMSE*, 63–72.

Halpern, J. Y. 2003. *Reasoning about uncertainty*. The MIT press.

Huber, F., and Schmidt-Petri, C. 2008. *Degrees of belief*, volume 342. Springer Science & Business Media.

Kanji, G. K. 2006. *100 statistical tests*. Sage.

Kawamoto, Y. 2019. Statistical epistemic logic. In *The Art of Modelling Computational Systems: A Journey from Logic and Concurrency to Security and Privacy*, volume 11760 of *LNCS*, 344–362. Springer.

Kawamoto, Y. 2020. An epistemic approach to the formal specification of statistical machine learning. *Software and Systems Modeling* 20(2):293–310.

Lang, T. A., and Altman, D. G. 2014. *Statistical Analyses and Methods in the Published Literature: The SAMPL Guidelines*. John Wiley & Sons, Ltd. chapter 25, 264–274.

Laverny, N., and Lang, J. 2005. From knowledge-based programs to graded belief-based programs, part I: on-line reasoning*. *Synth.* 147(2):277–321.

Nielson, H. R., and Nielson, F. 2007. *Semantics with Applications: An Appetizer (Undergraduate Topics in Computer Science)*. Berlin, Heidelberg: Springer-Verlag.

Reynolds, J. C. 2002. Separation logic: A logic for shared mutable data structures. In *Proc. LICS 2002*, 55–74. IEEE Computer Society.

Sober, E. 2008. *Evidence and evolution: The logic behind the science*. Cambridge University Press.

Suenaga, K., and Hasuo, I. 2011. Programming with infinitesimals: A while-language for hybrid system modeling. In *Proc. ICALP 2011, Part II*, volume 6756 of *LNCS*, 392–403. Springer.

Syverson, P. F., and Stubblebine, S. G. 1999. Group principals and the formalization of anonymity. In *World Congress on Formal Methods (1)*, 814–833.

von Wright, G. H. 1951. *An Essay in Modal Logic*. Amsterdam: North-Holland Pub. Co.

Wasserstein, R. L., and Lazar, N. A. 2016. The ASA statement on p-values: Context, process, and purpose. *The American Statistician* 70(2):129–133.

Winskel, G. 1993. *The Formal Semantics of Programming Languages—An Introduction*. The MIT Press.

Supplementary Materials

A Analyses of Example 2

We present more details on the reasoning about multiple comparison in Section 7.4 as follows.

Figure 5 shows the derivation tree for the judgment (3) in Example 2. It essentially represents the inference explained in Section 7.4 in the form of a derivation tree. Notice that $\Gamma \vdash \{\psi_{\mathsf{pre}}\}\ C_{12}\ \{\psi_{\mathsf{pre}} \wedge \mathbf{K}_y^{\alpha_{12}}\, \varphi_{12}\}$ is derived using TWO-Z (the instantiation of the schema TWO-T with the Z-test).

Figure 6 shows a part of the derivation tree in Figure 5. This derivation tree includes the derivation of

$$\Gamma \vdash \{\psi_{\mathsf{pre}} \wedge \mathbf{K}_y^{\alpha_{12}}\, \varphi_{12}\}\ C_{13}\ \{\mathbf{K}_y^{\alpha_{13}}\, \varphi_{13}\},$$

which is obtained by combining TWO-Z and CONSEQ. To use the postcondition $\mathbf{K}_y^{\alpha_{13}}\, \varphi_{13}$ of this judgment and the postcondition $\psi_{\mathsf{pre}} \wedge \mathbf{K}_y^{\alpha_{12}}\, \varphi_{12}$ of the judgment $\Gamma \vdash \{\psi_{\mathsf{pre}}\}\ C_{12}\ \{\psi_{\mathsf{pre}} \wedge \mathbf{K}_y^{\alpha_{12}}\, \varphi_{12}\}$ in combination, we derive $\Gamma \vdash \{\psi_{\mathsf{pre}} \wedge \mathbf{K}_y^{\alpha_{12}}\, \varphi_{12}\}\ \mathtt{skip}\ \{\mathbf{K}_y^{\alpha_{12}}\, \varphi_{12}\}$ first, which is followed by $\Gamma \vdash \{\psi_{\mathsf{pre}} \wedge \mathbf{K}_y^{\alpha_{12}}\, \varphi_{12}\}\ (\mathtt{skip} \| C_{13})\ \{\mathbf{K}_y^{\alpha_{12}}(\varphi_{12} \wedge \varphi_{13})\}$ derived by MULT-$\wedge$. By applying PAR-SKIPRML, we obtain $\Gamma \vdash \{\psi_{\mathsf{pre}} \wedge \mathbf{K}_y^{\alpha_{12}}\, \varphi_{12}\}\ C_{13}\ \{\mathbf{K}_y^{\leq\alpha}(\varphi_{12} \wedge \varphi_{13})\}$.

B Proofs for Technical Results

Next we show the proofs for all propositions in the paper.

