UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Nonparametric pragmatic hypothesis testing

Rodrigo Ferrari Lucas Lassance

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Rodrigo Ferrari Lucas Lassance

Nonparametric pragmatic hypothesis testing

Dissertation submitted to the Institute of Mathematics and Computer Science – ICMC-USP and to the Department of Statistics – DEs-UFSCar – in accordance with the requirements of the Statistics Interagency Graduate Program, for the degree of Master in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Rafael Bassi Stern

USP – São Carlos
July 2022

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

L346n

Lassance, Rodrigo Ferrari Lucas
Nonparametric pragmatic hypothesis testing / Rodrigo Ferrari Lucas Lassance; orientador Rafael Bassi Stern. -- São Carlos, 2022.
71 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022.

1. hipóteses pragmáticas. 2. função de dissimilaridade. 3. testes agnósticos. 4. bayesiana não-paramétrica. I. Bassi Stern, Rafael, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Rodrigo Ferrari Lucas Lassance

Testagem não-paramétrica de hipóteses pragmáticas

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Rafael Bassi Stern

USP – São Carlos
Julho de 2022

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Rodrigo Ferrari Lucas Lassance, realizada em 13/06/2022.

Comissão Julgadora:

Prof. Dr. Rafael Bassi Stern (UFSCar)

Prof. Dr. Dani Gamerman (UFRJ)

Prof. Dr. Luis Gustavo Esteves (USP)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

This work is dedicated to those willing to stick their necks out by trying to come up with better solutions to old problems, the ones people would rather hide under their carpets.
In particular, to all researchers of agnostic tests.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my parents, Meri and Antonio, for all the support they gave me throughout my life, and Olivia, my girlfriend, for a much needed emotional support and for being the best company I could ever have asked for.

I also thank my advisor, Dr. Rafael Bassi Stern, for the opportunity of working on this amazing project and for being such a reasonable and understanding person, while also challenging me to develop new ideas every day.

Some friends I already had before this endeavor were crucial to my success. They accompanied me since the beginning of this project, sharing their thoughts and their doubts, helping me make this work even greater than I could do myself. To Mateus and Deni, my deepest thanks, this work would not be the same without your inputs.

The friends I have made during my graduate studies also deserve mention. Especially during these times of pandemic, it has been more important than ever to be in touch with such wonderful people. In particular, I thank Ana Fernanda Noli, Benedito Faustinoni, Flávia Castro, Letícia Reis, Mateus Piovezan, Rafael Rocha and Rodrigo "Mili" Barroso.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise."

(John W. Tukey)

RESUMO

LASSANCE, R. F. L. **Testagem não-paramétrica de hipóteses pragmáticas**. 2022. 71 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

Na área de testagem estatística, uma hipótese pragmática amplia uma hipótese precisa, tomando casos na vizinhança da nula como sendo tão merecedores de consideração quanto ela. Ao contrário dos métodos tradicionais, hipóteses pragmáticas permitem ao usuário avaliar suposições mais relevantes e, simultaneamente, fornecem estratégias para lidar com *Big Data* de forma responsável, evitando complicadores usuais. Contudo, até o presente momento, tais procedimentos só foram aplicados em casos que já supõem uma família paramétrica para os dados. Nesta dissertação de mestrado, nós exploramos hipóteses pragmáticas em um contexto não-paramétrico, o que reduz drasticamente o número de suposições e fornece cenários mais realistas. Ao expandir a teoria em Coscrato *et al.* (2019) para um contexto não-paramétrico, delimitamos os diferentes tipos de hipóteses precisas de interesse, assim como os respectivos desafios que cada uma delas apresenta. Daí, derivamos dois tipos de testes para hipóteses não-paramétricas: um que adere aos procedimentos usuais e um que é agnóstico (que aceita, rejeita ou mantém a indecisão a respeito de uma hipótese específica), sendo que ambos seguem a propriedade de monotonicidade. Ao final, utilizamos o processo da árvore de Pólya para construir testes em múltiplas aplicações, demonstrando como o tamanho da amostra, níveis de confiança/credibilidade e o limiar de uma hipótese pragmática impactam na decisão do teste.

Palavras-chave: hipóteses pragmáticas, testes agnósticos, função de dissimilaridade, bayesiana não-paramétrica.

ABSTRACT

LASSANCE, R. F. L. **Nonparametric pragmatic hypothesis testing**. 2022. 71 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

In statistical testing, a pragmatic hypothesis is an extension of a precise one, taking cases on the vicinity of the null as being equally worthy of appraisal. Unlike standard procedures, pragmatic hypotheses allow the user to evaluate more relevant assumptions and, at the same time, provide strategies to tackle Big Data responsibly, avoiding common drawbacks. However, up until now, these procedures have been applied only when a parametric family is assumed for the data. In this master's thesis, we explore pragmatic hypotheses in a nonparametric setting, which drastically reduces the number of presuppositions and provides more realistic scenarios. By expanding the theory in Coscrato *et al.* (2019) to a nonparametric context, we delimit the different types of precise hypotheses of interest and the respective challenges each of them presents. Then, we derive two kinds of tests for nonparametric pragmatic hypotheses: one that adheres to standard procedures and one that is agnostic (which accepts, rejects or remains undecided on a given hypothesis), both obeying the property of monotonicity. Lastly, we use the Pólya tree process for building tests in a multitude of applications, showing how sample size, confidence/credible levels and the threshold of a pragmatic hypothesis impact the decision of the test.

Keywords: pragmatic hypotheses, agnostic tests, dissimilarity function, Bayesian nonparametrics.

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

H_0 — Null hypothesis

$\mathcal{D}$ — Set of possible outcomes of an agnostic test

$\mathcal{X}$ — Sample space

$\phi(\cdot)$ — Agnostic test function

$R(\cdot)$ — Region estimator function

$\mathcal{P}(\cdot)$ — Power set

Θ — Parametric space

$Pg(H_0)$ — Pragmatic hypothesis

$\theta, \theta_0, \theta^*$ — Element of the parametric space

$d_Z(\cdot,\cdot), d(\cdot,\cdot), d_1(\cdot,\cdot), d^*(\cdot,\cdot)$ — Dissimilarity function

$f(\cdot), g(\cdot)$ — Density functions

Z — Theoretical future observation

ε — Threshold

$\mathbb{H}$ — Hypothesis space

H_a — Alternative hypothesis

h, h^* — Element of $\mathbb{H}$

$\mathbb{F}$ — Space of probability distributions

$\mathcal{H}(\boldsymbol{x})$ — Sampler that draws objects of $\mathbb{H}$ based on a sample $\boldsymbol{x}$

G — Probability distribution, usually representing cases under H_0

F — Probability distribution, in some cases represents the Pólya Tree process

$h(\cdot)$ — Functional of a random variable

$\mathbb{F} \times \mathbb{F}$ — Cartesian product of the probability distribution space

α — Significance level

Ω — Sample space

$B_i^{(m)}, B_{\epsilon_m}$ — Partition of the sample space

E — Set $\{0,1\}$

E^m — Set $\{0,1\}^m$

E^* — Set $\bigcup_{m=0}^{\infty} E^m$

Π — Collection of separable binary trees

$\mathcal{A}$ — Collection of hyperparameters of the Pólya Tree

$\alpha_{\boldsymbol{\epsilon}}$ — Hyperparameter of the Pólya Tree

$\boldsymbol{\epsilon}_m$ — Trajectory of the Pólya Tree

$\mathbb{E}$ — Expectation

CONTENTS

CHAPTER

1

INTRODUCTION

Although Science has brought countless contributions to humanity, it is important to highlight that its foundations are a product of its time and are subject to change, which tends to occur in moments of crisis. The fact that physicists only started to question the deterministic view of the world by the end of the 19th century is a clear demonstration of this phenomenon. This change in comprehension ensued because, although their measurements became more precise as time went by, physicists still could not eliminate errors ever present in the predictions of physical models (SALSBURG, 2002). Similarly to the structure of scientific revolutions (KUHN, 1962), it was only as these contradictions became more predominant and detrimental that a new paradigm emerged to replace the one before.

This new paradigm, statistical modeling, persists to this day and still sets the standards of what is considered as valid scientific research. While statistical models have become the predominant strategy of analysis in most areas of Science, statistical hypothesis testing in particular is the go-to procedure to identify when an assertion is backed up by the data or not. The reliance on Statistics has offered a robust procedure of analysis, since it is now possible to derive conclusions even when the data is subject to random effects or perturbations.

However, in recent years, there is a new set of challenges that threaten the credibility statistical tests have garnered in the scientific community. In particular, we highlight two problems whose consequences seem to be the most deleterious: meaningless tests that will always reject the null hypothesis and failure of replicating previous significant statistical findings. The first case is mostly a consequence of dealing with large datasets, since tests become so precise that they reject any hypothesis negligibly different from its theoretical result. The second case, although possibly present since the beginning, has become increasingly evident nowadays thanks to the coordinated effort of organizations (Open Science Collaboration, 2015).

Similarly to past iterations, these challenges point towards the need of proposing a new paradigm for scientific research. Not only should scientists change their views on hypothesis

testing, but the statistical methods themselves should be subject to adaptations as well (MAYO, 2018). When looking at the issues that standard tests present nowadays, Coscrato *et al.* (2019) elicits three that are particularly troublesome: **(i)** difficulty in interpreting the outcomes of tests, **(ii)** multiple hypothesis testing leading to logically incoherent conclusions and **(iii)** rejection of a precise hypothesis not being relevant from a practical perspective. While **(i)** relates to the attitudes of scientists towards testing, **(ii)** and **(iii)** are directly linked to the problems of replicability and Big Data aforementioned. The complications that arise from **(iii)** are particularly problematic, since precise hypotheses usually are the ones that interest scientists the most.

In order to face all of the three issues elicited, Coscrato *et al.* (2019) argue for the use of agnostic tests and pragmatic hypotheses. The latter, which solves issue **(iii)** and the difficulties presented by Big Data, can also be directly linked to the evolution of Science (ESTEVES *et al.*, 2019). These recent contributions offer a fresh view of statistical testing, providing a robust foundation that can be used by frequentists and Bayesians alike.

Given its novelty, there are still open questions when dealing with pragmatic hypotheses. Since they have only been applied in a context where the parametric family of the data is assumed, it is still necessary to delimit how to expand the theory to a nonparametric scenario. That is the main objective of this thesis. Also, given that they always reference a threshold of reasonable deviations from the original hypothesis - an information that sometimes is not fully available to researchers - it is pertinent to provide some possible rules of thumb to get to them as well.

The procedures used throughout this paper are based on the Pólya tree process (FERGUSON, 1974; LAVINE, 1992; LAVINE, 1994). This Bayesian nonparametric prior allows the user to draw distribution functions, which are used here to evaluate if the distribution of the data is sufficiently close to the null hypothesis. The proximity between the null hypothesis and the distributions sampled from the Pólya tree is evaluated through a dissimilarity function, which is the necessary tool to properly define the pragmatic space. We note, however, that using prior processes is not the only way to obtain conclusions from these tests. As long as one is able to provide a sampler of distribution functions based on data, the procedure remains feasible, even if it abides to a frequentist paradigm.

The structure of this proposal is as follows. Chapter 2 provides the background that has served as the foundation for our contribution until the present moment. Then, Chapter 3 expands the current literature on pragmatic hypothesis and describes all the required concepts (types of restriction on the null hypothesis, a sampler defined on the hypothesis space, new testing procedures), while also providing novel theorems and examples of cases with clear-cut solutions. Chapter 4 shows how to apply the developed tests in different applications, dealing with practical issues such as determining a threshold for the pragmatic hypothesis. Lastly, Chapter 5 elicits the current challenges and possible next steps for the research. There are also two sections on the appendix, Appendix A and Appendix B, which respectively details the Pólya tree process and presents the proofs of all the theorems developed.

CHAPTER

2

BACKGROUND

2.1 Agnostic Tests

When talking about statistical tests, there is a slight disconnect between its possible outcomes and what one actually wants to know about. After all, while a traditional statistical test either rejects a hypothesis or not, its use had the intention of finding out if the data corroborates the acceptance/refusal of the hypothesis or not. Thus, a traditional test amalgamates two credal states ("accept H_0" and "remain in doubt about H_0") into one single decision (not to reject H_0).

While this configuration could adhere to a more falsificationist view of Science (POPPER, 1934), its sensitivities have become more exposed through time, as mentioned in Chapter 1. In order to solve two of the major issues that a test presents (representing adequately the credal state and ensuring that multiple testing will not lead to a logically incoherent conclusion), Coscrato *et al.* (2019) presents the concept of an agnostic test.

Definition 1. *(COSCRATO et al., 2019) Take $\mathcal{D} = \{0, \frac{1}{2}, 1\}$ as the set of possible outcomes of a test, where 0 leads to acceptance of H_0, 1 to rejection and $\frac{1}{2}$ to neither, i.e., remaining undecided. Thus, if $\mathcal{X}$ denotes the sample space, an agnostic test is a function $\phi : \mathcal{X} \to \mathcal{D}$.*

From the formulation of Definition 1, although the credal states are respected, a new type of error occurs, such as Table 1 demonstrates. Every time the test reaches an undecided state, it commits a type III error. However, unlike the other error types, the type III error is always known. Still, such tests allow for control of errors type I and II simultaneously (COSCRATO; IZBICKI; STERN, 2020), a feature that is essential, but rarely evaluated with standard tests.

It automatically follows from Definition 1 that any standard test can be reframed as an agnostic test, as long as $Im[\phi] = \{0, 1\}$. Thus, agnostic tests can be considered as a generalization of standard tests, while having the benefit of equating the possible outcomes with the credal states of interest. This is sufficient to solve issue **(i)** presented in Chapter 1. As for ensuring logical

STATEMENT	TEST OUTCOME		
	Accept H_0	Remain agnostic	Reject H_0
H_0 is true	No error	Type III error	Type I error
H_0 is false	Type II error	Type III error	No error

Table 1 – Error types based on the outcome of an agnostic test

consistency between the outcomes, Coscrato *et al.* (2019) and Esteves *et al.* (2019) put forward the notion of a region estimator.

Definition 2. *(ESTEVES et al., 2019) Let $\mathcal{X}$ denote the sample space used to test a hypothesis. A region estimator is a function $R : \mathcal{X} \to \mathcal{P}(\Theta)$, where $\mathcal{P}(\Theta)$ is the power set of Θ, the parametric space.*

Based on Definition 2, it is possible to derive an agnostic test. After all, for a given data $\boldsymbol{x} \in \mathcal{X}$, one could set

$$\phi(\boldsymbol{x}) = \begin{cases} 0, & \text{if } R(\boldsymbol{x}) \subseteq H_0; \\ 1, & \text{if } R(\boldsymbol{x}) \subseteq H_0^c; \\ \frac{1}{2}, & \text{otherwise.} \end{cases} \tag{2.1}$$

as the decision boundaries. As Figure 1 demonstrates, this construction is reasonable, since $\phi(\boldsymbol{x})$ only makes an assertive decision if the whole region is contained either in H_0 (leading to its acceptance) or in H_0^c (leading to the rejection of the null).

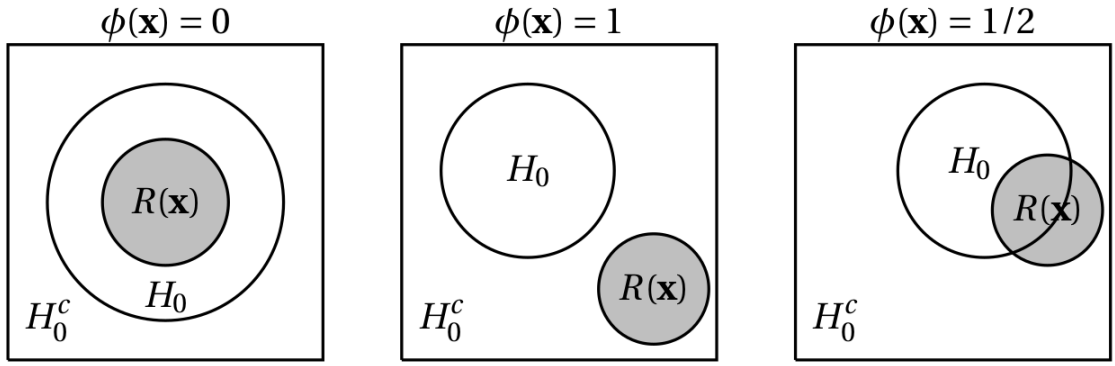


Figure 1 – Test outcome for each region estimate (source: Coscrato *et al.* (2019))

However advantageous this setup of agnostic tests is, it possesses a major drawback, which is not being able to evaluate precise hypotheses in the same manner. Precise hypotheses represent equality assumptions about the behavior of the data, such as $H_0 : \theta = \theta_0$, for example. Since H_0 in this case has a measure of 0, any region estimator will either reject or remain agnostic about H_0, leading us back to a standard test.

In order to sidestep this problem and, at the same time, provide a solution to issue **(iii)** presented in Chapter 1, the notion of a pragmatic hypothesis is developed.

2.2 Pragmatic Hypotheses

In a parametric context, i.e., when a parametric family is assumed for the data, a precise hypothesis is an equality assumption about one or more parameters of the statistical model. Then, a pragmatic hypothesis is an enlargement of the precise one, evaluating its vicinity as an equally viable assumption to be hold. If H_0 is the precise hypothesis, we define $Pg(H_0)$ as the pragmatic hypothesis derived from it. In order to establish what is the vicinity of H_0 that will form $Pg(H_0)$, we first need to set a dissimilarity function and a threshold of acceptable values for the dissimilarity.