B.1 Proofs for the Properties of Statistical Beliefs

Proposition 1 (Properties of $\mathbf{K}_{y,A}^{<\epsilon}$). *Let φ be a formula, A be a hypothesis test, $y \in \mathsf{Var}_{\mathsf{obs}}$, and $\varepsilon \in \mathbb{R}_{\geq 0}$.*

1. *Knowledge is also regarded as belief:* $\models \mathbf{K}\varphi \rightarrow \mathbf{K}_{y,A}^{<\epsilon}\, \varphi$.
2. *If we believe φ based on a test A, then we know this statistical belief; i.e.,* $\models \mathbf{K}_{y,A}^{<\epsilon}\, \varphi \rightarrow \mathbf{K}\, \mathbf{K}_{y,A}^{<\epsilon}\, \varphi$.
3. *If we failed to reject φ and think it possible, then we know this possibility; i.e.,* $\models \mathbf{P}_{y,A}^{<\epsilon}\, \varphi \rightarrow \mathbf{K}\, \mathbf{P}_{y,A}^{<\epsilon}\, \varphi$.
4. *If* $\epsilon \leq \epsilon'$, $\models \mathbf{K}_{y,A}^{<\epsilon}\, \varphi \rightarrow \mathbf{K}_{y,A}^{<\epsilon'}\, \varphi$ *and* $\models \mathbf{P}_{y,A}^{<\epsilon'}\, \varphi \rightarrow \mathbf{P}_{y,A}^{<\epsilon}\, \varphi$.

$$\dfrac{\begin{array}{c}\Gamma^{\mathsf{inv}} = \{\mu_1,\mu_2:\mathbb{R},\, x_1,x_2:\mathbb{DR}\}\cup \mathsf{Var}_{\mathsf{inv}}(\psi'),\quad \Gamma^{\mathsf{obs}}=\{\sigma_1,\sigma_2,n_1,n_2:\mathbb{N},\, y_1,y_2:\mathsf{list}\,\mathbb{R},\,\alpha:[0,1]\}\cup\mathsf{Var}_{\mathsf{obs}}(\psi'),\\ \alpha\notin\mathsf{fv}(\{\varphi_{\mathsf{L}},\varphi_{\mathsf{U}}\}),\ \ \psi':\tau^{\bowtie}\text{-free},\ \ \psi \stackrel{\text{def}}{=} (x_1\approx N(\mu_1,\sigma_1^2)\wedge x_2\approx N(\mu_2,\sigma_2^2)\wedge y_1 \underset{n_1}{\leftsquigarrow} x_1\wedge y_2 \underset{n_2}{\leftsquigarrow} x_2\wedge(\sigma_1=\sigma_2)\wedge\psi')\end{array}}{(\Gamma^{\mathsf{inv}},\Gamma^{\mathsf{obs}})\vdash\{\psi[\alpha\mapsto f_{A^{(\mathsf{T})}_{\varphi_0}}(y)]\wedge\mathbf{P}\varphi_{\mathsf{L}}\wedge\mathbf{P}\varphi_{\mathsf{U}}\}\ \alpha:=f_{A^{(\mathsf{T})}_{\varphi_0}}(y)\ \{\psi\wedge\mathbf{K}^{\alpha}_{y,A}\,\varphi_{\mathsf{T}}\}}\ (\textsc{Two-}Z)$$

Figure 4: The axiom for two-tailed Z-test for two population means (an instance of the schema TWO-T).

$$\dfrac{\dfrac{\vdots}{\Gamma\vdash\{\psi_{\mathsf{pre}}\}\,C_{12}\,\{\psi_{\mathsf{pre}}\wedge\mathbf{K}^{\alpha_{12}}_y\,\varphi_{12}\}}\quad \dfrac{\dfrac{\begin{array}{c}\dfrac{\vdots}{\Gamma\vdash\{\psi_{\mathsf{pre}}\wedge\mathbf{K}^{\alpha_{12}}_y\,\varphi_{12}\}\,C_{13}\,\{\varphi^{\mathsf{bel}}\}}\\ \Gamma\models(\psi_{\mathsf{pre}}\wedge\mathbf{K}^{\alpha_{12}}_y\,\varphi_{12}\wedge\alpha_{12}{\le}0.05)\to(\psi_{\mathsf{pre}}\wedge\mathbf{K}^{\alpha_{12}}_y\,\varphi_{12})\\ \Gamma\models\varphi^{\mathsf{bel}}\to(\varphi^{\mathsf{bel}}_{12}\to\varphi^{\mathsf{bel}})\end{array}}{\Gamma\vdash\{{}^{\psi_{\mathsf{pre}}\wedge\mathbf{K}^{\alpha_{12}}_y\,\varphi_{12}}_{\wedge\alpha_{12}\le0.05}\}\,C_{13}\,\{\varphi^{\mathsf{bel}}_{12}\to\varphi^{\mathsf{bel}}\}}\quad\dfrac{\begin{array}{c}\Gamma\models(\psi_{\mathsf{pre}}\wedge\mathbf{K}^{\alpha_{12}}_y\,\varphi_{12}\wedge\alpha_{12}{>}0.05)\to\neg\varphi^{\mathsf{bel}}_{12}\\ \Gamma\vdash\{\neg\varphi^{\mathsf{bel}}_{12}\}\ \mathtt{skip}\ \{\neg\varphi^{\mathsf{bel}}_{12}\}\\ \Gamma\models\neg\varphi^{\mathsf{bel}}_{12}\to(\varphi^{\mathsf{bel}}_{12}\to\varphi^{\mathsf{bel}})\end{array}}{\Gamma\vdash\{{}^{\psi_{\mathsf{pre}}\wedge\mathbf{K}^{\alpha_{12}}_y\,\varphi_{12}}_{\wedge\alpha_{12}>0.05}\}\ \mathtt{skip}\ \{\varphi^{\mathsf{bel}}_{12}\to\varphi^{\mathsf{bel}}\}}}{\Gamma\vdash\{\psi_{\mathsf{pre}}\wedge\mathbf{K}^{\alpha_{12}}_y\,\varphi_{12}\}\ \mathtt{if}\ \alpha_{12}{\le}0.05\ \mathtt{then}\ C_{13}\ \mathtt{else\ skip}\ \{\varphi^{\mathsf{bel}}_{12}\to\varphi^{\mathsf{bel}}\}}}{\Gamma\vdash\{\psi_{\mathsf{pre}}\}\ C_{12};\mathtt{if}\ \alpha_{12}{\le}0.05\ \mathtt{then}\ C_{13}\ \mathtt{else\ skip}\ \{\varphi^{\mathsf{bel}}_{12}\to\varphi^{\mathsf{bel}}\}}\ \textsc{Seq}$$

Figure 5: An outline of the proof for an illustrating program in Example 2 where $\varphi^{\mathsf{bel}}_{12}\stackrel{\text{def}}{=}\mathbf{K}^{\le 0.05}_y\,\varphi_{12}$, $\varphi^{\mathsf{bel}}_{13}\stackrel{\text{def}}{=}\mathbf{K}^{\le\alpha_{13}}_y\,\varphi_{13}$, $\alpha\stackrel{\text{def}}{=}\min(\alpha_{12},\alpha_{13})$, $\varphi^{\mathsf{bel}}\stackrel{\text{def}}{=}\mathbf{K}^{\le\alpha}_y(\varphi_{12}\wedge\varphi_{13})$, and $\varphi\stackrel{\text{def}}{=}(\varphi^{\mathsf{bel}}_{12}\to\varphi^{\mathsf{bel}})$. We first execute $C_{12}\stackrel{\text{def}}{=}(\alpha_{12}:=f_{Z_{\neg\varphi_{12}}}(y))$ for the Z-test with the alternative hypothesis $\varphi_{12}\stackrel{\text{def}}{=}(\mathsf{mean}(x_1)\neq\mathsf{mean}(x_2))$. If $\alpha_{12}<0.05$, then we execute $C_{13}\stackrel{\text{def}}{=}(\alpha_{13}:=f_{Z_{\neg\varphi_{13}}}(y))$ for the Z-test with the alternative hypothesis $\varphi_{13}\stackrel{\text{def}}{=}(\mathsf{mean}(x_1)\neq\mathsf{mean}(x_3))$, and report the p-value α for the whole test.

$$\dfrac{\dfrac{\begin{array}{c}\Gamma\models(\psi_{\mathsf{pre}}\wedge\mathbf{K}^{\alpha_{12}}_y\,\varphi_{12})\to\mathbf{K}^{\alpha_{12}}_y\,\varphi_{12}\\ \Gamma\vdash\{\mathbf{K}^{\alpha_{12}}_y\,\varphi_{12}\}\ \mathtt{skip}\ \{\mathbf{K}^{\alpha_{12}}_y\,\varphi_{12}\}\end{array}}{\Gamma\vdash\{\psi_{\mathsf{pre}}\wedge\mathbf{K}^{\alpha_{12}}_y\,\varphi_{12}\}\ \mathtt{skip}\ \{\mathbf{K}^{\alpha_{12}}_y\,\varphi_{12}\}}\quad\dfrac{\Gamma\models\psi_{\mathsf{pre}}\wedge\mathbf{K}^{\alpha_{12}}_y\,\varphi_{12}\to\psi_{\mathsf{pre}}\quad\dfrac{}{\Gamma\vdash\{\psi_{\mathsf{pre}}\}\,C_{13}\,\{\mathbf{K}^{\alpha_{13}}_y\,\varphi_{13}\}}\,\textsc{Two-}Z}{\Gamma\vdash\{\psi_{\mathsf{pre}}\wedge\mathbf{K}^{\alpha_{12}}_y\,\varphi_{12}\}\,C_{13}\,\{\mathbf{K}^{\alpha_{13}}_y\,\varphi_{13}\}}\,\textsc{Conseq}}{\dfrac{\Gamma\vdash\{\psi_{\mathsf{pre}}\wedge\mathbf{K}^{\alpha_{12}}_y\,\varphi_{12}\}\ (\mathtt{skip}\|C_{13})\ \{\mathbf{K}^{\le\min(\alpha_{12},\alpha_{13})}_y(\varphi_{12}\wedge\varphi_{13})\}}{\Gamma\vdash\{\psi_{\mathsf{pre}}\wedge\mathbf{K}^{\alpha_{12}}_y\,\varphi_{12}\}\ C_{13}\ \{\mathbf{K}^{\le\alpha}_y(\varphi_{12}\wedge\varphi_{13})\}}}\,\textsc{Mult-}\wedge$$

Figure 6: A part of the derivation of Figure 5.

5. *$\mathbf{K}^{<\epsilon}_{y,A}\,\varphi$ may represent a false belief. The alternative hypothesis φ we believe may be false, i.e., the rejected null hypothesis $\neg\varphi$ may be true: $\epsilon>0$ iff $\not\models\mathbf{K}^{<\epsilon}_{y,A}\,\varphi\to\varphi$.*
6. *$\models\mathbf{K}^{<\epsilon}_{y,A}\,\varphi\to\mathbf{K}^{<\epsilon}_y\,\varphi$.*

Proof. 1. Let w be a world such that $w\models\mathbf{K}\varphi$. Since $\mathbf{K}^{\bowtie\epsilon}_{y,A}\,\varphi$ is defined by $\mathbf{K}(\varphi\vee\tau^{\bowtie\epsilon}_{A_{\neg\varphi}}(y))$, we obtain $w\models\mathbf{K}^{<\epsilon}_{y,A}\,\varphi$. Therefore $\models\mathbf{K}\varphi\to\mathbf{K}^{<\epsilon}_{y,A}\,\varphi$.