Take $H_0 : \theta = \theta_0$ as the precise hypothesis of interest, where Θ is the parametric space of θ. As the name implies, the dissimilarity function evaluates how differently the data should behave when $\theta = \theta^*, \theta^* \in \Theta$, compared to the null hypothesis. Unlike a distance function, it does not need to have 0 as its smallest value, although it has to be strictly non-negative. One of the most commonly used dissimilarity functions is the classification dissimilarity, given by

$$d_Z(\theta_0, \theta^*) = 0.5 \left[\mathbb{P}_{\theta_0} \left(\frac{f(Z|\theta_0)}{f(Z|\theta^*)} > 1 \middle| \theta = \theta_0 \right) + \mathbb{P}_{\theta^*} \left(\frac{f(Z|\theta^*)}{f(Z|\theta_0)} > 1 \middle| \theta = \theta^* \right) \right], \tag{2.2}$$

where Z is a possible future observation of the data and $f(Z|\theta)$ is its probability density function. Since both terms in (2.2) are probabilities between 0.5 and 1 - after all, it is not possible for any probability distribution to be closer to the data than its true distribution - the classification dissimilarity only assumes values in $[0.5, 1]$. Citing Coscrato *et al.* (2019), "the classification dissimilarity is the highest achievable probability of correctly identifying which θ generated Z." Thus, the closest (2.2) is to 0.5, the more reasonable it is to assume that θ^* should belong to $Pg(H_0)$.

Definition 3. *Esteves et al. (2019) Let $H_0 : \theta = \theta_0$, d_Z be a predictive dissimilarity function and $\varepsilon > 0$. The pragmatic hypothesis for H_0, $Pg(H_0)$, is*

$$Pg(H_0) := \{\theta^* \in \Theta : d_Z(\theta_0, \theta^*) < \varepsilon\}.$$

As θ^* gets further apart from θ_0, the dissimilarity function gets higher or closer to its maximum value, it is possible to fully identify the set $\theta^* \in \Theta$ that belongs to $Pg(H_0)$. And since the choice of ε is entirely arbitrary, the pragmatic space can be as restrictive as one desires.

In case the sharp hypothesis is less restrictive than a point-wise assumption (such as $H_0 : \theta \in \Theta_0$, where $\dim(\Theta_0) < \dim(\Theta)$), Definition 3 can straightforwardly be adapted to

$$Pg(H_0) := \bigcup_{\theta_0 \in \Theta_0} \{\theta^* \in \Theta : d_Z(\theta_0, \theta^*) < \varepsilon\}. \tag{2.3}$$

With the addition of the pragmatic hypothesis, the theory of agnostic tests presented in section 2.1 can proceed exactly as described, since $Pg(H_0)$ occupies a continuous region on the hypothesis space. Not only that, but the use of pragmatic hypotheses allow for the user to explicitly define what is considered as a meaningful deviation from H_0 through the choice of ε. Thus, it solves issue **(iii)** presented in Chapter 1 and provides a strategy that works as expected even when dealing with a massive number of observations.

In closing, it is important to highlight that the choice of a predictive dissimilarity function d_Z and a threshold ε should be a reflection of the researcher's understanding of what is an irrelevant deviation from H_0. Still, given the novelty of the research and its additional mathematical complexity, a choice of dissimilarity that leads to a mathematically convenient solution can be valid in some cases.

CHAPTER

3

METHOD

Since the previous contributions have been explored in Chapter 2, it is time to elicit the novel results that this thesis has to offer. So far, both agnostic tests and pragmatic hypotheses have only been applied in contexts where a parametric family is assumed for the data. Particularly for pragmatic hypotheses, Equation 2.3 becomes meaningless in a nonparametric scenario, since $Pg(H_0)$ does not evaluate parameter values anymore, but more general concepts such as cumulative distribution functions. In this section, we expand the notion of pragmatic hypotheses to a nonparametric setting, providing a general framework to derive and apply new tests. The proofs of all pertinent mathematical results are presented in Appendix B.

3.1 Nonparametric Pragmatic Hypotheses

We begin the section by providing an updated definition of the pragmatic hypothesis, in order to highlight similarities and differences with the original concept.

Definition 4. *(Nonparametric Pragmatic Hypothesis) Set $\mathbb{H} = H_0 \cup H_a$ as the hypothesis space, where H_0 is a precise hypothesis and $H_a = H_0^c$. Take $h, h^* \in \mathbb{H}$ as elements of such space. For a given dissimilarity function $d(\cdot,\cdot)$ and a threshold $\varepsilon > 0$, a nonparametric pragmatic hypothesis is defined as*

$$Pg(H_0) = \bigcup_{h \in H_0} B(h, \epsilon) = \bigcup_{h \in H_0} \{h^* \in \mathbb{H} : d(h, h^*) < \varepsilon\} = \left\{ h^* \in \mathbb{H} : \inf_{h \in H_0} d(h, h^*) < \varepsilon \right\},$$

where $B(h, \epsilon)$ is an open ball (KREYSZIG, 1978).

From now on, we will use the initials NPHT to refer to a nonparametric pragmatic hypothesis test.

From Definition 4, it is clear that many elements of a parametric pragmatic hypothesis are being recycled with few significant changes. Similarly to Equation 2.3, there is still an interest in

the elements of the space that are close enough to H_0, based on a dissimilarity function. However, in the nonparametric case, $\mathbb{H}$ is more abstract than Θ, even in settings where the null hypothesis can be reframed as a region constrained to a specific parametric space. Also, the notion of a dissimilarity function is still essential to determine negligible deviations from the null, even if thinking about a future observation Z is not always relevant in the nonparametric case (which is why the subscript of d_Z was dropped).

Unlike its parametric counterpart, it is not possible to fully specify all the elements of $\mathbb{H}$ that belong to $Pg(H_0)$ in the NPHT. Still, in any practical setting, it suffices to check if a specific $h^* \in \mathbb{H}$ (which should be corroborated by the data) is such that $\inf_{h \in H_0} d(h, h^*) < \varepsilon$. For now, we assume that an element h^* that adequately represents the data is known. In section 3.2, we suggest a procedure to draw these reasonable candidates.

To ensure that the research could be completed in accordance to the time constraints of a master's thesis, the scope of the problem was reduced to cases where the data is independent and identically distributed, univariate and presented no covariates. In general, but not always, we will restrict ourselves to cases where the hypothesis space is the space of distribution functions, represented by $\mathbb{F}$.

3.2 Building the NPHT

3.2.1 Procedure overview

Now that $Pg(H_0)$ has been defined in the context of the NPHT, it is time to propose accessible guidelines that could allow for a researcher to devise a test. In this section, we present the building blocks of a test that, while not necessarily being logically coherent, guarantees that at least the property of monotonicity is achieved and is not affected by issues **(i)** and **(iii)** presented at Chapter 1.

The first step is to identify the type of restriction that is being described in H_0. This is important for multiple reasons: it allows one to identify what the hypothesis space is, to establish what are the more intuitive choices for the dissimilarity function and also to find $\inf_{h \in H_0} d(h, h^*)$ for a specific $h^* \in \mathbb{H}$. Except for ε, this is enough to determine all the relevant parts of $Pg(H_0)$.

Once the hypothesis space has been determined, the next step is to assign a model that is capable of adequately representing the data in relationship to $\mathbb{H}$. Given the scope restrictions proposed in section 3.1, it is enough to provide a faithful representation in $\mathbb{F}$, a feasible objective as long as one uses a method that draws cumulative distribution functions based on data. To achieve this, one valid approach is to resort to prior processes (PHADIA, 2016), although there are no prohibitions on using frequentist methods as well. In general, we use $\mathcal{H}(\boldsymbol{x})$ to represent a model that draws objects in $\mathbb{H}$ based on a sample $\boldsymbol{x}$.

Lastly, it is necessary to propose a test that can reach a decision based on how reasonable

it is to assume that the data was generated by an element of $Pg(H_0)$. Of course, this can be done in multiple ways and should reflect the main interests of the researcher. One possibility is to adhere to a setting that more closely resembles current standard tests. For a given significance level α and a threshold ε, one would reject the null hypothesis if

$$\mathbb{P}\left(\inf_{h\in H_0} d(h,\mathcal{H}(\boldsymbol{x})) < \varepsilon\right) < \alpha, \tag{3.1}$$

i.e., if the probability that $\mathcal{H}(\boldsymbol{x})$ belongs to $Pg(H_0)$ is less than α. This test alone is enough to solve issue **(iii)** presented in Chapter 1, since in this case the researcher directly informs which cases constitute a considerable departure from H_0 and which offer negligible differences based on one's practical setting.

Considering all issues in standard tests brought up in Chapter 1, we also put forth the idea of using an agnostic test that is based on a region estimator, since they have the potential to solve issues **(i)**-**(iii)** altogether. So far, we were able to propose a test that fully addresses **(i)** and **(iii)**, while partially addressing **(ii)** (it obeys monotonicity, i.e., it reaches coherent conclusions for a pair of hypotheses when one is a subset of the other). To achieve this, we choose two quantiles of $\inf_{h\in H_0} d(h,\mathcal{H}(\boldsymbol{x}))$, with respective probabilities α_1 and α_2 $(\alpha_1 \leq \alpha_2)$. Then, if $q_p(\cdot)$ represents the p-th quantile,

$$\phi(\boldsymbol{x}) = \begin{cases} 0 \text{ (accept } H_0), & \text{if } \varepsilon \geq q_{\alpha_2}\left[\inf_{h\in H_0} d(h,\mathcal{H}(\boldsymbol{x}))\right]; \\ 1 \text{ (reject } H_0), & \text{if } \varepsilon < q_{\alpha_1}\left[\inf_{h\in H_0} d(h,\mathcal{H}(\boldsymbol{x}))\right]; \\ \frac{1}{2} \text{ (remain agnostic)}, & \text{otherwise.} \end{cases} \tag{3.2}$$

By choosing $\alpha_2 = \alpha_1 = \alpha$, Equation 3.2 provides the exact same conclusion as Equation 3.1.

3.2.2 Part 1: Restriction types on H_0

We now turn our attention to specific restrictions over H_0, so as to evaluate how easily one can derive $Pg(H_0)$. It is important to note, however, that this list may not be exhaustive, but they were sufficient to describe all kinds of nonparametric hypotheses that the authors thought about during the research. For each restriction type, an example is provided to help develop an intuition in every case.

3.2.2.1 Parametrical

This is the case that most closely resembles the context of Definition 3, so we develop the first intuition of the theory from it. We start by noticing that the null hypothesis $H_0 : \theta = \theta_0$ could be reframed as an assumption about the distribution of the data itself. If we take G_X to be the distribution function assumed for the data, then the following equivalence applies:

$$H_0 : \theta = \theta_0 \Leftrightarrow H_0 : X \sim G_{X|\theta_0}. \tag{3.3}$$

Although, conceptually speaking, the null hypothesis has not been changed, the right side of (3.3) implies that the alternative hypothesis H_a has become wider than in the original setting. Now, H_a encompasses any case where the distribution of the data is different than $G_{X|\theta_0}$, which implies that $Pg(H_0)$ is a set that contains $G_{X|\theta_0}$ and a subset of distribution functions on $\mathbb{H}$. Thus, if $\mathbb{F}$ is the space that represents all possible distribution functions for the data X, Z is a future observation, d_Z is a dissimilarity function between probability distributions and $\varepsilon > 0$ is a given threshold, the pragmatic space is described by

$$Pg(H_0) = \left\{F \in \mathbb{F} : d_Z(G_{Z|\theta_0}, F_Z) < \varepsilon\right\}. \tag{3.4}$$

If we return to Definition 4, (3.4) is valid because $\mathbb{H} = \mathbb{F}$, and thus

$$Pg(H_0) = \left\{F \in \mathbb{F} : \inf_{G \in H_0} d(G, F) < \varepsilon\right\} = \left\{F \in \mathbb{F} : d_Z(G_{Z|\theta_0}, F) < \varepsilon\right\},$$

since $G_{X|\theta_0}$ is the only distribution function that belongs to H_0.

We now present a parametrical restriction on H_0 in its full generality. If we were to expand Equation 3.4 and evaluate a hypothesis that more closely resembles Equation 2.3, it would then become

$$Pg(H_0) = \left\{F \in \mathbb{F} : \inf_{G \in H_0} d(G, F) < \varepsilon\right\} = \left\{F \in \mathbb{F} : \inf_{\theta \in \Theta_0} d_Z(G_{Z|\theta}, F) < \varepsilon\right\}. \tag{3.5}$$

It could be assumed that $\Theta_0 = \Theta$, the whole parametric space, since $H_0 : X \sim G_{X|\theta}$, $\theta \in \Theta$, is also a sharp hypothesis in this scenario. Hence, when the null hypothesis assumes that the data obeys a specific distribution function or a parametric family, then it is said that there is a parametrical restriction on H_0.

Identifying if a candidate $F \in \mathbb{F}$ belongs to $Pg(H_0)$ is rather straightforward in this case, since Equation 3.5 could be translated into an optimization procedure. For every given F, we find $\hat{\theta} \in \Theta_0$ such that $\inf_{\theta \in \Theta_0} d_Z(G_{Z|\theta}, F) = d_Z(G_{Z|\hat{\theta}}, F)$. Then, if $d_Z(G_{Z|\hat{\theta}}, F) < \varepsilon$, we could then conclude that $F \in Pg(H_0)$. Since this procedure is rather general, any combination of parametric families and dissimilarity functions could be evaluated this way.

Given how complex and plural the space $\mathbb{F}$ is, there is no direct method of explicitly declaring the subfamily of $\mathbb{F}$ that forms $Pg(H_0)$, i.e., it is not possible to elicit all functions that belong to $Pg(H_0)$. This implies that we must know of a restricted (discrete) group of functions beforehand, one that would suffice to reach a valid conclusion. In practice, our interest resides in checking if the distribution function of a given sample is sufficiently close to the parametric family assumed at H_0. Thus, in this particular case, the NPHT can be reframed as a problem of density estimation.

In many cases, a dissimilarity function that works on the context presented in section 2.2 can be easily adapted to the fully restricted case. Take, for example, the classification dissimilarity in (2.2). Then, its nonparametric counterpart is

$$d_Z(G, F) = 0.5\left[\mathbb{P}_G\left(\frac{g(Z)}{f(Z)} > 1 \middle| Z \sim G\right) + \mathbb{P}_F\left(\frac{f(Z)}{g(Z)} > 1 \middle| Z \sim F\right)\right], \tag{3.6}$$

where $f(.)$ and $g(.)$ are, respectively, the probability density functions of F and G.

Example 1 ($H_0 : N(t), t \in \mathbb{R}^+$, is a Poisson process)**.** *The Poisson process (ROSS, 2009) is perhaps the most widely known stochastic process in the literature. It is a counting process that assumes that $N(t) \sim Poisson(\lambda t), \forall t \in \mathbb{R}^+$. In this example, the interest here is twofold: to test if the data behaves like a Poisson process and to find what are reasonable values for λ.*

If $(X_1, \cdots, X_n)$ is a sample of the moment in time each observation has occurred, it follows from the properties of the process that, if $T_0 = 0$ and $T_i = X_i - X_{i-1}$, $i \in \{1, \cdots, n\}$, then $T_i \overset{iid}{\sim} Exp(1/\lambda)$. Thus, the null hypothesis can be reframed as $H_0 : T \sim Exp(1/\lambda)$, $\lambda \in \mathbb{R}^+$, which is itself a parametrical restriction. Hence,

$$Pg(H_0) = \left\{ F \in \mathbb{F} : \inf_{\lambda \in \mathbb{R}^+} d_Z(G_{Z|\lambda}, F_Z) < \varepsilon \right\},$$

where $G_{Z|\lambda} \equiv Exp(1/\lambda)$ in this case.

We now turn our attention to the choice of the dissimilarity function. If we were to use the dissimilarity function that was just presented in Equation 3.6, it would be expressed as

$$d_Z(G_{Z|\lambda}, F_Z) = 0.5 \left[\mathbb{P}_G \left(\frac{\exp(-Z/\lambda)}{\lambda f(Z)} > 1 \middle| Z \sim G \right) + \mathbb{P}_F \left(\frac{\lambda f(Z)}{\exp(-Z/\lambda)} > 1 \middle| Z \sim F \right) \right].$$

Consequently, if $F \in \mathbb{F}$,

$$F \in Pg(H_0) \Leftrightarrow \exists \lambda \in \mathbb{R}^+ : \mathbb{P}_G \left(\frac{\exp(-Z/\lambda)}{\lambda f(Z)} > 1 \middle| Z \sim G \right) + \mathbb{P}_F \left(\frac{\lambda f(Z)}{\exp(-Z/\lambda)} > 1 \middle| Z \sim F \right) < 2\varepsilon. \tag{3.7}$$

If we were to accept the hypothesis, i.e., if the data consistently provided candidates for F such that we could find a $\lambda \in \mathbb{R}^+$ that obeyed Equation 3.7, a possible next step would be to obtain a point-estimate for λ, or at least a set of reasonable values for it. A natural candidate for this would be the maximum likelihood estimator, which in this case is given by $\hat{\lambda} = \sum_{i=1}^{n} \frac{t_i}{n}$, where t_i is the observed value of T_i, $i \in \{1, \cdots, n\}$. Then, we could use the same procedure to test how trustworthy $\hat{\lambda}$ is in this case by simplifying Equation 3.7 to

$$F \in Pg(H_0) \Leftrightarrow \mathbb{P}_G \left(\frac{\exp(-Z/\hat{\lambda})}{\hat{\lambda} f(Z)} > 1 \middle| Z \sim G \right) + \mathbb{P}_F \left(\frac{\hat{\lambda} f(Z)}{\exp(-Z/\hat{\lambda})} > 1 \middle| Z \sim F \right) < 2\varepsilon.$$

3.2.2.2 Functional

From this point on, no specific parametric family is assumed for H_0. In this case, we consider cases like $H_0 : h(X) = \theta_0$, i.e., where we restrict H_0 to probability distributions that obey a specific functional restriction. Unlike subsubsection 3.2.2.1, closed-form solutions may vary depending on the null hypothesis and the dissimilarity function of choice.

If one attempts to follow a similar strategy as in subsubsection 3.2.2.1, this time it is necessary to find a function G_X that provides the smallest dissimilarity while at the same time

being such that $h(X) = \theta_0$. If d is a metric over the space, then using functional analysis to find a projection seems to a viable solution, although complex.

Another possible solution resides in considerably restricting the dissimilarity function defined on $\mathbb{F}$, such that it could be reshaped into a dissimilarity that only takes the functional into account. In this particular case,

$$\inf_{G \in H_0} d(G,F) \equiv d^*(\theta_0, h(Z|F)).$$

This is of course a poor decision, since it reduces the problem to one that may be exceedingly simple, but it serves as a preliminary solution if one is unsure of how to derive a test.

Example 2 ($H_0 : G(x_0) = p_0$)**.** *This example illustrates the case of a quantile test, without assuming any particular distribution for the data. For this case, we will choose the L1-distance as the dissimilarity function. Thus, for any two probability distributions $F, G \in \mathbb{F}$,*

$$d_1(G,F) := \|G - F\|_1 := \int_{\mathbb{R}} |G(x) - F(x)| dx.$$

Here, we take G as a distribution function such that H_0 is true and F as a potential candidate for a distribution function that belongs to $Pg(H_0)$.

The following theorem provides a straightforward procedure for obtaining the infimum of the dissimilarity function presented in Example 2.

Theorem 1 (Quantile test)**.** Take $H_0 : G(x_0) = p_0$ as the null hypothesis, F as a distribution function and

$$d(G,F) = \int_{-\infty}^{\infty} |G(x) - F(x)| dx$$

as the dissimilarity function. Then, if $a = \min(F^{-1}(p_0), x_0)$ and $b = \max(F^{-1}(p_0), x_0)$,

$$\delta = \inf_{G \in H_0} d(G,F) = \inf_{G \in H_0} \int_{-\infty}^{\infty} |G(x) - F(x)| dx = \int_a^b |p_0 - F(x)| dx. \tag{3.8}$$

Example 2 (continued)**.** *In case $F(x_0) = p_0$, then F also belongs to H_0, which implies that $\inf_{G \in H_0} d_1(G,F) = 0$. Now, if $F(x_0) < p_0$, (3.8) becomes*

$$\inf_{G \in H_0} d_1(G,F) = \int_{x_0}^b (p_0 - F(x)) dx = (b - x_0)p_0 - \int_{x_0}^b F(x) dx.$$

Equivalently, if $F(x_0) > p_0$,

$$\inf_{G \in H_0} d_1(G,F) = \int_a^{x_0} (F(x) - p_0) dx = \int_a^{x_0} F(x) dx - (x_0 - a)p_0.$$

Then, if $\int_a^b F(x)dx$ cannot be explicitly obtained, a Monte Carlo integration procedure (ROBERT; CASELLA, 2005) could be applied.

3.2.2.3 Contextual

This restriction type is by far the most conceptually challenging, since it no longer equates $\mathbb{H}$ to $\mathbb{F}$ and does not assume that the data is identically distributed (even if it could be split into parts that are i.i.d.). More specifically, we assume that multiple datasets are provided, each of them pertaining to a specific random variable, and that the hypothesis of interest is described through a contextual relationship between these variables. Some hypotheses that belong to this category are:

- Equality of the distribution functions of two populations;
- Independence between two random variables;
- Equality of variance of k samples (a nonparametric adaptation of Levene's test).

In general, if the null hypothesis assumes a relationship between $k > 1$ random variables, then $\mathbb{H} = \bigotimes_{i=1}^{k} \mathbb{F}_i$, where $\mathbb{F}_i$ represents the set of valid distribution functions of the i-th variable.

It is important to highlight that, even when the null hypothesis exhibits a functional assumption, it is the contextual relationship between the variables that truly defines it. Take for example the nonparametric Levene's test: although it is assumed that the variance of each variable is the same, their distribution function can assume any behavior and their variance can take on any value, as long as it is the same for all. This is the reason why, unlike the other restriction types, it is harder to propose a unified approach that, for a specific element of $\mathbb{H}$, is able to identify if it belongs to $Pg(H_0)$. Not only are the comparisons more complex (since this time we have to propose dissimilarity functions that work for higher dimensional objects), but it may also prove difficult to find the infimum due to how loosely defined the null hypothesis is.

Example 3 ($H_0 : G_X = G_Y$)**.** *Tests that compare the equivalence between the distribution functions of two populations are popular and widespread in the literature, with slight variations being proposed through time (such as Holmes et al. (2015), Inácio, Izbicki and Salasar (2020) and Ceregatti, Izbicki and Salasar (2021) to name a few). Here, we propose the NPHT version of it.*

We highlight that $\mathbb{H} = \mathbb{F} \times \mathbb{F}$, i.e., the hypothesis space in this case is composed by the Cartesian product of the distribution function space. In Figure 2, a visualization is provided to give an idea of the peculiarities of such space. Each axis of the figure represents the distribution function of a specific population. Then, the blue line represents the null hypothesis, where both distributions are equal. Thus, while the red dot is an element of $\mathbb{H}$ (i.e., a given pair of distribution functions), the red arrow represents the closest distance between such element and H_0.

The next theorem offers an intuitive solution for Example 3. Under rather general conditions, instead of requiring the examination of pairs of distribution functions defined on $\mathbb{F} \times \mathbb{F}$, it allows one to just compare the distance between the two distribution functions that are deemed as representative of the data - F_X and F_Y.

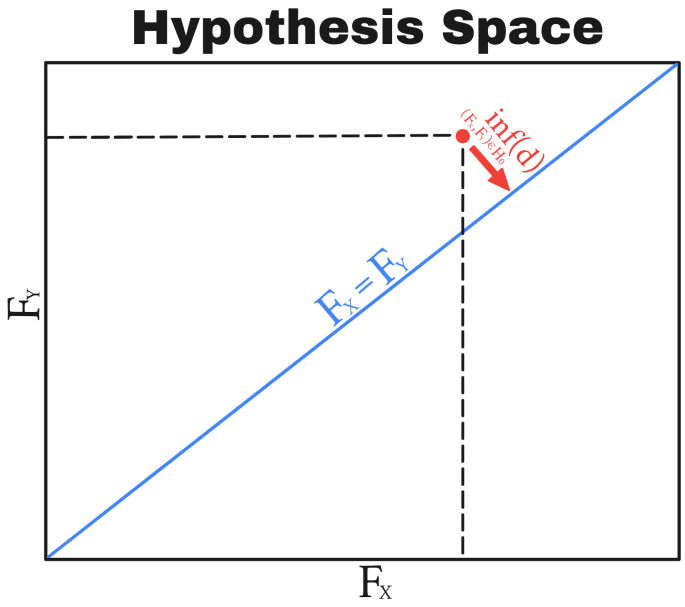


Figure 2 – Representation of the two population NPHT

Theorem 2 (Two sample test)**.** Take $H_0 : G_X = G_Y$ as the null hypothesis, F_X and F_Y as two distribution functions and

$$d[(G_X, G_Y), (F_X, F_Y)] = d^*(G_X, F_X) + d^*(G_Y, F_Y),$$

where $d^*(\cdot, \cdot)$ is a distance function. Then

$$\delta = \inf_{(G_X, G_Y) \in H_0} d[(G_X, G_Y), (F_X, F_Y)] = d^*(F_X, F_Y).$$

Example 3 (continued)**.** *Based on Theorem 2, it is possible to use any dissimilarity function to compare two pairs of distribution functions of interest, as long as the former is also a distance function. This requirement is easily achievable by subtracting the dissimilarity function by its minimum value. For example, if one were to use the classification dissimilarity of Equation 3.6, it would have to be corrected to*

$$d^*(F_X, F_Y) = 0.5 \left[\mathbb{P}_{F_X} \left(\frac{f_X(Z)}{f_Y(Z)} > 1 \middle| Z \sim F_X \right) + \mathbb{P}_{F_Y} \left(\frac{f_Y(Z)}{f_X(Z)} > 1 \middle| Z \sim F_Y \right) \right] - 0.5, \tag{3.9}$$

where f_X and f_Y are the respective density functions of F_X and F_Y.

3.2.3 Part 2: Establishing $\mathcal{H}(\cdot)$

In this section, the main focus resides in determining elements of $\mathbb{H}$ that are somewhat representative of the dataset. To achieve this, we propose a sampler $\mathcal{H}(\cdot)$, which is responsible to draw elements of $\mathbb{H}$ based on the data it receives. Hence, it is only at this point that any information contained in a collected sample is actually used in the test. Once a sample $\boldsymbol{x}$ is used to inform how $\mathcal{H}(\boldsymbol{x})$ should behave, no further references to the data are required for the test.

Although there are conditions on what constitutes $\mathcal{H}(\cdot)$, its structure is general enough to allow for methods both frequentist and Bayesian. In subsection 3.2.2, it was shown for the purposes of this thesis that $\mathbb{H}$ either represents the space of distribution functions or the Cartesian of multiple distribution spaces. Furthermore, given the independence assumption between observations, most nonparametric density estimation methods should be robust enough to be used in this context.

There is, however, an important property of $\mathcal{H}(\cdot)$ that should not be overlooked. Contrary to section 2.2, where the whole uncertainty resided on the parametric space, in the NPHT case it resides in the hypothesis space $\mathbb{H}$. We call $\mathcal{H}(\cdot)$ a sampler instead of a function per se due to the importance of making explicit the uncertainty around which element of $\mathbb{H}$ is the true source of the data. Thus, by treating $\mathcal{H}(\cdot)$ as random, it directly follows that the dissimilarity function will also be random, a feature that will be of utmost importance when deriving both standard and agnostic tests (subsection 3.2.4).

As long as it can provide valid candidates of elements of $\mathbb{H}$ that are representative of the data, there are no actual impediments on which statistical school of thought to abide by. In frequentist statistics, density estimation methods with bootstrapping procedures usually provide adequate candidates, while in Bayesian statistics there are prior processes which naturally keep the uncertainty about the data generating process (FERGUSON, 1973; FERGUSON, 1974).

Example 4 (Basic distribution function estimator)**.** *Given a sample* $\boldsymbol{x} = (x_1, \cdots, x_n)$*, the simplest distribution function estimator is perhaps*

$$\hat{F}_{Y|\boldsymbol{x}}(y) = \sum_{i=1}^{n} \frac{\mathbb{I}(x_i \leq y)}{n}, \tag{3.10}$$

where $\mathbb{I}(\cdot)$ *is the indicator function. Thus, if* $\boldsymbol{x}^B$ *is a random bootstrap sample of* $\boldsymbol{x}$ *and* $\mathbb{H} = \mathbb{F}$*, a valid proposal for the sampler would be* $\mathcal{H}(\boldsymbol{x}) = \hat{F}_{Y|\boldsymbol{x}^B}(\cdot)$*. This, however, could be a poor choice depending on the context. If* $\boldsymbol{x}$ *was related to a continuous variable, then* $\mathcal{H}(\boldsymbol{x})$ *would always draw discrete distribution functions, thus not adequately representing* $\mathbb{F}$*. And since no density function could be derived from* $\mathcal{H}(\boldsymbol{x})$*, it would not be possible to use a dissimilarity function such as the one presented in Equation 3.6.*

In general, there are at least three desirable properties one should take into account when considering proposals for $\mathcal{H}(\cdot)$:

- Adaptability: the proposal easily adapts to data and provides draws that actually belong to the space $\mathbb{H}$ of interest;
- Generality: the proposal can be applied no matter the type of random variable one is examining, such as discrete, absolutely continuous or singular continuous;
- Scalability: the proposal scales well even when dealing with larger sample sizes. In particular, it is desirable for the method to be done in parallel if possible.

3.2.4 Part 3: Testing procedures

This section details the testing procedures that have already been developed. All of the code was developed in R (R Core Team, 2020) and is available upon request.

The first method proposed is the application of a pragmatic hypothesis in a standard testing procedure. If $d(\cdot,\cdot)$ is the elected dissimilarity function, $\varepsilon > 0$ is a threshold for the dissimilarity and α is a significance level, we could reject the null hypothesis if

$$\mathbb{P}\left(\inf_{h\in H_0} d(h,\mathcal{H}(\boldsymbol{x})) < \varepsilon\right) < \alpha, \tag{3.11}$$

i.e., when the probability that the dissimilarity function does not exceed ε is less than a significance α. Thus, Equation 3.11 mirrors standard procedures of statistical testing, making it highly intuitive. Then, even if one does not wish to use an agnostic test, it is still possible to evaluate a pragmatic hypothesis.

There are, however, some conceptual differences between this test and standard procedures that should be highlighted. Unlike p-values, there is neither a need to condition the probability on H_0 nor a clear definition of what would be considered "more extreme" than the observed results. Instead, through the use of the random function $\mathcal{H}(\boldsymbol{x})$, the probability is built from reasonable candidates of the true distribution function based on the sample data $\boldsymbol{x}$. Also, since

$$\mathbb{P}\left(\inf_{h\in H_0} d(h,\mathcal{H}(\boldsymbol{x})) < \varepsilon\right) = \mathbb{P}\left(\mathcal{H}(\boldsymbol{x}) \in Pg(H_0)\right),$$

this probability intrinsically carries the definition of a rejection region. After all, $\mathcal{H}(\boldsymbol{x}) \notin Pg(H_0)$ automatically implies that the differences between the sampler and H_0 are not negligible and, since all information our data provides is expressed through $\mathcal{H}(\boldsymbol{x})$, it should lead to the rejection of the null. Thus, once one sets ε and chooses through α how improbable a setting has to be to lead to its rejection, there is no further need to control for other factors such as sample size. This effect is shown in Example 5.

Example 5 (Effect of sample size on decision)**.** *The purpose of this example is to demonstrate how a standard test and a NPHT are affected as the sample size increases. To achieve this and to keep the example as simple as possible, we sample* $(X_1,\cdots,X_n) \overset{iid}{\sim} N(\mu,1)$*, where* $n \in \{50, 500, \cdots, 5\times10^7\}$ *and* $\mu \in \{0.2, 0.02, 0.002\}$.

For the standard test, we assume that the parametric family of the data is known and use a Student's t-test to check $H_0 : \mu = 0$*, registering its p-value. As for the NPHT, we test* $H_0 : X \sim N(0,1)$*, using the classification dissimilarity proposed in Equation 3.6 and* $\varepsilon = 0.55$*. As for* $\mathcal{H}(\cdot)$*, we chose the Pólya tree process (LAVINE, 1992; LAVINE, 1994), using* $N(0,1)$ *as the centering distribution.*

Table 2 presents two terms in each cell, all represented as percentages: the first is the p-value of the t-test, while the second is the smallest α *required to reject the null hypothesis of*

the NPHT. Each cell represents a combination of the true value of μ and the sample size n. It is clear that, for a large enough sample size, the t-test starts to reject the null hypothesis, no matter how corroborated H_0 seemed to be for smaller values of n. In opposition, the NPHT may be less consistent for smaller sample sizes (which should be due to the prior choice and the uncertainty surrounding the true distribution function, unlike the t-test that already takes it for granted), but for larger samples it can adequately separate if the difference is too large ($\mu = 0.2$) or small enough ($\mu = 0.02$ and $\mu = 0.002$).

	True μ		
Sample size (n)	0.2	0.02	0.002
5×10^1	(31.7875, 21.9365)	(66.6656, 22.3331)	(22.9480, 19.6941)
5×10^2	(0.0245, 25.0548)	(66.3828, 13.6322)	(23.2528, 18.3166)
5×10^3	(0.00, 12.3468)	(7.7806, 0.1401)	(45.7200, 0.0837)
5×10^4	(0, 0.5186)	(0.0037, 0.2402)	(28.2979, 0.2080)
5×10^5	(0, 0)	(0.00, 16.0659)	(2.1192, 19.5977)
5×10^6	(0, 0)	(0, 53.7031)	(0.2567, 51.8039)
5×10^7	(0, 0)	(0, 53.9987)	(0, 55.5592)

Table 2 – Comparison between the p-value of a Student's t-test and the smallest α required to reject the hypothesis in the NPHT, all probabilities are written as percentages

Inspired by Coscrato *et al.* (2019) and Esteves *et al.* (2019), we also propose an adaptation of agnostic tests to a nonparametric pragmatic hypothesis. In a similar setting as Equation 3.11, we derive a confidence/credible interval for the dissimilarity function $\inf_{h\in H_0} d(h,\mathcal{H}|\boldsymbol{x})$, whose behavior will indicate what an adequate conclusion is.

By choosing the probabilities α_1 and α_2 ($\alpha_1 \leq \alpha_2$) and setting a threshold ε, we obtain the respective quantiles of $\inf_{h\in H_0} d(h,\mathcal{H}(\boldsymbol{x}))$. Then, the following procedure applies:

- If $\varepsilon < q_{\alpha_1}\left[\inf_{h\in H_0} d(h,\mathcal{H}(\boldsymbol{x}))\right]$, we reject the hypothesis;
- If $\varepsilon \geq q_{\alpha_2}\left[\inf_{h\in H_0} d(h,\mathcal{H}(\boldsymbol{x}))\right]$, we accept the hypothesis;
- Otherwise, we remain agnostic.

In general, we assume that $\alpha := \alpha_1 = 1-\alpha_2$ and use a symmetrical interval instead. Still, if $\alpha_2 = \alpha_1$, this test would be equivalent to the one at Equation 3.11.

Being an agnostic test, this procedure automatically solves issue **(i)**. Also, being derived from a pragmatic hypothesis, it addresses issue **(iii)**. And, even though it does not resolve issue **(ii)**, the test adheres to monotonicity, which is a property of logically coherent procedures.