2. Let w be a world such that $w\models\mathbf{K}^{<\epsilon}_{y,A}\,\varphi$. Then $w\models\mathbf{K}(\varphi\vee\tau^{<\epsilon}_{A_{\neg\varphi}}(y))$. By the axiom (5) of $\mathbf{K}$, we obtain $w\models\mathbf{K}\mathbf{K}(\varphi\vee\tau^{<\epsilon}_{A_{\neg\varphi}}(y))$, hence $w\models\mathbf{K}\,\mathbf{K}^{<\epsilon}_{y,A}\,\varphi$.

3. Let w be a world such that $w\models\mathbf{P}^{<\epsilon}_{y,A}\,\varphi$. Then $w\models\mathbf{P}(\varphi\wedge\neg\tau^{<\epsilon}_{A_{\varphi}}(y))$. By the axiom (4) of $\mathbf{K}$, we obtain $w\models\mathbf{K}\mathbf{P}(\varphi\wedge\neg\tau^{<\epsilon}_{A_{\varphi}}(y))$, and thus $w\models\mathbf{K}\,\mathbf{P}^{<\epsilon}_{y,A}\,\varphi$.

4. Assume that $\epsilon\le\epsilon'$. Then we obtain the following formulas on the strength of confidence levels: $\models\tau^{<\epsilon}_{A_{\neg\varphi}}(y)\to\tau^{<\epsilon'}_{A_{\neg\varphi}}(y)$. By definition, we obtain $\models\mathbf{K}^{<\epsilon}_{y,A}\,\varphi\to\mathbf{K}^{<\epsilon'}_{y,A}\,\varphi$ and $\models\mathbf{P}^{<\epsilon'}_{y,A}\,\varphi\to\mathbf{P}^{<\epsilon}_{y,A}\,\varphi$.

5. The claim is immediate from the definition of $\mathbf{K}^{<\epsilon}_{y,A}$.

6. The claim is immediate from the definitions of $\mathbf{K}^{<\epsilon}_{y,A}$ and $\mathbf{K}^{<\epsilon}_y$.

$\square$

B.2 Remark on Parallel Compositions

In this subsection, we show for parallel composition $C_1\|C_2$, the world $w'\in[\![C_1\|C_2]\!](w)$ is convertible to a pair of $w_1\in[\![C_1]\!](w)$ and $w_2\in[\![C_1]\!](w)$ and vise versa.

Recall that we imposed the restriction $\mathsf{upd}(C_b)\cap\mathsf{Var}(C_{3-b})=\emptyset$ for $b=1,2$. For any possible world w, we may assume $\langle C_b,w\rangle\longrightarrow^* w;u_b$ for $b=1,2$, and write

$$u_b=(m^b[0],a^b[0]),\cdots,(m^b[n_b],a^b[n_b])$$

where $n_b=\mathsf{len}(w_b)-1$.

Since both sequences u_1 and u_2 of states are generated in executions of C_1 and C_2, we can write $m^b[l]=m\rho^b_l$ with substitutions $\rho^b_l=[v_1\mapsto k^l_1,\ldots,v_s\mapsto k^l_s]$ of variables where (m,a) is the last state of the possible world w. Then, we can take

$$u'{=}u_1;(m^1[n_1]\rho^2_0,a^2[0]),\ldots,(m^1[n_1]\rho^2_{n_2},a^2[n_2]).$$

Lemma 1. *We have $\langle C_1\|C_2,w\rangle\longrightarrow^* w;u'$.*

Proof. First, from the assumption and inference rules on $C_1 \| C_2$, we obtain $\langle C_1 \| C_2, w\rangle \longrightarrow^* \langle C_2, w; u_1\rangle$. We next show $\langle C_2, w; u_1\rangle \longrightarrow^* w; u'$ by induction on C_2.

- If C_2 is $\texttt{skip}$ then $u_2 = (m, \{\texttt{skip}\})$. We obtain $\langle C_2, w; u_1\rangle \longrightarrow^* w; u_1; (m^1[n_1], \{\texttt{skip}\})$.
- If C_2 is $v := e$ then $u_2 = (m[v \mapsto [\![e]\!]_m], \{v := e\})$. Since $\mathsf{upd}(C_2) \cap \mathsf{Var}(C_1) = \emptyset$ for $b = 1, 2$, we have $m^1[n_1][v \mapsto [\![e]\!]_m] = m^1[n_1][v \mapsto [\![e]\!]_{m^1[n_1]}]$. Hence,

$$\begin{aligned}&\langle C_2, w; u_1\rangle \\ &\longrightarrow^* w; u_1; (m^1[n_1][v \mapsto [\![e]\!]_{m^1[n_1]}], \{v := e\}).\end{aligned}$$

- If C_2 is a sequential composition $C_3; C_4$, we can decompose $u_2 = u_3; u_4$ such that $\langle C_3; C_4, w\rangle \longrightarrow^* \langle C_4, w; u_3\rangle$ and $\langle C_4, w; u_3\rangle \longrightarrow^* w; u_3; u_4$ (by a simple induction, we can prove it). By induction hypothesis, we have $\langle C_2, w; u_1\rangle \longrightarrow^* w; u'$.
- In other cases, we immediately have the desired statement by induction hypothesis.

□

We check the converse. Assume $\langle C_1 \| C_2, w\rangle \longrightarrow^* w; u'$. We write

$$u' = (m'[0], a'[0]), \cdots, (m'[n'], a'[n'])$$

where $n' = \mathsf{len}(u') - 1$. Since u' is generated by executions of $C_1 \| C_2$, we can write $m'[0] = m\rho'_0$ and $m'[l+1] = m[l]\rho'_{l+1}$ with substitutions $\rho'_l = [v_1 \mapsto k^l_1, \ldots, v_s \mapsto k^l_s]$ of variables. Since $\mathsf{upd}(C_b) \cap \mathsf{Var}(C_{3-b}) = \emptyset$ for $b = 1, 2$, each ρ'_l changes variables in *either* $\mathsf{upd}(C_1)$ or $\mathsf{upd}(C_2)$.