Theorem 3 (Monotonicity property)**.** Take $Pg(H_0)$ as the nonparametric pragmatic hypothesis of H_0 and $\mathcal{H}(\boldsymbol{x})$ as a method that randomly draws elements of $\mathbb{H}$ based on a sample $\boldsymbol{x}$. Then, for a precise hypothesis $H_0 \subset \mathbb{H}$ and $\alpha_1, \alpha_2 \in (0,1)$ where $\alpha_2 \geq \alpha_1$, if a test is such that

- If $\varepsilon < q_{\alpha_1}\left[\inf_{h\in H_0} d(h,\mathcal{H}(\boldsymbol{x}))\right]$, we reject the hypothesis;
- If $\varepsilon \geq q_{\alpha_2}\left[\inf_{h\in H_0} d(h,\mathcal{H}(\boldsymbol{x}))\right]$, we accept the hypothesis;
- Otherwise, we remain agnostic;

it obeys the property of monotonicity, i.e., it is logically coherent for any two precise hypotheses $H_0^1, H_0^2 \subset \mathbb{H}$ if $H_0^1 \supseteq H_0^2$, as shown in Table 3.

Table 3 – Possibility of each combination of decisions

Decisions	Reject H_0^1	Undecided on H_0^1	Accept H_0^1
Reject H_0^2	Possible	Possible	Possible
Undecided on H_0^2	Impossible	Possible	Possible
Accept H_0^2	Impossible	Impossible	Possible

3.3 Commentaries on the choice of ε

In the whole theory of pragmatic hypotheses, one of the greater points of concern resides in the choice of ε. Of course, this is a problem that should become less relevant as time goes on, since the more researchers apply these tests, the easier it will get to develop intuitions about reasonable values for ε. However, since its choice adds yet another layer of subjectivity, a procedure that automatically suggests values of ε would be greatly appreciated.

In some cases (such as in subsection 4.2.1) there is available information which allows one to directly derive a value for ε. In others, the researcher may be able to explicitly define what are negligible deviations from H_0 or to propose an educated guess, which opens room for more subjective evaluations. There still may be cases in which there is available information, but not enough to fully identify ε, or where there is no information at all.

If, for some reason, there is a need of a general procedure that could help delimit ε, some possibilities are

- Sampling data from H_0 multiple times and choosing the smallest ε such that, for a fixed $\alpha \in (0,1)$, the test accepts the hypothesis with frequency $(1-\alpha)$. This strategy has the drawback of requiring previous knowledge of the sample size that will be used and may not be feasible for all restriction types. Ideally, $Pg(H_0)$ should be established without any dependence on the sample size.
- Choosing two (or more) cases in the hypothesis space that have a negligible difference and equate the dissimilarity between them (or the highest one observed) as an estimate for ε. This represents a lower bound for the true ε and, in case the test leads to the acceptance of H_0, provides exactly the same conclusion that ε would.

- Ascribing a grid of reasonable values for ε (or sampling multiple values based on a probability distribution), registering the conclusion for each value and taking the most common decision as the final one, like a voting procedure. Even though this adds yet another layer of subjectivity, it does not require the researcher to compromise to a specific ε. Such procedure follows a similar inspiration than that of a prior distribution, an essential concept in Bayesian inference (ROBERT, 2007).

- Plot the decision as a function of α and ε and, instead of compromising to a specific value of ε, evaluate what is the reasonable decision to make based on the figure. Although this alternative might not apply for all contexts, it can completely sidestep the issue of choosing ε in some clear-cut cases, such as the one presented in subsection 4.2.3.

CHAPTER

4

APPLICATIONS

This section details multiple applications of the NPHT, covering different hypothesis types and dissimilarity functions. For all nonparametric tests, the method chosen for $\mathcal{H}(\cdot)$ was the Pólya tree process (LAVINE, 1992; LAVINE, 1994), even though any other method (frequentist or Bayesian) that draws from $\mathbb{H}$ based on data is valid. This prior process has all the desirable properties presented at subsection 3.2.3 (adaptability, generality and scalability) and possesses the attribute of conjugacy as well, justifying its choice. We mention, however, that such process has an infinite number of parameters, so the partially specified Pólya tree introduced in section A.4 was used instead. Since it is possible to control how close it can get to the true process, this caveat was not considered to be particularly detrimental.

4.1 Application 1: Neuron Data Analysis

This application is meant as an illustrative example of the use of NPHT and does not necessarily reflect the assumptions and procedures of neuroscientists. Rather, it uses the context of neurons to derive tests that adhere to the limitations proposed in this thesis and to demonstrate how to deal with practical issues that may arise throughout the analysis. The data used is publicly available and can be accessed through the original article (FARAUT *et al.*, 2018).

In the original dataset, 42 human epilepsy patients were exposed to a series of experiments where pictures of different settings were used as stimuli. For each patient, their brain activity on the amygdala and hippocampus during the experiments was recorded through the use of depth electrodes, which allowed for the identification of clusters of activity, which were assumed to be of individual neurons. When combining all patients, a total of 1576 individual neurons were identified.

For each neuron, the dataset provides timestamps (in microseconds) for each time its activity spiked throughout a given experiment (the same experiment could be performed in multiple sessions, so we use the notation “a-b” to represent session b of experiment a). Thus,

two questions that naturally arise are: **1.** how to model the spiking behavior and **2.** if the neuron activity is the same across experiments (i.e., if the spiking frequency is affected by the type of picture being shown to the patient). Since neuron spikes through time are a type of counting process, a Poisson process such as presented in Example 1 seems like a valid candidate. And, as for the spike frequency between experiments, evaluating if the same neuron keeps the same distribution function when exposed to different experiments might be too strict. Rather, it seems better to check if the median time between spikes (to avoid the influence of outliers) is consistent throughout experiments.

For both tests, we first need to choose a model $\mathcal{H}(\cdot)$ that can adapt to the data. As was mentioned at the beginning of the chapter, a Pólya tree process (LAVINE, 1992; LAVINE, 1994) will be used, but there is still a need to establish a distribution that will be used as its starting point. As Figure 3 shows, the mean of the time between spikes for each combination of neuron and experiment varies considerably, so trying to come up with a "one-size-fits-all" centering distribution is ill-advised. Instead, we chose to do the following: for each neuron, an experiment is selected at random and removed from the original sample to inform about the centering distribution. We used a gamma distribution and the maximum likelihood estimator based on such data for its parameters.

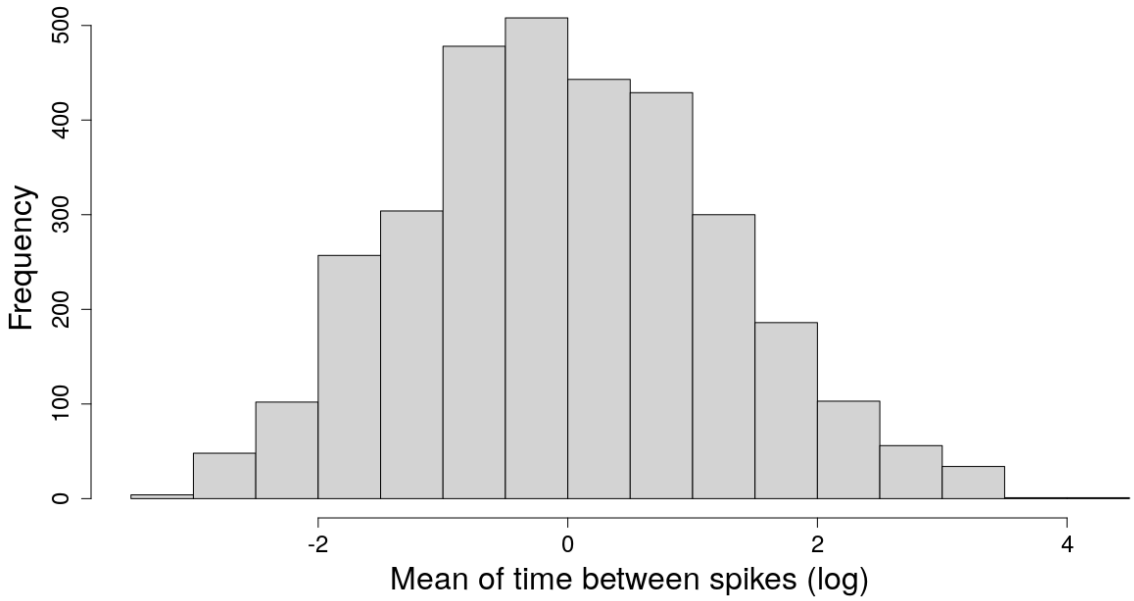


Figure 3 – Histogram of the mean of the time between spikes (log scale)

Since this application is meant to be illustrative rather than exhaustive, we restricted the analysis to a single neuron between multiple experiments. The chosen neuron was the one with index "2494" due to it having a high number of experiments applied (8 in total) and a reasonably high sample size in each experiment (minimum of 693, maximum of 2691). Figure 4 shows the estimated density (no model applied yet) for each experiment. This plot alone already puts the

assumption of a Poisson process into question, since some cases exhibit a bimodal behavior with peaks not that close to 0.

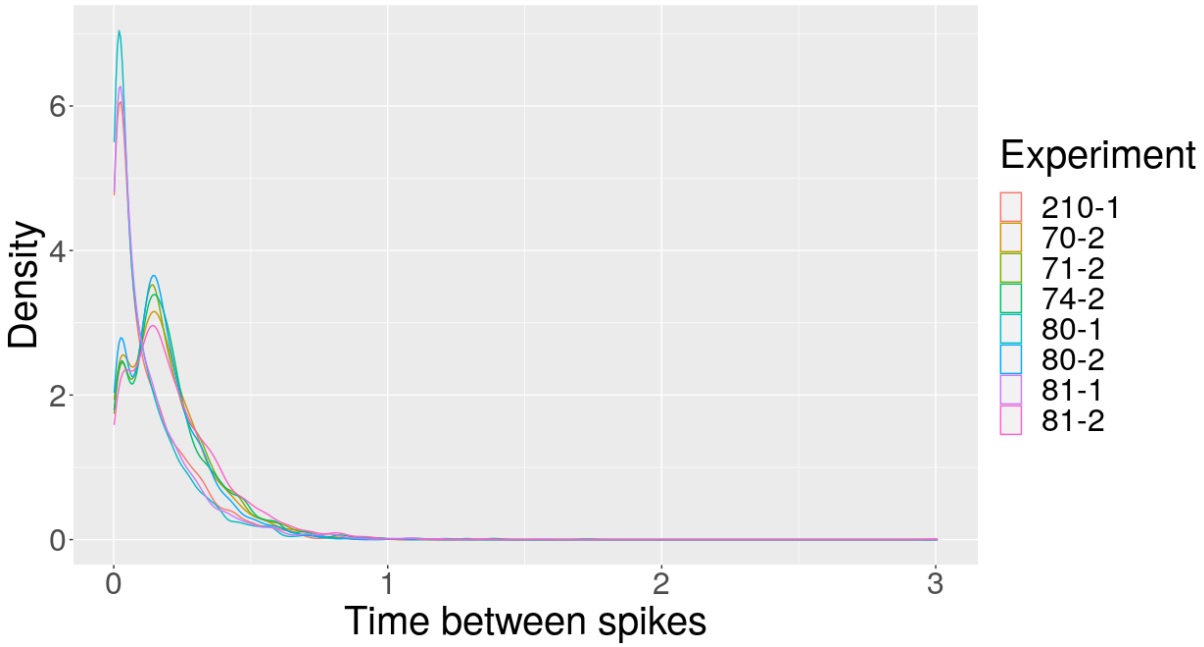


Figure 4 – Density of the time between spikes of neuron "2494" for each experiment

4.1.1 First test: Poisson process

If X is the random variable that represents the time between spikes, then

$$H_0 : \text{The data follows a Poisson process} \iff H_0 : X \overset{iid}{\sim} Exp(\lambda), \lambda \in \mathbb{R}^+.$$

This hypothesis offers a parametrical restriction (subsubsection 3.2.2.1), which implies that the pragmatic hypothesis is given by

$$Pg(H_0) = \left\{ F \in \mathbb{F} : \inf_{\lambda \in \mathbb{R}^+} d(G_\lambda, F) < \varepsilon \right\}.$$

Then, we chose the classification dissimilarity (Equation 3.6) and sampled 500 distributions for each experiment. Then, for each distribution, we found the value for λ that minimized the dissimilarity function. As for ε, instead of choosing it directly, we fixed $\alpha = 0.05$ and evaluated what value ε should take to avoid the rejection of H_0. Table 4 shows that this would require a value of ε of over 0.6, which means that we would be able to correctly identify the true distribution of a new observation 60% of the time, a considerably high number.

Experiment	70-2	71-2	74-2	80-1	80-2	81-1	81-2	210-1
ε	0.7875	0.7690	0.6950	0.7820	0.7895	0.7495	0.6005	0.7900

Table 4 – Minimum value of ε needed to not reject H_0 ($\alpha = 0.05$).

4.1.2 Second test: Median of time between spikes

Since this second test is a particular case of Example 2 ($p_0 = 0.5$), the testing procedure itself is rather straightforward, with only two challenges remaining: what value to assume for x_0 and how to derive a minimally adequate proposal for ε (unlike subsection 4.1.1, there is no clear intuition of what values of the dissimilarity function could be considered "too large"). For the former, we use once again the experiment that was removed from the original data, which provides a sample median of around 0.1787 second between spikes, implying that the null hypothesis can be expressed approximately as

$$H_0 : G(0.1787) = 0.5, \quad G \in \mathbb{F}.$$

As for the latter, we follow the second suggestion of section 3.3 and derive a lower bound for ε based on cases which are sure to be indistinguishable in practice.

To obtain such lower bound, we use the fact that a neuron spike typically lasts for 1 millisecond, thus no spike could happen in that interval (this is corroborated by the fact that, in the whole dataset, the minimum time observed between spikes is 0.0016 second, i.e, 1.6 milliseconds). Thus, by proposing two distribution functions, F_A and F_B, that can reasonably represent the data and whose difference could be attributed to the 1 millisecond threshold, it would then be possible to obtain the lower bound for ε by taking $d_1(F_A, F_B)$.

Since in subsection 4.1.1 we were led to reject the hypothesis that the time between spikes is exponential, we model F_A and F_B based on gamma distributions. Even though, as shown in Figure 4, part of the data presents a bimodal distribution, we judged that this behavior was not so severe as to invalidate the gamma density as an approximation. Then, if $tol = \pm 0.001$ (the tolerance between spikes is of 1 millisecond), the following procedure was used:

1. Take $A \sim Gamma(\hat{\alpha}, \hat{\beta})$, where $\hat{\alpha}$ and $\hat{\beta}$ are the maximum likelihood estimates obtained from the experiment that was removed from the sample;

2. Take $B \sim Gamma(\alpha^*, \beta^*)$ such that $\mathbb{E}[B] = \mathbb{E}[A] + tol$ (i.e., the means differ by at most 1 millisecond) and $\mathbb{V}[B] = \mathbb{V}[A]$ (i.e., the variance is not affected);

3. Obtain $d_1(F_A, F_B)$ for all possible tolerance values and take the greatest result as the lower bound for ε.

Such procedure implies that $\beta^* = \hat{\beta} + tol \times (\hat{\beta}^2/\hat{\alpha})$ and $\alpha^* = \hat{\alpha} \times (\beta^*/\hat{\beta}) + tol \times \beta^*$. With this information at hand, the lower bound obtained is of about 0.00101, quite close to the tolerance we started with in the first place.

As Table 5 shows, experiment 81-2 seems to be the one that more closely adheres to the null hypothesis, requiring $\alpha \geq 0.126$ to lead to the rejection of H_0, which implies that it should not be rejected. All other experiments seem to lean towards the rejection of H_0, with the

Experiment	Sample size	Sample median	α for rejecting H_0
70-2	**693**	**0.1651**	**0.020**
71-2	2388	0.1668	0.000
74-2	**1834**	**0.1718**	**0.014**
80-1	2487	0.0693	0.000
80-2	1919	0.1601	0.000
81-1	2691	0.0785	0.000
81-2	**1547**	**0.1793**	**0.126**
210-1	2279	0.0795	0.000

Table 5 – Comparison between experiments based on the sample median and the smallest value of α that would lead to the rejection of H_0.

conclusion of experiments 70-2 and 74-2 being conditional on the choice of α (if $\alpha = 0.01$, both would not be rejected). This seems to be due to the fact that their sample medians are not that distant from the one in H_0 and their sample sizes are smaller than most of the others (notice how the value of α for experiment 70-2 is bigger than the one for experiment 74-2, even though the latter has a sample median closer to H_0. This indicates that the higher value of α in 70-2 is a consequence of its smaller sample size, that there is still great uncertainty on the function generating process $\mathcal{H}(\boldsymbol{x})$).

In general, Table 5 points out that there really is a difference on neuron spikes based on the experiment being performed. Still, in some cases, there could be an equivalence between tests for a given neuron, which in turn could suggest that some procedures are redundant and could be eliminated without a significant loss of information, saving time and resources.

4.2 Application 2: The Water Droplet Experiment

The free falling water droplet experiment (DUGUID, 1969) was a study that seeked to evaluate the behavior of small water droplets (ranging from 3 to 9 micrometers) as they fall through a tube. More specifically, one of the main interests of the research was to test the validity of Fick's law of diffusion in a controlled setting. In this particular case, the law posits that, as the droplet falls through the drift tube, its radius decreases linearly through time. Of course, controlling all other variables is essential to ensure that the effect is genuinely described. Thus, to ensure that all other factors of influence remain constant throughout the experiment, the apparatus shown in Figure 5 was used.

The top part of the apparatus (diffusion cloud chamber) is responsible for generating droplets within the specified radius of interest. To produce the droplets, the inside of the apparatus requires humidity and air particles. However, the particles should not be too big, or else the droplets will become too contaminated by them. Thus, room air is pumped into the apparatus through the nuclei inlet, but only particles small enough to pass through the filter actually get inside, reducing the contamination. As for the humidity, a humidifier was inserted at the bottom

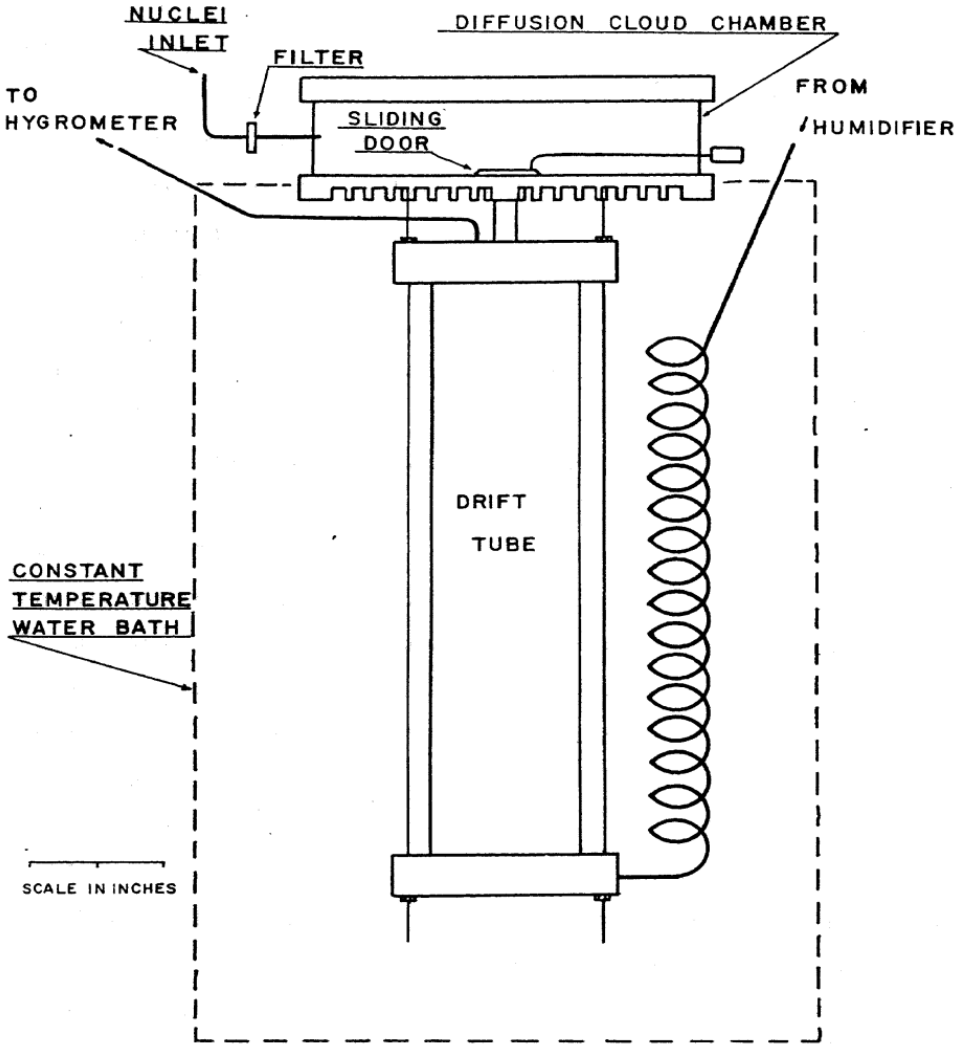


Figure 5 – Drop generator and drift tube (Source: Duguid (1969))

of the apparatus.

Once the droplet is formed, the sliding door is opened, letting the water fall through the drift tube. To ensure that temperature variations do not influence the observations, the apparatus was put on a water bath with constant temperature.

As the droplet falls, a camera takes pictures of the apparatus every 0.5 second. Since this was not enough to infer the actual position of the drop, a grid with precise measurements was put in front of the apparatus so as to be used as a ruler and provide an estimate. This, of course, is a source of measurement error, leading to an error of around $\delta = \pm 0.07$ for ΔS, the change in position between picture frames.

We highlight the fact that, although the main interest resides in retrieving the radius of the droplet as it falls, it was never directly observed in the experiment. To actually obtain it, Stokes' law was applied, which asserts that

$$V_T = \frac{a^2}{K_s} \Longrightarrow a = \sqrt{V_T K_s}, \tag{4.1}$$

where V_T is the terminal velocity of the object, a is the radius and K_s is a known constant that depends on factors such as temperature and humidity. This is not a serious issue, since the Stokes' law adequately works for small objects such as the droplets of the experiment. Hence, we assume that the formula holds true for this case.

This does not mean, however, that using (4.1) was enough to provide a trustworthy estimate of the radius. In the experiment of Duguid (1969), the mean velocity ($V_M = \frac{\Delta S}{\Delta t} = 2\Delta S$) was used instead of the terminal velocity, leading to yet another source of measurement error. Unfortunately, no estimate of this error was provided in the original paper.

With all these caveats in mind, Table 6 provides the data of a single water droplet as it falls. In this particular case, $K_s = 8.446$.

Table 6 – Data for an evaporating water droplet (Source: Duguid (1969))

Film frame	Position of drop (S, in mm)	ΔS	Elapsed time (sec)
1	4.7	NA	0
2	8.1	3.4	0.5
3	11.5	3.4	1
4	14.5	3.0	1.5
5	17.4	2.9	2
6	19.9	2.5	2.5
7	22.4	2.5	3
8	24.6	2.2	3.5
9	26.6	2.0	4
10	28.5	1.9	4.5
11	30.1	1.6	5
12	31.5	1.4	5.5
13	32.8	1.3	6
14	33.8	1.0	6.5
15	34.6	0.8	7

4.2.1 Margin of error of the experiment

Now that the experiment has been fully described, it is necessary to identify a margin of error for the radius of the droplet. As it was mentioned, the two main sources of error are in the estimation of ΔS and in exchanging V_T for V_M in Stokes' law. If $T = \{0, 0.5, \cdots, 6.5, 7\}$ and $\eta = \max_{t \in T} |V_T(t) - V_M(t)|$, then it follows from Equation 4.1 that

$$a(t) = \sqrt{K_s V_T(t)} = \sqrt{K_s (V_M(t) \pm \eta)} = \sqrt{K_s (2\Delta S(t) \pm \delta) \pm \eta)}. \tag{4.2}$$

If $\hat{a}(t) = \sqrt{2K_s \Delta S(t)}$ is our estimate of the radius at time $t \in T$, the margin of error for the radius is $\varepsilon = \max_{t \in T} |a(t) - \hat{a}(t)|$. Thus, the only remaining element left to be identified is η.

The terminal velocity V_T is the instantaneous velocity of the droplet, which in turn is the derivative of the drop's position through time. Hence, $V_T = \frac{dS}{dt}$, which means that using a

method that is capable of estimating derivatives should shed light on what would be a reasonable margin of error. In this case, we follow the ideas presented in Wang and Lin (2015), which use symmetric differences of the response and the explanatory variable as a new response variable and apply locally weighted least squares regression to provide the estimate. For a fixed moment in time, k symmetric differences are calculated and used to estimate the derivative. In our case, due to the small sample and the suggestion in Wang and Lin (2015) that $k < n/4$, we opted to use $k = 4$. From this choice, it follows that the first and last k derivatives need to be calculated through a slightly different method due to lack of data. This change results in poorer estimates, which, along with the fact that the sample size is small for our case, suggests that they should not be as highly regarded as the estimates in the middle.

Table 7 shows the respective errors for both $|V_T(t) - V_M(t)|$ and $|a(t) - \hat{a}(t)|$. Since the method proposed by Wang and Lin (2015) results in poorer estimates at the boundaries, the first and last k estimates were disregarded in the analysis, so $\eta = \max_{t \in T} |V_T(t) - V_M(t)| \simeq 0.356$. Then, by plugging δ and η in $|a(t) - \hat{a}(t)|$, we observed the highest value for each $t \in T$, which then shows that $\varepsilon = \max_{t \in T} |a(t) - \hat{a}(t)| \simeq 0.622$. This means that, for practical purposes, any difference between radius below 0.622 micrometer is negligible for the analysis.

Table 7 – Margin of error for each $t \in T$

Film frame	$V_M = 2\Delta S$	$\lvert V_T(t) - V_M(t)\rvert$	$\lvert a(t) - \hat{a}(t)\rvert$	Elapsed time (sec)
1	NA	NA	NA	0
2	6.8	0.779	0.281	0.5
3	6.8	1.186	0.281	1
4	6.0	0.836	0.300	1.5
5	5.8	0.051	0.306	2
6	5.0	**0.356**	0.330	2.5
7	5.0	0.028	0.330	3
8	4.4	0.164	0.354	3.5
9	4.0	0.149	0.372	4
10	3.8	0.095	0.382	4.5
11	3.2	0.104	0.419	5
12	2.8	1.143	0.451	5.5
13	2.6	0.893	0.470	6
14	2.0	1.064	0.545	6.5
15	1.6	1.057	**0.622**	7

To end the discussion of this subsection, it is valid to ask why we do not use the estimated V_T instead of V_M. Some of the reasons are:

- V_T has not been directly observed, while V_M has
- Due to the small sample size, the estimates of V_T are questionable, especially at the boundaries. Using V_T could result in an even smaller sample.

- This change would provide a different setting than the one in Duguid (1969), which is our reference.

4.2.2 First test: validity of Fick's law

The first test consists of a parametric pragmatic test, so the methods presented in Chapter 2 apply in this case. In the context of the application, we assume that the data can be correctly described by a polynomial model of, at most, degree p. And, since we are interested in testing the validity of Fick's law for this case, i.e., that the data could be described by a line, then

$$a(t) = \sum_{i=0}^{p} \beta_i t^i + \epsilon(t), \quad \epsilon \overset{iid}{\sim} N(0, \sigma^2) \quad \Longrightarrow \quad H_0 : \beta_2 = \beta_3 = \cdots = \beta_p = 0.$$

This implies that, if H_0 is true, $(\boldsymbol{\beta}_0, \sigma_0) = (\beta_0, \beta_1, \beta_2, \cdots, \beta_p, \sigma_0) \in \Theta_0 = \mathbb{R}^2 \times \{0\}^{p-1} \times \mathbb{R}_+$. Then, the pragmatic hypothesis based on H_0 should be such that, for all $t \in T$, the predictive dissimilarity between any $(\boldsymbol{\beta}_0, \sigma_0) \in \Theta_0$ and a candidate $(\boldsymbol{\beta}^*, \sigma^*) \in \Theta$ should not surpass a threshold ν, i.e.,

$$Pg(H_0) = \bigcup_{(\boldsymbol{\beta}_0, \sigma_0) \in \Theta_0} \left\{ (\boldsymbol{\beta}^*, \sigma^*) \in \Theta : \max_{t \in T} d_Z[(t, \boldsymbol{\beta}_0, \sigma_0), (t, \boldsymbol{\beta}^*, \sigma^*)] \leq \nu \right\}. \tag{4.3}$$

Hence, to identify $Pg(H_0)$ and propose a testing procedure, the challenge is then to choose a predictive dissimilarity function that could allow us to relate ν and ε.

Definition 5. *(ESTEVES et al., 2019) Let $\hat{\mathbf{Z}} : \Theta \to \mathcal{X}$, where $\mathcal{X}$ represents all possible future observations, be such that $\hat{\mathbf{Z}}(\theta_0)$ is the best prediction for $\mathbf{Z}$ given that $\theta = \theta_0$. For example, one can take*

$$\hat{\mathbf{Z}}(\theta_0) = \arg\min_{z \in \mathcal{Z}} \delta_{\mathbf{Z}, \theta_0}(z),$$

where $\delta_{\mathbf{Z}, \theta_0} : \mathcal{X} \to \mathbb{R}$ is such that $\delta_{\mathbf{Z}, \theta_0}(z)$ measures how bad z predicts $\mathbf{Z}$ when $\theta = \theta_0$. The "best prediction dissimilarity", $BP_{\mathbf{Z}}(\theta_0, \theta^*)$, measures how badly $\hat{\mathbf{Z}}(\theta^*)$ predicts $\mathbf{Z}$ relatively to $\hat{\mathbf{Z}}(\theta_0)$ when $\theta = \theta_0$. Formally,

$$BP_{\mathbf{Z}}(\theta_0, \theta^*) = g\left(\frac{\delta_{\mathbf{Z}, \theta_0}(\hat{\mathbf{Z}}(\theta^*)) - \delta_{\mathbf{Z}, \theta_0}(\hat{\mathbf{Z}}(\theta_0))}{\delta_{\mathbf{Z}, \theta_0}(\hat{\mathbf{Z}}(\theta_0))} \right), \tag{4.4}$$

where $g : \mathbb{R} \to \mathbb{R}$ is a monotonic function. The choice of g in a particular setting aims at improving the interpretation of the best prediction dissimilarity criterion.

In particular, for the univariate case, if $g(x) = \sqrt{x}$, $\delta_{Z, \theta_0}(z) = \mathbb{E}[(Z - z)^2 | \theta = \theta_0]$ and Z is Gaussian, then (4.4) simplifies to

$$BP_Z(\theta_0, \theta^*) = \sigma_0^{-1} |\theta_0 - \theta^*|. \tag{4.5}$$

The best prediction dissimilarity, also known as BP, allows one to evaluate differences between predictions of the response variable directly. Then, based on Equation 4.5, for a fixed $t, \boldsymbol{\beta}_0, \sigma$ and $\boldsymbol{\beta}^*$,

$$BP_Z[(t,\boldsymbol{\beta}_0,\sigma_0),(t,\boldsymbol{\beta}^*,\sigma^*)] = \sigma_0^{-1}|(\boldsymbol{\beta}_0-\boldsymbol{\beta}^*)'\boldsymbol{t}_p| \leq \sigma_0^{-1}\varepsilon,$$

where $\boldsymbol{t}_p = (1,t,t^2,\cdots,t^p)'$. This implies that, if $\nu = \sigma_0^{-1}\varepsilon$, then Equation 4.3 simplifies to

$$Pg(H_0) = \bigcup_{\boldsymbol{\beta}_0 \in H_0} \left\{ (\boldsymbol{\beta}^*,\sigma^*) \in \Theta : \max_{t\in T} |(\boldsymbol{\beta}_0-\boldsymbol{\beta}^*)'\boldsymbol{t}_p| \leq \varepsilon \right\}. \tag{4.6}$$

Given that the model consists of a polynomial regression with normally distributed errors, the MLE and its probability distribution is known. If

$$T_p = ((1,0.5,0.5^2,\cdots,0.5^p)',(1,1,1^2,\cdots,1^p)',\cdots,(1,7,7^2,\cdots,7^p)')',$$

$H = (T_p'T_p)^{-1}T_p'$ and $\boldsymbol{a}$ is the vector of the response variable (radius of the droplet), then $\hat{\boldsymbol{\beta}} = H\boldsymbol{a}$ is the MLE of $\boldsymbol{\beta}$ and

$$\hat{\boldsymbol{\beta}} \sim N_{p+1}\left[\boldsymbol{\beta}, \sigma^2(T_p'T_p)^{-1}\right] \Rightarrow (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'\boldsymbol{t}_p \sim N\left[(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{t}_p, \sigma^2\boldsymbol{t}_p'(T_p'T_p)^{-1}\boldsymbol{t}_p\right] \tag{4.7}$$

Thus, for a fixed $\boldsymbol{\beta}_0$ and a confidence level of $1-\alpha$, the confidence interval is

$$CI[(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{t}_p]_{(1-\alpha)} = (H\boldsymbol{a}-\boldsymbol{\beta}_0)'\boldsymbol{t}_p \pm q_t(\alpha/2, n-(p+1))\sqrt{\hat{\sigma}^2\boldsymbol{t}_p'(T_p'T_p)^{-1}\boldsymbol{t}_p}, \tag{4.8}$$

where $q_t(.,n-(p+1))$ is the quantile function of the Student's-t distribution with $n-(p+1)$ degrees of freedom and $\hat{\sigma}^2$ is the MLE of σ^2. With this information in mind, we can then derive an agnostic test that will:

- Accept H_0 if $\exists \boldsymbol{\beta}_0 \in H_0$ such that $CI[(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{t}_p]_{(1-\alpha)} \subseteq [-\varepsilon,\varepsilon]$, $\forall t \in T$;
- Remain agnostic about H_0 if the above condition is not true, but $\exists \boldsymbol{\beta}_0 \in H_0$ such that $CI[(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{t}_p]_{(1-\alpha)} \cap [-\varepsilon,\varepsilon] \neq \emptyset$, $\forall t \in T$;
- Reject H_0 if, $\forall \boldsymbol{\beta}_0 \in H_0$ and $t \in T$, $CI[(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{t}_p]_{(1-\alpha)} \cap [-\varepsilon,\varepsilon] = \emptyset$.

Of course, $\boldsymbol{\beta}_0 \in H_0$ represents an infinite set of pairs $(\beta_0,\beta_1) \in \mathbb{R}^2$, which could be complex to evaluate in the test. Some possible solutions are:

- Reduce the complexity of the problem by evaluating the test only for $\hat{\boldsymbol{\beta}}_0$, i.e., the MLE of the model under H_0.

- Use the context of the problem to derive a grid of reasonable values of (β_0, β_1). In the case of the experiment, given that we know that the radius of the droplet ranges between 3-9 micrometers and that it only decreases through time, we could check the hypothesis only for a grid of (β_0, β_1) that does not result in unreasonable values. Thus, $\beta_0 \in [3, 9]$ and $\beta_1 \in [-\beta_0/\max(t), 0] = [-\beta_0/7, 0]$.
- For a large enough M, sample M values from the distribution of $\hat{\boldsymbol{\beta}}_0$ and check the result for each of them. Since they are all reasonable candidates of the true $\boldsymbol{\beta}_0$, these comparisons should be enough to get to a trustworthy statement.