We decompose $\{0, 1, \ldots, n'\}$ into two sub-sequences L_1 and L_2 such that $l \in L_b$ if and only if the substitution ρ'_l changes only variables in $\mathsf{upd}(C_b)$. For $b = 1, 2$, we define:

$$\begin{aligned}u_b = {}&(m\rho'_{L_b[0]}, a_{L_b[0]}), (m\rho'_{L_b[0]}\rho'_{L_b[1]}, a_{L_b[1]}), \\ &\ldots, (m\rho'_{L_b[0]} \cdots \rho'_{L_b[\mathsf{len}(L_b)-1]}, a_{L_b[\mathsf{len}(L_b)-1]})\end{aligned}$$

Lemma 2. *We have $\langle C_b, w\rangle \longrightarrow^* w; u_b$ for $b = 1, 2$.*

Proof. We show the lemma in the case of $b = 1$ by induction on $\mathsf{len}(L_1)$ as follows. (The case of $b = 2$ can be derived similarly.)

If $\mathsf{len}(L_1) = 0$, then there is no C_1 with $\langle C_1 \| C_2, w\rangle \longrightarrow^* w; u'$ because each terminating program adds at least a state to possible world.

If $\mathsf{len}(L_1) = k + 1$, then we can take the first pair $(m\rho'_{L_1[0]}, a_{L_1[0]})$ of L_1. If this is also the first pair of u', we have one of:

$$\begin{aligned}&\langle C_1 \| C_2, w\rangle \longrightarrow \langle C'_1 \| C_2, w; (m\rho'_{L_1[0]}, a_{L_1[0]})\rangle \\ &\qquad \text{with } \langle C_1 w\rangle \longrightarrow \langle C'_1 w; (m\rho'_{L_1[0]}, a_{L_1[0]})\rangle \qquad \text{(A)} \\ &\langle C_1 \| C_2, w\rangle \longrightarrow \langle C_2, w; (m\rho'_{L_1[0]}, a_{L_1[0]})\rangle \\ &\qquad \text{with } \langle C_1 w\rangle \longrightarrow w; (m\rho'_{L_1[0]}, a_{L_1[0]}) \qquad \text{(B)}\end{aligned}$$

If (A) holds, by induction hypothesis, we obtain:

$$\begin{aligned}&\langle C'_1, w; (m\rho'_{L_1[0]}, a_{L_1[0]})\rangle \\ &\longrightarrow^* w; (m\rho'_{L_1[0]}, a_{L_1[0]}), \\ &\qquad (m\rho'_{L_1[1]}, a_{L_1[1]}), \ldots, (m\rho'_{L_1[\mathsf{len}(L_1)]}, a_{L_1[\mathsf{len}(L_1)]}) \\ &= w; u_1\end{aligned}$$

If (B) holds, we already have $\langle C_1, w\rangle \longrightarrow^* w; u_1$.

If $(m\rho'_{L_1[0]}, a_{L_1[0]})$ is not the first state, then u' can be decomposed into $u_4; (m'[L_1[0]-1]\rho'_{L_1[0]}, a_{L_1[0]}); u_5$ where u_4 is a nonempty sub-sequence of L_2. By similar discussion as above, we have one of:

$$\begin{aligned}&\langle C_1 \| C_2, w\rangle \longrightarrow^* \langle C_1 \| C'_2, w; u_4\rangle \\ &\qquad \text{with } \langle C_2, w\rangle \longrightarrow^* \langle C'_2, w; u_4\rangle \qquad \text{(A')} \\ &\langle C_1 \| C_2, w\rangle \longrightarrow^* \langle C_1, w; u_4\rangle \\ &\qquad \text{with } \langle C_2, w\rangle \longrightarrow^* w; u_4 \qquad \text{(B')}\end{aligned}$$

In both case, by replacing w with $w; u_4$, we can apply the case that $(m\rho'_{L_1[0]}, a_{L_1[0]})$ is the first state. □

Lemma 3. *For any possible world $w' \in [\![C_1 \| C_2]\!](w)$, there is a pair of worlds $w_1 \in [\![C_1]\!](w)$ and $w_2 \in [\![C_2]\!](w)$ such that for any pair of $\tau^{\bowtie}$-free assertions φ_1 and φ_2 satisfying $\mathsf{upd}(C_b) \cap \mathsf{fv}(\varphi_{3-b}) = \emptyset$ for $b = 1, 2$, we have:*

$$w' \models \varphi_1 \land \varphi_2 \quad \textit{iff} \quad w_1 \models \varphi_1 \textit{ and } w_2 \models \varphi_2.$$

Proof. Let $w' \in [\![C_1 \| C_2]\!](w)$. By the definition of semantics, we have $w' = w; u'$ for some u'. By Lemma 2, we can decompose u' into u_1 and u_2 such that $w; u_1 \in [\![C_1]\!](w)$ and $w; u_2 \in [\![C_2]\!](w)$. Remark that for $b = 1, 2$, all updates in u' done by C_b occur in u_b.

For each $b = 1, 2$, by the assumption $\mathsf{upd}(C_{3-b}) \cap \mathsf{fv}(\varphi_b) = \emptyset$ and the $\tau^{\bowtie}$-freeness of φ_b, we have:

$$w' \models \varphi_b \quad \text{iff} \quad w; u_b \models \varphi_b.$$

Otherwise, φ_b depends on variables in $\mathsf{upd}(C_{3-b})$, and this contradicts the assumption.