In our case, using the first strategy was enough, since it led to the acceptance of the null hypothesis. Figure 6 shows the confidence interval for each value of $t \in T$ when $\boldsymbol{\beta}_0 = \hat{\boldsymbol{\beta}}_0$. Since no interval leaves the tolerance region (which is represented by the blue area), we can safely accept the hypothesis that Fick's law is valid for the water droplet experiment.

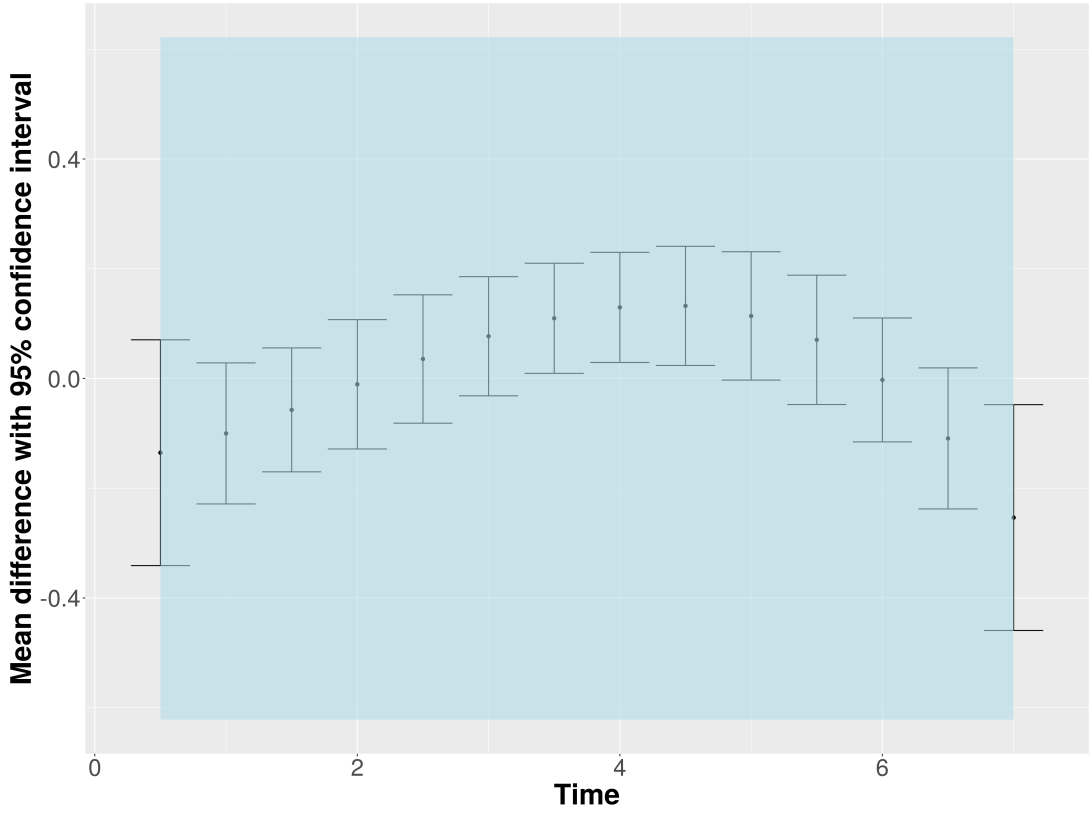


Figure 6 – Confidence interval for the mean difference between the MLE under $p = 3$ and $p = 1$ (the blue area represents the tolerance region of $[-\varepsilon, \varepsilon]$).

As for the Bayesian case, if we take the noninformative prior $p(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$, then

$$\boldsymbol{\beta}|\sigma^2, \boldsymbol{a}, T_p \sim N_{p+1}\left(\hat{\boldsymbol{\beta}}, \sigma^2 (T_p' T_p)^{-1}\right) \Rightarrow (\boldsymbol{\beta} - \boldsymbol{\beta}_0)' \boldsymbol{t}_p | \sigma^2, \boldsymbol{a}, T_p \sim N\left((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' \boldsymbol{t}_p, \sigma^2 \boldsymbol{t}_p' (T_p' T_p)^{-1} \boldsymbol{t}_p\right),$$

which implies that, by integrating σ out, we conclude that

$$\left. \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}_0)' \boldsymbol{t}_p - (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' \boldsymbol{t}_p}{\sqrt{\hat{\sigma}^2 \boldsymbol{t}_p' (T_p' T_p)^{-1} \boldsymbol{t}_p}} \right| \boldsymbol{a}, T_p \sim t_{n-(p+1)}.$$

Hence, for this prior choice, both the frequentist confidence interval and the Bayesian credible interval coincide in its values, leading then to the same conclusion in both cases. This is interesting because it shows that it is possible to simply use this procedure to derive conclusions no matter which Statistics' school of thought the researcher adheres to.

4.2.3 Second test: residual normality

After subsection 4.2.2, in which the hypothesis that the droplet radius decreases linearly through time was accepted, it is time to test the validity of different aspects of the model, such as the normality of the residuals. This is a case of NPHT, since our hypothesis of interest is $H_0 : \epsilon \sim N(0, \sigma^2), \sigma^2 \in \mathbb{R}^+$. It is a clear case of a parametrical restriction as presented in subsubsection 3.2.2.1, so all that remains to be done is to choose a dissimilarity function, a prior process and a threshold ε.

For this context, we opted to use the classification dissimilarity specified in Equation 3.6 and a Pólya tree process with the centering distribution being a Student-t with 3 degrees of freedom. The first choice was due to the intuitive appeal of the classification dissimilarity, while the second was because such centering distribution provides longer tails while having a variance of 1 (this is a controlled experiment, so it is natural to assume that the variance of the variable is relatively small).

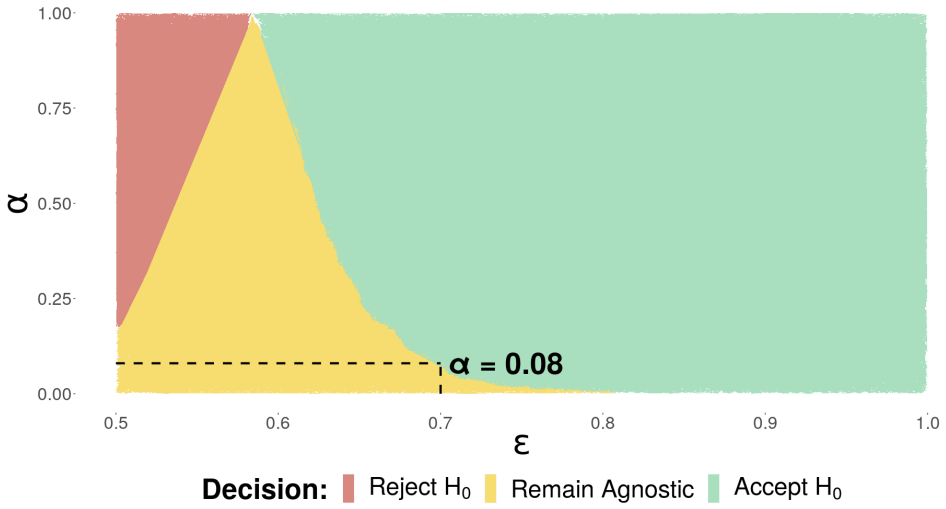


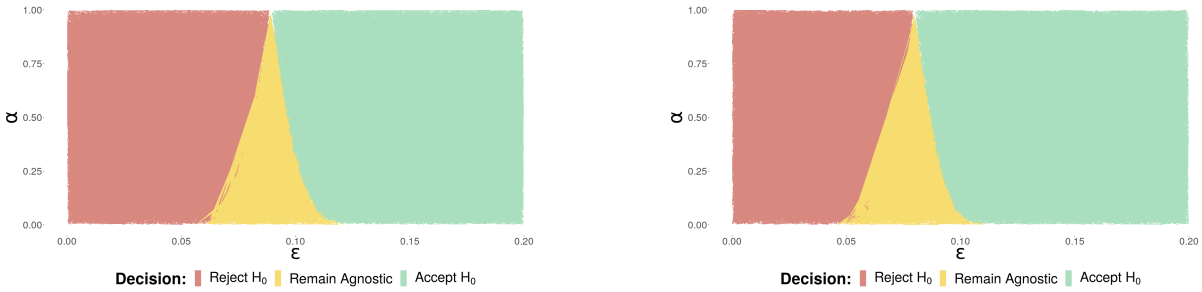
Figure 7 – Decision region for H_0 as a function of α and ϵ

Figure 7 shows the decision region as a function of α and ε. As it can be seen, rejecting H_0 would require α to be around 0.2, which is considerably high. As for accepting H_0, a considerably high ε would be required to reach this conclusion for a reasonable α. The figure shows that, if $\alpha = 0.08$, we would only accept the hypothesis if we were willing to correctly identify the data generating process above 70% of the time. Thus, remaining agnostic seems to be the most reasonable conclusion in this case, which makes sense due to the fact that the sample size is rather small.

4.3 Application 3: Pseudorandom Number Generators

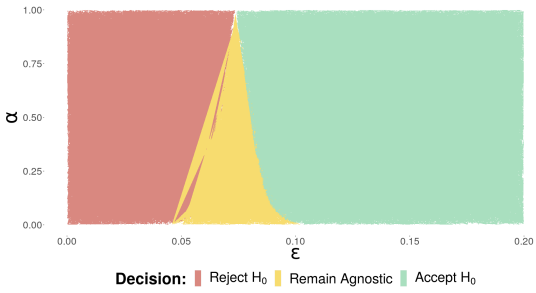
Pseudorandom number generators (PRNG) are methods that allow the user to draw numbers that behave similarly to a truly random number generator (RNG). In general, the objective of such procedures are to generate numbers that closely resemble those of a $U(0,1)$, due to the fact that transforming these results allow for draws similar to those of many other distributions. Using PRNG instead of RNG is usually due to two factors: reproducibility of the results and that PRNG can draw numbers considerably faster than RNG.

Since PRNG are by design deterministic, it is of utmost importance to ensure that they provide results considerably similar to a true $U(0,1)$, especially when it comes to correlated results. Thus, this section proposes to compare how similar different procedures are, i.e., if one procedure generates a sample $\boldsymbol{X}$ and the other generates a sample $\boldsymbol{Y}$, we wish to test if $G_X = G_Y$. This procedure is exactly the one described in Example 3.

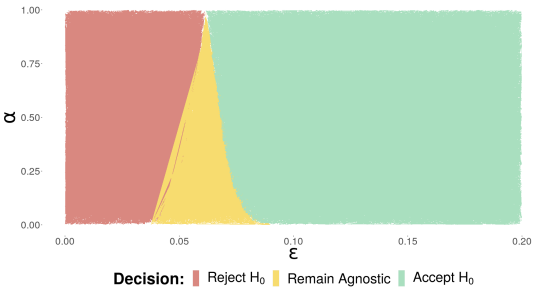


(a) $n = 500$

(b) $n = 1000$



(c) $n = 2500$



(d) $n = 5000$

Figure 8 – Decision region of the comparison between pseudorandom number generators at different sample sizes

Figure 8 shows the decision region when comparing the Mersenne-Twister and the Marsaglia-Multicarry methods for different sample sizes. We once more used a Pólya tree process, this time centered around a $U(0,1)$. As for the distance function, we simply subtracted 0.5 to the classification dissimilarity. It is clear that, as the sample size increases, the agnostic region gets smaller, but also that there always seem to remain a difference between the two methods. Thus, if one had a sample of a truly uniform distribution available, they could use this procedure as yet another method of testing how trustworthy a given PRNG is.

CHAPTER

5

CONCLUSION AND FUTURE WORK

Agnostic tests and pragmatic hypotheses offer a new paradigm for hypothesis testing, one that is simultaneously theoretically defensible and adaptable to practical settings. In particular, the NPHT (its nonparametric counterpart) has already shown promising results, offering adaptations to tests that are widely used throughout the literature. Still, since this field of research is relatively recent, there are still many gaps to be filled, which are the next steps this research intends to take.

Throughout Chapter 4, figures of the decision region were used to derive conclusions about each agnostic test. Of course, if a confidence/credible level α and a threshold ε are already set, such figures are unnecessary, since there can be a clear-cut conclusion. Still, we find such visualizations compelling, since they provide a more complete picture of the test being performed. In cases such as Figure 7, the agnostic region was so wide that the only reasonable conclusion seemed to be to remain agnostic. This implies that, to reach a more assertive decision, one should collect more data, repeat the experiment or change the centering distribution of the prior process (this third option is not ideal, since it could be seen as meddling with the test to achieve a desired result). However, if for some reason one must achieve an assertive result (either rejecting or accepting H_0) without changing attributes of the test, then it is a matter of deciding which factor is worth sacrificing more, α or ε.

So far, all of the proposed tests were able to tackle issues **(i)** and **(iii)**, while only partially dealing with **(ii)**. This is due to the fact that logical coherence can only be ensured if, instead of proposing intervals for $\inf_{h\in H_0} d(h, \mathcal{H}(\boldsymbol{x}))$, we were able to devise intervals for $\mathcal{H}(\boldsymbol{x})$ directly. Even in simpler cases such as $\mathcal{H}(\boldsymbol{x}) = F|\boldsymbol{x}$, where F is a Pólya tree process, proposing credible bands for such process requires finding distribution functions F_L and F_U such that, for a specific ν,

$$\mathbb{P}\left(F_L(y) \leq F_{Y|\boldsymbol{x}}(y) \leq F_U(y), \forall y \in \mathcal{X}\right) = \nu.$$

This is the natural next step of the research, since solving issue **(ii)** completely has the potential to make such methods even more appealing.

Another direction the research should follow is to delve deeper into theories that may help generalize the procedures even further. As was evident throughout subsubsection 3.2.2.2, there are still many tests which have not yet been explored, since the theory still lacks general guidelines. Promising areas that should help close these gaps, even if applicable in a limited set of problems, are functional analysis (KREYSZIG, 1978) and entropy methods (ABBAS; CADENBACH; SALIMI, 2017).

Lastly, frequentist methods still need to be proposed and evaluated for this context. Bayesian nonparametrics, while useful, is certainly not the only available method for drawing distributions based on data. One of the greatest advantages of the procedures proposed in this paper is that they work no matter the method used to sample distributions from. Thus, as long as one can draw distribution functions, either through frequentist or Bayesian methods, all tests should work exactly as intended. Cases where the context demands multivariate models are of special interest, given that only univariate solutions have been developed so far.

BIBLIOGRAPHY

ABBAS, A. E.; CADENBACH, A. H.; SALIMI, E. A kullback–leibler view of maximum entropy and maximum log-probability methods. **Entropy**, v. 19, n. 5, 2017. Citation on page 58.

CEREGATTI, R. de C.; IZBICKI, R.; SALASAR, L. E. B. Wiks: a general bayesian nonparametric index for quantifying differences between two populations. **TEST**, v. 30, p. 274–291, 2021. Citation on page 35.

COSCRATO, V.; ESTEVES, L. G.; IZBICKI, R.; STERN, R. B. **Interpretable hypothesis tests**. 2019. Citations on pages 11, 13, 15, 24, 25, 26, 27, and 39.

COSCRATO, V.; IZBICKI, R.; STERN, R. B. Agnostic tests can control the type I and type II errors simultaneously. **Brazilian Journal of Probability and Statistics**, Brazilian Statistical Association, v. 34, n. 2, p. 230 – 250, 2020. Citation on page 25.

DUGUID, H. A. **A study of the evaporation rates of small freely falling water droplets**. Master's Thesis (Master's Thesis) — Missouri University of Science and Technology, Rolla, Missouri, USA, 1969. Citations on pages 15, 17, 47, 48, 49, and 51.

ESTEVES, L. G.; IZBICKI, R.; STERN, J. M.; STERN, R. B. Pragmatic hypotheses in the evolution of science. **Entropy**, v. 21, n. 9, 2019. ISSN 1099-4300. Available: <https://www.mdpi.com/1099-4300/21/9/883>. Citations on pages 24, 26, 27, 39, and 51.

FARAUT, M. C.; CARLSON, A. A.; SULLIVAN, S.; TUDUSCIUC, O.; ROSS, I.; REED, C. M.; CHUNG, J. M.; MAMELAK, A. N.; RUTISHAUSER, U. Dataset of human medial temporal lobe single neuron activity during declarative memory encoding and recognition. **Scientific Data**, v. 5, n. 1, 2018. Citation on page 43.

FERGUSON, T. S. A Bayesian Analysis of Some Nonparametric Problems. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 1, n. 2, p. 209 – 230, 1973. Available: <https://doi.org/10.1214/aos/1176342360>. Citation on page 37.

____. Prior Distributions on Spaces of Probability Measures. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 2, n. 4, p. 615 – 629, 1974. Available: <https://doi.org/10.1214/aos/1176342752>. Citations on pages 24 and 37.

HOLMES, C. C.; CARON, F.; GRIFFIN, J. E.; STEPHENS, D. A. Two-sample Bayesian Nonparametric Hypothesis Testing. **Bayesian Analysis**, International Society for Bayesian Analysis, v. 10, n. 2, p. 297 – 320, 2015. Available: <https://doi.org/10.1214/14-BA914>. Citations on pages 35 and 61.

INáCIO, M. H. de A.; IZBICKI, R.; SALASAR, L. E. Comparing two populations using bayesian fourier series density estimation. **Communications in Statistics - Simulation and Computation**, Taylor & Francis, v. 49, n. 1, p. 261–282, 2020. Available: <https://doi.org/10.1080/03610918.2018.1484480>. Citation on page 35.

KREYSZIG, E. **Introductory Functional Analysis with Applications**. [S.l.]: Wiley, 1978. (Wiley classics library). ISBN 9780471507314. Citations on pages 29, 58, and 70.

KUHN, T. S. **The Structure of Scientific Revolutions**. [S.l.]: Chicago: University of Chicago Press, 1962. Citation on page 23.

LAVINE, M. Some aspects of polya tree distributions for statistical modelling. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 20, n. 3, p. 1222–1235, 1992. ISSN 00905364. Available: <http://www.jstor.org/stable/2242010>. Citations on pages 24, 38, 43, 44, 61, 62, and 63.

____. More aspects of polya tree distributions for statistical modelling. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 22, n. 3, p. 1161–1176, 1994. ISSN 00905364. Citations on pages 24, 38, 43, 44, 61, 63, and 65.

MAYO, D. G. **Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars**. [S.l.]: Cambridge University Press, 2018. Citation on page 24.

Open Science Collaboration. Estimating the reproducibility of psychological science. **Science**, American Association for the Advancement of Science, v. 349, n. 6251, 2015. Citation on page 23.

PHADIA, E. G. **Prior Processes and Their Applications: Nonparametric Bayesian Estimation**. 2. ed. [S.l.]: Springer International Publishing, 2016. Citations on pages 30, 64, and 66.

POPPER, K. **The Logic of Scientific Discovery**. [S.l.]: Routledge, 1934. Citation on page 25.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2020. Available: <https://www.R-project.org/>. Citation on page 38.

ROBERT, C. **The Bayesian Choice**. [S.l.]: Springer, 2007. Citation on page 41.

ROBERT, C. P.; CASELLA, G. **Monte Carlo statistical methods**. 2. ed. Berlin: Springer, 2005. (Springer texts in statistics). Citation on page 34.

ROSS, S. M. **A First Course in Probability**. 8. ed. [S.l.]: Pearson, 2009. Citation on page 33.

SALSBURG, D. **The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century**. [S.l.]: Henry Holt and Company, 2002. Citation on page 23.

WANG, W.; LIN, L. Derivative estimation based on difference sequence via locally weighted least squares regression. **Journal of Machine Learning Research**, v. 16, n. 81, p. 2617–2641, 2015. Available: <http://jmlr.org/papers/v16/wang15b.html>. Citation on page 50.

APPENDIX

A

PÓLYA TREE PROCESS

This section is based on the results presented in Lavine (1992) and Lavine (1994). However, the notation used follows more closely that of Holmes *et al.* (2015), since it is better suited for the context of density estimation. It is important to remind that, since the Pólya Tree is a nonparametric prior, the distribution function is a random variable itself and, in this particular case, requires an infinite number of parameters. Thus, in order to sample from the data, one must either obtain a fixed distribution first or integrate the parameters out.

A.1 Formalization of the Model

In order to fully specify the model, we first need to establish the set Π of probability measures of interest. In the case of the Pólya tree, Π is a collection of separable binary trees of partitions of Ω, the sample space. This means that, at the m-th level of the partition, we have a collection of sets $\{B_i^{(m)}; i = 0, \cdots, 2^m - 1\}$ such that

$$\bigcup_{i=0}^{2^m-1} B_i^{(m)} = \Omega; \quad B_i^{(m)} \cap B_j^{(m)} = \emptyset, \forall i \neq j; \quad B_i^{(m)} = B_{2i}^{(m+1)} \cup B_{2i+1}^{(m+1)}.$$

Figure 9 provides an example of such partitions. Take the blue dot as the point that divides the sample space, providing a partition. Once the sample space has been split, the point that splits the set in two persists at all following levels. Then, at the first level, there is a single partition, providing two interval sets that represent the sample space. At the second level, there are four sets. At the third level, there are eight sets, and so on.

From now on, we'll represent the subsets of Ω through base 2, allowing us to drop the superscript. Thus, B_0 represents the first set at the first level, B_{11} the fourth and last set of the

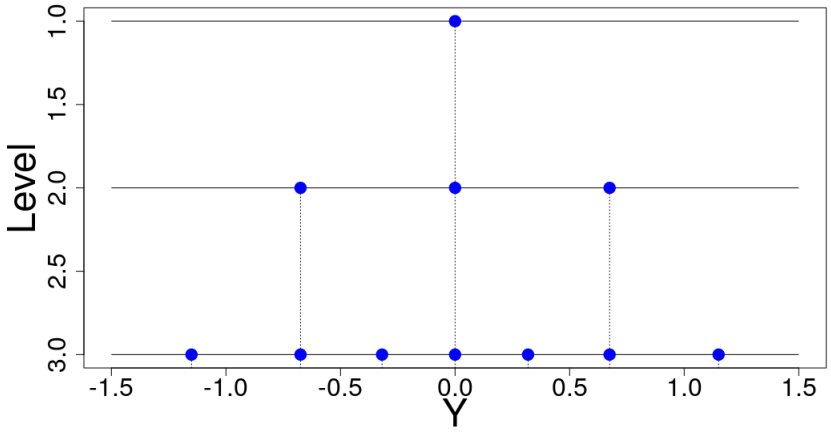


Figure 9 – Partition Example

second level, B_{010} the third set of the third level and so on. Let $B_\emptyset = \Omega$ and take

$$E = \{0,1\}, \quad E^m = \overbrace{E \times \cdots \times E}^{m \text{ times}} = \{0,1\}^m, \quad E^0 = \emptyset, \quad E^* = \bigcup_{m=0}^{\infty} E^m.$$

Now, we can define the Pólya tree process.

Definition 6. *Lavine (1992). Let* Π *be a collection of separable binary trees of partitions of* Ω*. Take* $\mathcal{A} = \{\alpha_{\boldsymbol{\epsilon}} : \boldsymbol{\epsilon} \in E^*\}$ *and* $\Theta = \{\theta_{\boldsymbol{\epsilon}} : \boldsymbol{\epsilon} \in E^*\}$*. Also take* $\boldsymbol{\epsilon}_m = \{\epsilon_{m1}, \cdots, \epsilon_{mm}\}$ *as the record of the whole trajectory up to m, i.e.,* $\epsilon_{mi} = 0$ *if, at the i-th layer, we chose the set on the left and* $\epsilon_{mi} = 1$ *if we chose the set on the right.*

F is said to have a Pólya tree distribution with parameters $(\Pi, \mathcal{A})$*, i.e.,* $F \sim PT(\Pi, \mathcal{A})$*, if the following conditions are satisfied:*

1. *All random variables in* Θ *are mutually independent;*
2. *For every* $m \in \{1,2,\cdots\}$ *and every* $\boldsymbol{\epsilon}_m \in E^m$*,* $\theta_{\boldsymbol{\epsilon}_m} \sim Beta(\alpha_{\boldsymbol{\epsilon}_m 0}, \alpha_{\boldsymbol{\epsilon}_m 1})$*;*
3. *For every* $m \in \{1,2,\cdots\}$ *and every* $\boldsymbol{\epsilon}_m \in E^m$*,*

$$P(B_{\boldsymbol{\epsilon}_m}|\Theta) = \prod_{i=1}^{m} (\theta_{\boldsymbol{\epsilon}_{i-1}})^{\epsilon_{ii}} (1 - \theta_{\boldsymbol{\epsilon}_{i-1}})^{1-\epsilon_{ii}}.$$

In the case of the Pólya tree, it is allowed for $\theta_{\boldsymbol{\epsilon}_m}$ to have a degenerate distribution ($\alpha_{\boldsymbol{\epsilon}_m 0} = 0$ or $\alpha_{\boldsymbol{\epsilon}_m 1} = 0$). In this case, one would choose the set on the left or on the right with probability one.

A.2 Model Properties

One of the most important aspects of the Pólya tree is the fact that its posterior is conjugate. Based on a sample $\boldsymbol{x} = (x_1, \cdots, x_n)$, let $n_{\boldsymbol{\epsilon}_m} = \sum_{j=1}^{n} \mathbb{I}(x_j \in B_{\boldsymbol{\epsilon}_m})$, for all $m \in \{1,2,\cdots\}$, where $\mathbb{I}(.)$

is the indicator function. Then, we can ensure that

$$F|\boldsymbol{x} \sim PT(\Pi, \mathcal{A}^*); \quad \mathcal{A}^* = \{\alpha^*_{\boldsymbol{\epsilon}} : \boldsymbol{\epsilon} \in E^*\}; \quad \alpha^*_{\boldsymbol{\epsilon}_m} = \alpha_{\boldsymbol{\epsilon}_m} + n_{\boldsymbol{\epsilon}_m}. \tag{A.1}$$

This shows one of the greatest advantages of the process, since it remains tractable and is updated in a sequential manner. If any $\alpha_{\boldsymbol{\epsilon}_m}$ is updated by an observation, then, at the next level, it suffices to check if $\alpha_{\boldsymbol{\epsilon}_{m0}}$ or $\alpha_{\boldsymbol{\epsilon}_{m1}}$ should be updated as well. The conjugacy of the model remains valid even when the data is partially observed, that is, if it presents any sort of censoring or is only known up to an interval. In such a case, only finitely many parameters would need to be updated.

Based on the conjugacy statement and given the fact that all parameters on Θ are mutually independent, we can derive the distribution of $\boldsymbol{X} = (X_1, \cdots, X_n)$ either with a fixed set of Θ or integrating Θ out. In this case,

$$\begin{aligned} P(X_1 \in B_{\boldsymbol{\epsilon}_{m_1}}, \cdots, X_n \in B_{\boldsymbol{\epsilon}_{m_n}} | \Theta, \Pi, \mathcal{A}) &= \prod_{j=1}^{n} P(Y_j \in B_{\boldsymbol{\epsilon}_{m_j}} | \Theta, \Pi, \mathcal{A}) \\ &= \prod_{j=1}^{n} \prod_{i=1}^{m_j} \theta_{\boldsymbol{\epsilon}_i}^{n_{\epsilon_i 0}} (1 - \theta_{\boldsymbol{\epsilon}_i})^{n_{\epsilon_i 1}}; \end{aligned} \tag{A.2}$$

$$P(X_1 \in B_{\boldsymbol{\epsilon}_{m_1}}, \cdots, X_n \in B_{\boldsymbol{\epsilon}_{m_n}} | \Pi, \mathcal{A}) = \prod_{j=1}^{n} \prod_{i=1}^{m_j} \left[\frac{\Gamma(\alpha_{\boldsymbol{\epsilon}_i 0} + \alpha_{\boldsymbol{\epsilon}_i 1})}{\Gamma(\alpha_{\boldsymbol{\epsilon}_i 0})\Gamma(\alpha_{\boldsymbol{\epsilon}_i 1})} \frac{\Gamma(\alpha_{\boldsymbol{\epsilon}_i 0} + n_{\boldsymbol{\epsilon}_i 0})\Gamma(\alpha_{\boldsymbol{\epsilon}_i 1} + n_{\boldsymbol{\epsilon}_i 1})}{\Gamma(\alpha_{\boldsymbol{\epsilon}_i 0} + n_{\boldsymbol{\epsilon}_i 0} + \alpha_{\boldsymbol{\epsilon}_i 1} + n_{\boldsymbol{\epsilon}_i 1})} \right].$$

The result in (A.2) highlights the intuition behind the Pólya tree. With a fixed Θ, an observation behaves as a particle that falls throughout the tree. With probability $\theta_{\boldsymbol{\epsilon}_i}$, it goes to the left side and, with probability $1 - \theta_{\boldsymbol{\epsilon}_i}$, it goes to the right side. Then, after repeating this process infinitely many times, the end of the trajectory results in a new observation. Thus, Π and Θ represent a fixed distribution function, providing sufficient information to sample from X. One can expect that, as n grows large, the posterior of F gets arbitrarily close to the true distribution of the data. This is in fact the case, and the proximity argument can be proved in at least two different contexts, as shown in Theorems 4 and 5:

Theorem 4. Lavine (1992). Let $\Omega = \mathbb{R}$ and take G as the true distribution function of the data. Then, for any $\varepsilon > 0$ and any $\eta \in (0, 1)$, there exists a Pólya tree process F such that

$$P\left(\bigcap_{y \in \Omega} |F(y) - G(y)| < \varepsilon \right) > \eta.$$

Theorem 5. Lavine (1994). Let G be a probability measure with probability density function $g(\cdot)$. For any $\varepsilon > 0$ and any $\eta \in (0, 1)$, there exists a Pólya tree process F with density $f(\cdot)$ such that

$$P\left(ess \sup_{y \in \Omega} \left| \log\left(\frac{f(y)}{g(y)} \right) \right| < \varepsilon \right) > \eta.$$

A.3 Centering Distribution and Partition Choice

Another feature of this prior process that differs from many others in the literature is the fact that it allows for discrete, continuous and singular continuous random variables as input. Based on results reproduced in Phadia (2016), the following choice for the parameters on $\mathcal{A}$ can guarantee that, with probability one, the distribution function is of the same type as that of X:

$$\text{For all } m \in \{1,2,\cdots\}, \quad \alpha_{\epsilon_m 0} = \alpha_{\epsilon_m 1} = \begin{cases} \frac{1}{2^m}, & \text{if } X \text{ is discrete;} \\ 1, & \text{if } X \text{ is continuous singular;} \\ m^2, & \text{if } X \text{ is absolutely continuous.} \end{cases} \tag{A.3}$$

With the information of (A.3) at hand, one can fully specify the parameters according to the type of distribution function one seeks. Now, it is time to specify an adequate collection of partitions to accomodate the prior uncertainty about the model.

Unlike other usual nonparametric priors, such as the Dirichlet process, the Pólya tree process is considerably dependent on the chosen set of partitions, since an ill-chosen set can slow down considerably the process of fitting the data. While this can be seen as a disadvantage, it allows one to place greater weights on regions deemed appropriate, reaffirming the value of an adequate prior.

We now turn our attention to the conditional distribution $B_{\epsilon_m}|B_{\epsilon_{m-1}}$. For a fixed Θ, F is known and the conditional distribution is given by

$$F(B_{\epsilon_m}|B_{\epsilon_{m-1}},\Theta) = (\theta_{\epsilon_{m-1}})^{\epsilon_m}(1-\theta_{\epsilon_{m-1}})^{1-\epsilon_m}. \tag{A.4}$$

Thus, we can use (A.4) to obtain the expectation (expressed as $\mathbb{E}(.)$) of the process:

$$\begin{aligned} \mathbb{E}_F[F(B_{\epsilon_m}|B_{\epsilon_{m-1}})] &= \mathbb{E}_\Theta\{\mathbb{E}_F[F(B_{\epsilon_m}|B_{\epsilon_{m-1}},\Theta)]\} = \mathbb{E}_\Theta\left[(\theta_{\epsilon_{m-1}})^{\epsilon_m}(1-\theta_{\epsilon_{m-1}})^{1-\epsilon_m}\right] \\ &= \left(\frac{\alpha_{\epsilon_{m-1}0}}{\alpha_{\epsilon_{m-1}0}+\alpha_{\epsilon_{m-1}1}}\right)^{\epsilon_m}\left(\frac{\alpha_{\epsilon_{m-1}1}}{\alpha_{\epsilon_{m-1}0}+\alpha_{\epsilon_{m-1}1}}\right)^{1-\epsilon_m} = \frac{1}{2}, \end{aligned} \tag{A.5}$$

since $\alpha_{\epsilon_m 0} = \alpha_{\epsilon_m 1}$.

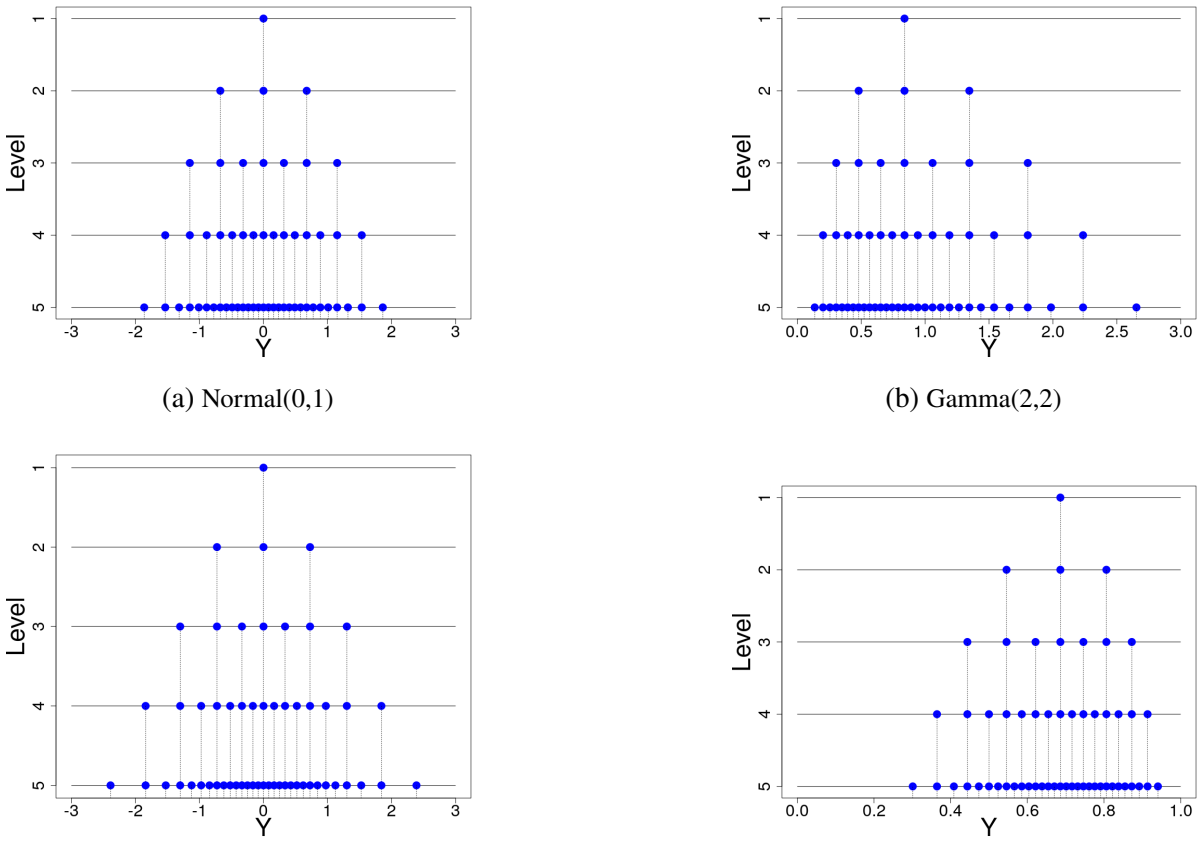
Based on (A.5), one can set a centering distribution F_0 with a known quantile function $q_{F_0}(.)$ and choose the partitions accordingly, such that $F_0(B_0) = F_0(B_1) = \frac{1}{2}$ and, $\forall m = \{1,2,\cdots\}$, $F_0(B_{\epsilon_m 0}|B_{\epsilon_m}) = F_0(B_{\epsilon_m 1}|B_{\epsilon_m}) = \frac{1}{2}$. By setting partitions that guarantee that such a result is true, we can then verify that $\mathbb{E}_F(F) = F_0$.