By letting $w_b = w; u_b$ for $b = 1, 2$, we conclude:

$$w' \models \varphi_1 \land \varphi_2 \quad \text{iff} \quad w_1 \models \varphi_1 \text{ and } w_2 \models \varphi_2.$$

□

B.3 Proof for the Soundness of BHL

We present a proof for the soundness of belief Hoare logic (Theorem 1). We obtain the validity of the axioms and rules for basic constructs in Figure 2 as usual, except for PAR-SKIPADDL, PAR-SKIPADDR, PAR-SKIPRML and PAR-SKIPRMR. Thus, to prove the soundness, it is sufficient to show the validity of these four skip rules, and the axiom schemas and rules specific to hypothesis tests in Figure 3 as follows.

Validity of Par-SkipAddL and Par-SkipAddR We first show the validity of the rule Par-SkipAddL from Lemma 3. (The validity of Par-SkipAddR can be proven in a similar way.)

$$\frac{\Gamma \vdash \{\psi\}\, C\, \{\varphi\}}{\Gamma \vdash \{\psi\}\, \mathtt{skip} \| C\, \{\varphi\}} \quad (\textsc{Par-SkipAddL})$$

Proposition 2 (Validity of Par-SkipAddL)**.** *The rule* Par-SkipAddL *is valid.*

Proof. By the rules (Skip) and (Conseq), we have $\Gamma \vdash \{\psi\}\ \mathtt{skip}\ \{\top\}$. Let w be a possible world such that $w \models \psi$. Let $w' \in [\![\mathtt{skip} \| C]\!](w)$. Since $\top$ has no variable, by Lemma 3, there exist $w'_1 \in [\![\mathtt{skip}]\!](w)$ and $w'_2 \in [\![C]\!](w)$ such that:

$$w' \models \varphi \text{ iff } w'_2 \models \varphi.$$

By the premise of the rule, we have $w'_2 \models \varphi$. Therefore $w' \models \varphi$. □

Validity of Par-SkipRmL and Par-SkipRmR Next we show the validity of the rule Par-SkipRmL from Lemma 3. (The validity of Par-SkipRmR can be proven in a similar way.)

$$\frac{\Gamma \vdash \{\psi\}\, \mathtt{skip} \| C\, \{\varphi\}}{\Gamma \vdash \{\psi\}\, C\, \{\varphi\}} \quad (\textsc{Par-SkipRmL})$$

Proposition 3 (Validity of Par-SkipRmL)**.** *The rule* Par-SkipRmL *is valid.*

Proof. Let w be a possible world such that $w \models \psi$. Let $w' \in [\![C]\!](w)$, that is, $\langle C, w\rangle \longrightarrow^* w'$. This implies that

$$\langle \mathtt{skip} \| C, w\rangle \longrightarrow^* \langle \mathtt{skip}, w'\rangle \longrightarrow w'; (m[\mathsf{len}(w')-1], \mathtt{skip}).$$

By the premise of the rule, we have $w'; (m[\mathsf{len}(w') - 1], \mathtt{skip}) \models \varphi$.

Here, we have $w'; (m[\mathsf{len}(w')-1], \mathtt{skip}) \in [\![\mathtt{skip}]\!](w')$. By the rule (Skip), we have $\Gamma \vdash \{\neg\varphi\}\ \mathtt{skip}\ \{\neg\varphi\}$. Hence,

$$\text{if } w' \not\models \varphi \text{ then } w'; (m[\mathsf{len}(w') - 1], \mathtt{skip}) \not\models \varphi.$$

Since $w'; (m[\mathsf{len}(w') - 1], \mathtt{skip}) \models \varphi$, we have $w' \models \varphi$. □

Validity of axiom schemas and rules for hypothesis tests Next we present the proofs for the validity of the axiom schemas and rules specific to hypothesis tests as follows.

Proposition 4 (Validity of Two-T, Low-T, Up-T)**.** *The schemas* Two-T*,* Low-T*, and* Up-T *are valid.*

Proof. We show the validity of the schema Two-T for the two-tailed test as follows. We denote the precondition by $\psi_{\mathsf{pre}} \stackrel{\text{def}}{=} (x \approx P(\xi, \theta) \wedge y \leftarrow_n x \wedge \psi'[\alpha \mapsto f_{A^{(\mathsf{T})}_{\varphi_0}}(y)] \wedge \mathbf{P}\varphi_{\mathsf{L}} \wedge \mathbf{P}\varphi_{\mathsf{U}})$. Let w be a possible world such that $w \models \psi_{\mathsf{pre}}$. For unknown parameters ξ and known parameters θ, $P(\xi, \theta)$ is the statistical model for the population $w(x)$. In the world w, we have sampled a dataset $w(y)$ from $w(x)$. By $w \models \mathbf{P}\varphi_{\mathsf{L}} \wedge \mathbf{P}\varphi_{\mathsf{U}}$, the hypothesis test should be two-tailed. Let $w' = [\![\alpha := f_{A^{(\mathsf{T})}_{\varphi_0}}(y)]\!](w)$ where $f_{A^{(\mathsf{T})}_{\varphi_0}}$ is the program that executes the two-tailed test $A^{(s)}_{\varphi_0} = (\varphi_0, t, D_{t_\theta,\varphi_0}, s, P(\xi, \theta))$ with the null hypothesis φ_0. By Equation (5), we obtain:

$$\alpha = \Pr_{d \sim D_{t_\theta,\varphi_0}} [\, d \preccurlyeq^{(\mathsf{T})}_{t_\theta} t_\theta(w(y)) \,].$$

By the semantics for $\mathbf{K}^{\alpha}_{y,A}$, we obtain $w' \models \mathbf{K}^{\alpha}_{y,A}\, \varphi_{\mathsf{T}}$. Since only the variable α is updated by the hypothesis test command, we have $w' \models x \approx P(\xi; \theta) \wedge y \leftarrow_n x \wedge \psi'$. Therefore the judgment $\Gamma \vdash \{\psi_{\mathsf{pre}}\}\ \alpha := f_{A^{(\mathsf{T})}_{\varphi_0}}(y)\ \{x \approx P(\xi; \theta) \wedge y \leftarrow_n x \wedge \psi' \wedge \mathbf{K}^{\alpha}_{y,A}\, \varphi_{\mathsf{T}}\}$ is valid.

The validity of Low-T and Up-T are proven analogously. □

Proposition 5 (Validity of Mult-$\vee$)**.** Mult-$\vee$ *is valid.*

Proof. Let $\Gamma^{\mathsf{inv}} = \Gamma^{\mathsf{inv}}_1 \cup \Gamma^{\mathsf{inv}}_2$, $\Gamma^{\mathsf{obs}} = \Gamma^{\mathsf{obs}}_1 \cup \Gamma^{\mathsf{obs}}_2$. Let w be a possible world such that $w \models \psi_1 \wedge \psi_2$.

Let $w' \in [\![C_1 \| C_2]\!](w)$. By $\mathsf{upd}(C_1) \cap \mathsf{Var}(C_2) = \mathsf{upd}(C_2) \cap \mathsf{Var}(C_1) = \emptyset$, the two programs C_1 and C_2 do not update common variables in $\mathsf{Var}(C_1) \cap \mathsf{Var}(C_2)$. Then, by Lemma 2, the world w' can be split into w_1 and w_2 such that $w_1 \in [\![C_1]\!](w)$ and $w_2 \in [\![C_1]\!](w)$.

By the premises of Mult-$\vee$, we have $w_1 \models \psi'_1$ and $w_2 \models \psi'_2$. Since $\mathsf{upd}(C_b) \cap \mathsf{fv}(\psi'_{3-b}) = \emptyset$ holds for $i = 1, 2$ and ψ'_1 and ψ'_2 are $\tau^{\bowtie}$-free, it follows from Lemma 3 that $w' \models \psi'_1 \wedge \psi'_2$.

Finally, we show $w' \models \mathbf{K}^{\leq \alpha_1+\alpha_2}_{(y_1,y_2)}(\varphi_1 \vee \varphi_2)$ as follows. For $i = 1, 2$, let $A_{\varphi_{0i}} = (\varphi_{0i}, t_i, D_{t_i,\varphi_{0i}}, \preccurlyeq^{(s_i)}_{t_i})$ be the hypothesis test executed in C_i with a null hypothesis φ_{0i} and an alternative hypothesis φ_i. For each $i = 1, 2$, by the premises of Mult-$\vee$, we have $w_i \models \mathbf{K}^{\alpha_i}_{y_i}\, \varphi_i$, that is, we have the statistical belief that the alternative hypothesis φ_i is true with the significance level α_i; i.e.,

$$\Pr_{d \sim D_{t_i,\varphi_{0i}}} [\, d \preccurlyeq^{(s_i)}_{t_i} t_i(w_i(y_i)) \,] = \alpha_i. \tag{6}$$

By $\mathsf{upd}(C_i) \cap \{y_1, y_2\} = \emptyset$ for $i = 1, 2$, the programs C_1 and C_2 do not change the datasets y_1 or y_2, hence:

$$w'(y_1) = w_1(y_1) \text{ and } w'(y_2) = w_2(y_2). \tag{7}$$

Recall that in Section 4.4, the disjunctive combination is defined by $A_{\varphi_{01} \vee \varphi_{02}} = (\varphi_{01} \vee \varphi_{02}, t, D, \preccurlyeq^{(s_1,s_2)}_t)$ where $t(y_1, y_2) = (t_1(y_1), t_2(y_2))$, $D = D_{t_1,\varphi_{01}} \times D_{t_2,\varphi_{02}}$, and $(d_1, d_2) \preccurlyeq^{(s_1,s_2)}_t (d'_1, d'_2)$ iff either $d_1 \preccurlyeq^{(s_1)}_{t_1} d'_1$ or $d_2 \preccurlyeq^{(s_2)}_{t_2} d'_2$.

Then we obtain:

$$\Pr_{(d_1,d_2)\sim D}[\,(d_1,d_2) \preccurlyeq_t^{(s_1,s_2)}\ t(w'(y_1), w'(y_2))\,]$$
(by the definition of $\preccurlyeq_t^{(s_1,s_2)}$)
$$= \Pr_{(d_1,d_2)\sim D}[\,d_1 \preccurlyeq_{t_1}^{(s_1)}\ t_1(w'(y_1)) \vee d_2 \preccurlyeq_{t_2}^{(s_2)}\ t_2(w'(y_2))\,]$$
$$\leq \Pr_{d_1\sim D_{t_1,\varphi_{01}}}[\,d_1 \preccurlyeq_{t_1}^{(s_1)}\ t_1(w'(y_1))\,] + \Pr_{d_2\sim D_{t_2,\varphi_{02}}}[\,d_2 \preccurlyeq_{t_2}^{(s_2)}\ t_2(w'(y_2))\,]$$
$$= \Pr_{d_1\sim D_{t_1,\varphi_{01}}}[\,d_1 \preccurlyeq_{t_1}^{(s_1)}\ t_1(w_1(y_1))\,] + \Pr_{d_2\sim D_{t_2,\varphi_{02}}}[\,d_2 \preccurlyeq_{t_2}^{(s_2)}\ t_2(w_2(y_2))\,]$$
(by Equation (7))
$$= \alpha_1 + \alpha_2$$
(by Equation (6)).

Hence $w' \models \mathbf{K}^{\leq \alpha_1+\alpha_2}_{(y_1,y_2)}(\varphi_1 \vee \varphi_2)$.

Therefore MULT-$\vee$ is valid. $\square$

Proposition 6 (Validity of MULT-$\wedge$)**.** MULT-$\wedge$ *is valid.*

Proof. Let $\Gamma^{\mathsf{inv}} = \Gamma_1^{\mathsf{inv}} \cup \Gamma_2^{\mathsf{inv}}$, $\Gamma^{\mathsf{obs}} = \Gamma_1^{\mathsf{obs}} \cup \Gamma_2^{\mathsf{obs}}$. Let w be a possible world such that $w \models \psi_1 \wedge \psi_2$.

Let $w' \in [\![C_1 \| C_2]\!](w)$. By $\mathsf{upd}(C_1) \cap \mathsf{Var}(C_2) = \mathsf{upd}(C_2) \cap \mathsf{Var}(C_1) = \emptyset$, the two programs C_1 and C_2 do not update common variables in $\mathsf{Var}(C_1) \cap \mathsf{Var}(C_2)$. Then, by Lemma 2, the world w' can be split into w_1 and w_2 such that $w_1 \in [\![C_1]\!](w)$ and $w_2 \in [\![C_1]\!](w)$.

By the premises of MULT-$\wedge$, we have $w_1 \models \psi'_1$ and $w_2 \models \psi'_2$. Since $\mathsf{upd}(C_b) \cap \mathsf{fv}(\psi'_{3-b}) = \emptyset$ holds for $i = 1, 2$ and ψ'_1 and ψ'_2 are $\tau^{\bowtie}$-free, it follows from Lemma 3 that: $w' \models \psi'_1 \wedge \psi'_2$.

Finally, we show $w' \models \mathbf{K}^{\leq \min(\alpha_1,\alpha_2)}_{(y_1,y_2)}(\varphi_1 \wedge \varphi_2)$ as follows. For $i = 1, 2$, let $A_{\varphi_{0i}} = (\varphi_{0i}, t_i, D_{t_i,\varphi_{0i}}, \preccurlyeq_{t_i}^{(s_i)})$ be the hypothesis test executed in C_i with a null hypothesis φ_{0i} and an alternative hypothesis φ_i. For each $i = 1, 2$, by the premises of MULT-$\wedge$, we have $w_i \models \mathbf{K}^{\alpha_i}_{y_i}\,\varphi_i$, that is, we have the statistical belief that the alternative hypothesis φ_i is true with the significance level α_i; i.e.,

$$\Pr_{d\sim D_{t_i,\varphi_{0i}}}[\,d \preccurlyeq_{t_i}^{(s_i)}\ t_i(w_i(y_i))\,] = \alpha_i. \tag{8}$$

By $y_1 \notin \mathsf{upd}(C_1)$, the programs C_1 does not change the datasets y_1:

$$w'(y_1) = w_1(y_1). \tag{9}$$

Recall that in Section 4.4, the conjunctive combination is defined by $A_{\varphi_{01}\wedge\varphi_{02}} = (\varphi_{01} \wedge \varphi_{02}, t, D, \preccurlyeq_t^{(s_1,s_2)})$ where $t(y_1, y_2) = (t_1(y_1), t_2(y_2))$, $D = D_{t_1,\varphi_{01}} \times D_{t_2,\varphi_{02}}$, and $(d_1, d_2) \preccurlyeq_t^{(s_1,s_2)} (d'_1, d'_2)$ iff $d_1 \preccurlyeq_{t_1}^{(s_1)} d'_1$ and $d_2 \preccurlyeq_{t_2}^{(s_2)} d'_2$.

Then we obtain:

$$\varepsilon \stackrel{\text{def}}{=} \Pr_{(d_1,d_2)\sim D}[\,(d_1,d_2) \preccurlyeq_t^{(s_1,s_2)}\ t(w'(y_1), w'(y_2))\,]$$
(by the definition of $\preccurlyeq_t^{(s_1,s_2)}$)
$$= \Pr_{(d_1,d_2)\sim D}[\,d_1 \preccurlyeq_{t_1}^{(s_1)}\ t_1(w'(y_1)) \wedge d_2 \preccurlyeq_{t_2}^{(s_2)}\ t_2(w'(y_2))\,]$$
$$\leq \Pr_{(d_1,d_2)\sim D}[\,d_1 \preccurlyeq_{t_1}^{(s_1)}\ t_1(w'(y_1))\,]$$
$$= \Pr_{(d_1,d_2)\sim D}[\,d_1 \preccurlyeq_{t_1}^{(s_1)}\ t_1(w_1(y_1))\,] \quad \text{(by Equation (9))}$$
$$= \alpha_1 \quad \text{(by Equation (8)).}$$

A similar inequality holds for α_2. Thus $\varepsilon \leq \min(\alpha_1, \alpha_2)$. Hence $w' \models \mathbf{K}^{\leq \min(\alpha_1,\alpha_2)}_{(y_1,y_2)}(\varphi_1 \wedge \varphi_2)$.

Therefore MULT-$\wedge$ is valid. $\square$