As an example, take $F_0 \equiv \mathcal{N}(0,1)$, i.e., F_0 is a standard normal distribution. Then, at the first level, $B_0 = (-\infty, 0]$ and $B_1 = (0,\infty)$, since $F_0(B_0) = F_0(B_1) = \frac{1}{2}$. Now, at the second level, B_{00}, B_{01} must be such that

$$\begin{aligned} F_0(B_{00}) = F_0(B_{00}|B_0)F_0(B_0) = \frac{1}{4} &\Rightarrow B_{00} = (-\infty, q_{F_0}(0.25)]; \\ F_0(B_{01}) = F_0(B_{01}|B_0)F_0(B_0) = \frac{1}{4} &\Rightarrow B_{01} = (q_{F_0}(0.25), 0]. \end{aligned}$$

From the same argument, we conclude that $B_{10} = (0, q_{F_0}(0.75)]$ and $B_{11} = (q_{F_0}(0.75), \infty)$. The partitions at the next levels abide by the same logic.

Figure 10 demonstrates the construction of the partitions for different centering distributions. It clearly shows the pivotal role of the choice of the distribution. Even in lower layers, the partition sets are considerably diverse.



(a) Normal(0,1)

(b) Gamma(2,2)

(c) Student's t(5)

(d) Beta(4,2)

Figure 10 – Partitions on the first five levels for different centering distributions

A.4 Partially Specified Pólya Tree

Now that the model is fully specified and its properties have been explored, it is time to look at its limitations. Since the Pólya tree process possesses an infinite number of parameters, it is simply impossible to update them all. (LAVINE, 1994) provides two arguments to defend the updating of the parameters only up to a level M:

1. Pólya trees may be constructed so that the parameters in $\mathcal{A}$ rapidly increase at each level. Then, given a sample of size n and a margin of error, one could set M levels to be updated and ignore the deeper levels, as long as the error based on this choice does not surpass the established margin.

2. It is unreasonable to expect of a professional to be able to fully specify the Pólya tree structure, especially on the tails. However, it could be reasonable to assume that it is possible to specify finitely many parameters responsibly.

Even if one is not in full agreement with these arguments, it is still possible to use a Pólya tree that is specified only up to a level M, which is called a partially specified Pólya tree. Phadia (2016) reproduces that, as a rule of thumb, one could take M to be of order of approximately $\log_2(n)$, where n is the sample size. This alternative allows one to not only obtain a completely specified posterior, but also to sample data from such posterior. If, after level M, one tries to obtain a probability, it can be done by taking $F(B_{\boldsymbol{\epsilon}_{M+1}}|B_{\boldsymbol{\epsilon}_M}) = F_0(B_{\boldsymbol{\epsilon}_{M+1}}|B_{\boldsymbol{\epsilon}_M})$, i.e., as long as one knows the truncated distribution of F_0, obtaining a full probability distribution and sampling are possible. In this case, the prior process is still such that $\mathbb{E}_F(F) = F_0$.

APPENDIX

B

PROOFS

B.1 Quantile test

Theorem 1 (Quantile test)**.** Take $H_0 : G(x_0) = p_0$ as the null hypothesis, F as a distribution function and

$$d(G,F) = \int_{-\infty}^{\infty} |G(x) - F(x)|dx$$

as the dissimilarity function. Then, if $a = \min(F^{-1}(p_0), x_0)$ and $b = \max(F^{-1}(p_0), x_0)$,

$$\delta = \inf_{G \in H_0} d(G,F) = \inf_{G \in H_0} \int_{-\infty}^{\infty} |G(x) - F(x)|dx = \int_a^b |p_0 - F(x)|dx. \tag{3.8}$$

Proof. The proof is done in parts.

- If $F(x_0) = p_0$, then $\inf_{G \in H_0} d(G,F) = \int_a^b |p_0 - F(x)|dx = \int_{x_0}^{x_0} |p_0 - F(x)|dx = 0$.

  Subproof. This case is trivial. If $F(x_0) = p_0$, then $F \in H_0$. If that is the case,

  $$\inf_{G \in H_0} \int_{-\infty}^{\infty} |G(x) - F(x)|dx = \int_{-\infty}^{\infty} |F(x) - F(x)|dx = \int_{-\infty}^{\infty} 0dx = 0.$$

  ∎

- If $F(x_0) < p_0$, then $\inf_{G \in H_0} d(G,F) = \int_a^b |p_0 - F(x)|dx$.

  Subproof. We begin by highlighting the fact that, since $F(x_0) < p_0$, then $a = x_0$. Now, let us define the probability function $F^*(\cdot)$ such that

  $$F^*(x) = \begin{cases} p_0, & \text{if } x \in [x_0, b]; \\ F(x), & \text{otherwise.} \end{cases}$$

Thus, proving that the argument is true is equivalent to showing that

$$\inf_{G\in H_0} d(G,F) = d(F^*,F), \text{ since } d(F^*,F) = \int_a^b |p_0 - F(x)|dx.$$

The proof is done by contradiction. Suppose there exists a probability function $F' \in \mathbb{F}$ such that $F'(x_0) = p_0$ and $d(F^*,F) > d(F',F)$. Then, it necessarily follows that

$$\int_a^b |p_0 - F(x)|dx > \int_a^b |F'(x) - F(x)|dx.$$

After all,

$$\int_{-\infty}^{\infty} |F'(x) - F(x)|dx \geq \int_a^b |F'(x) - F(x)|dx,$$

so it should be guaranteed that $F'(\cdot)$ provides a smaller dissimilarity at $[a,b]$. Then,

$$\int_a^b |F'(x) - F(x)|dx = \int_a^b |[F'(x) - p_0] - [F(x) - p_0]|dx.$$

For $x \in [a,b]$, $F'(x) \geq p_0$ and $F(x) \leq p_0$. Then, $[F'(x) - p_0] \geq 0$ and $-[F(x) - p_0] \geq 0$. Thus, since $[F'(x) - p_0] - [F(x) - p_0] = |F'(x) - p_0| + |F(x) - p_0|$,

$$\begin{aligned}\int_a^b |F'(x) - F(x)|dx &= \int_a^b |F'(x) - p_0| + |F(x) - p_0|dx \\ &= \int_a^b |F'(x) - p_0|dx + \int_a^b |F(x) - p_0|dx \geq \int_a^b |p_0 - F(x)|dx,\end{aligned}$$

which is a contradiction. Thus, $\inf_{G\in H_0} d(G,F) = \int_a^b |p_0 - F(x)|dx$. ■

- If $F(x_0) > p_0$, then $\inf_{G\in H_0} d(G,F) = \int_a^b |p_0 - F(x)|dx$.

Subproof. In this case, we highlight that, since $F(x_0) > p_0$, then $b = x_0$. We now define a sequence $(F_n^*)_{n\geq 1}$ of probability functions such that

$$F_n^*(x) = \begin{cases} p_0, & \text{if } x \in [a, x_0 + \frac{1}{n}); \\ F(x), & \text{otherwise.} \end{cases}$$

By construction, $F_n^* \in H_0, \forall n \geq 1$ and

$$d(F_n^*,F) = \int_a^{x_0+\frac{1}{n}} |p_0 - F(x)|dx = \int_a^{x_0} |p_0 - F(x)|dx + \int_{x_0}^{x_0+\frac{1}{n}} |p_0 - F(x)|dx,$$

which converges decreasingly to $\int_a^{x_0} |p_0 - F(x)|dx$ as $n \to \infty$.

Once again, the proof follows by contradiction. Suppose there exists $F' \in \mathbb{F}$ in H_0 such that $\inf_{G\in H_0} d(G,F) = d(F',F) \neq \int_a^b |p_0 - F(x)|dx$. By a similar argument than the one in the previous subproof,

$$\begin{aligned}\int_a^b |F'(x) - F(x)|dx &= \int_a^b |[F(x) - p_0] - [F'(x) - p_0]|dx \\ &= \int_a^b |F(x) - p_0|dx + \int_a^b |F'(x) - p_0|dx \geq \int_a^b |p_0 - F(x)|dx,\end{aligned}$$

which leads us to conclude that $d(F', F) > \int_a^b |p_0 - F(x)|dx$. Then, there exists $\eta > 0$ such that $d(F', F) = \int_a^b |p_0 - F(x)|dx + \eta$.

However, since $d(F_n^*, F) \underset{n\to\infty}{\longrightarrow} \int_a^b |p_0 - F(x)|dx$, then for the same $\eta > 0$ there exists $n_0 \in \mathbb{N}$ such that, for $n \geq n_0$,

$$\left| d(F_n^*, F) - \int_a^b |p_0 - F(x)|dx \right| = d(F_n^*, F) - \int_a^b |p_0 - F(x)|dx < \eta$$

$$\Rightarrow d(F_n^*, F) < \int_a^b |p_0 - F(x)|dx + \eta,$$

which contradicts the fact that $\inf_{G\in H_0} d(G, F) = d(F', F)$. Then, we conclude that $\inf_{G\in H_0} d(G, F) = \int_a^b |p_0 - F(x)|dx$. ■

Based on each of the subproofs presented, we can safely conclude that

$$\inf_{G\in H_0} d(G, F) = \int_a^b |p_0 - F(x)|dx.$$

□

B.2 Comparison of the distribution of two samples

Theorem 2 (Two sample test)**.** Take $H_0 : G_X = G_Y$ as the null hypothesis, F_X and F_Y as two distribution functions and

$$d[(G_X, G_Y), (F_X, F_Y)] = d^*(G_X, F_X) + d^*(G_Y, F_Y),$$

where $d^*(\cdot, \cdot)$ is a distance function. Then

$$\delta = \inf_{(G_X,G_Y)\in H_0} d[(G_X, G_Y), (F_X, F_Y)] = d^*(F_X, F_Y).$$

Proof. We begin by noticing that, since H_0 presents a contextual restriction related to two random variables, it follows that $\mathbb{H} = \mathbb{F} \times \mathbb{F}$. After all, if X and Y were defined on different sample spaces, it would make no sense to compare if their respective distribution functions are the same. Furthermore, $\mathbb{H}$ has this specific structure because the alternative hypothesis is that G_X and G_Y can take any element of $\mathbb{F}$ as long as $G_X \neq G_Y$. Thus, formally,

$$\begin{cases} H_0 : (G_X, G_Y) \in \mathbb{F} \times \mathbb{F} : G_X(z) = G_Y(z), \forall z \in \mathcal{X}; \\ H_1 : (G_X, G_Y) \in \mathbb{F} \times \mathbb{F} : \exists z \in \mathcal{X} \text{ such that } G_X(z) \neq G_Y(z). \end{cases}$$

Since H_0 assumes that $G_X = G_Y$, we can drop the subscript and conclude that the pragmatic hypothesis is given by

$$Pg(H_0) = \left\{ (F_X, F_Y) \in \mathbb{F} \times \mathbb{F} : \inf_{(G_X,G_Y)\in H_0} d[(G_X, G_Y), (F_X, F_Y)] < \varepsilon \right\} \tag{B.1}$$

$$= \left\{ (F_X, F_Y) \in \mathbb{F} \times \mathbb{F} : \inf_{G\in\mathbb{F}} d[(G, G), (F_X, F_Y)], < \varepsilon \right\}. \tag{B.2}$$

From (B.2), we now need to establish how exactly these pairs of functions will be compared. Given the statement of the proposition,

$$\inf_{G\in\mathbb{F}} d[(G,G),(F_X,F_Y)] = \inf_{G\in\mathbb{F}} [d^*(G,F_X)+d^*(G,F_Y)]. \tag{B.3}$$

Now, since d^* from (B.3) is a distance function, it follows from the properties of symmetry and triangle inequality (KREYSZIG, 1978) that

$$\inf_{G\in\mathbb{F}} [d^*(G,F_X)+d^*(G,F_Y)] = \inf_{G\in\mathbb{F}} [d^*(F_X,G)+d^*(G,F_Y)] \geq d^*(F_X,F_Y). \tag{B.4}$$

Thus, if there exists $G \in \mathbb{F}$ such that $d^*(F_X,G)+d^*(G,F_Y) = d^*(F_X,F_Y)$, it will necessarily be the infimum. However, since $F_X \in \mathbb{F}$, if $G = F_X$ the equality is guaranteed. □

B.3 Monotonicity property of the test

Theorem 3 (Monotonicity property)**.** Take $Pg(H_0)$ as the nonparametric pragmatic hypothesis of H_0 and $\mathcal{H}(\boldsymbol{x})$ as a method that randomly draws elements of $\mathbb{H}$ based on a sample $\boldsymbol{x}$. Then, for a precise hypothesis $H_0 \subset \mathbb{H}$ and $\alpha_1, \alpha_2 \in (0,1)$ where $\alpha_2 \geq \alpha_1$, if a test is such that

- If $\varepsilon < q_{\alpha_1}\left[\inf_{h\in H_0} d(h,\mathcal{H}(\boldsymbol{x}))\right]$, we reject the hypothesis;
- If $\varepsilon \geq q_{\alpha_2}\left[\inf_{h\in H_0} d(h,\mathcal{H}(\boldsymbol{x}))\right]$, we accept the hypothesis;
- Otherwise, we remain agnostic;

it obeys the property of monotonicity, i.e., it is logically coherent for any two precise hypotheses $H_0^1, H_0^2 \subset \mathbb{H}$ if $H_0^1 \supseteq H_0^2$, as shown in Table 3.

Table 8 – Possibility of each combination of decisions

Decisions	Reject H_0^1	Undecided on H_0^1	Accept H_0^1
Reject H_0^2	Possible	Possible	Possible
Undecided on H_0^2	Impossible	Possible	Possible
Accept H_0^2	Impossible	Impossible	Possible

Proof. To observe that the monotonicity property is obeyed, take any two precise hypotheses $H_0^1, H_0^2 \subset \mathbb{H}$ such that $H_0^1 \supseteq H_0^2$. Then, it follows that

$$\inf_{h\in H_0^1} d(h,h^*) = \min\left[\inf_{h\in H_0^2} d(h,h^*), \inf_{h\in(H_0^1-H_0^2)} d(h,h^*)\right] \leq \inf_{h\in H_0^2} d(h,h^*), \forall h^* \in \mathbb{H}$$

Since $\mathcal{H}(\boldsymbol{x})$ provides elements in $\mathbb{H}$, then $\inf_{h\in H_0^1} d(h,\mathcal{H}(\boldsymbol{x})) \leq \inf_{h\in H_0^2} d(h,\mathcal{H}(\boldsymbol{x}))$ as well.

Now, take $Y = \mathbb{I}\left(\inf_{h\in H_0^1} d(h,\mathcal{H}(\boldsymbol{x})) \leq q\right)$ and $Z = \mathbb{I}\left(\inf_{h\in H_0^2} d(h,\mathcal{H}(\boldsymbol{x})) \leq q\right)$, where $\mathbb{I}(\cdot)$ is the indicator function and $q \in \mathbb{R}^+$. Therefore, $Y \geq Z$, which implies that $\mathbb{E}[Y] \geq \mathbb{E}[Z]$. However, due to the definition of Y and Z,

$$\mathbb{E}[Y] \geq \mathbb{E}[Z] \Rightarrow \mathbb{P}\left(\inf_{h\in H_0^1} d(h,\mathcal{H}(\boldsymbol{x})) \leq q\right) \geq \mathbb{P}\left(\inf_{h\in H_0^2} d(h,\mathcal{H}(\boldsymbol{x})) \leq q\right), \quad \forall q \in \mathbb{R}^+.$$

Hence, by taking $q = q_\alpha\left[\inf_{h\in H_0^1} d(h,\mathcal{H}(\boldsymbol{x}))\right]$, i.e., q is the α quantile of $\inf_{h\in H_0^1} d(h,\mathcal{H}(\boldsymbol{x}))$, then

$$\alpha \geq \mathbb{P}\left(\inf_{h\in H_0^2} d(h,\mathcal{H}(\boldsymbol{x})) \leq q_\alpha\left[\inf_{h\in H_0^1} d(h,\mathcal{H}(\boldsymbol{x}))\right]\right), \quad \forall \alpha \in (0,1),$$

leading to the conclusion that $q_\alpha\left[\inf_{h\in H_0^1} d(h,\mathcal{H}(\boldsymbol{x}))\right] \leq q_\alpha\left[\inf_{h\in H_0^2} d(h,\mathcal{H}(\boldsymbol{x}))\right], \forall \alpha \in (0,1)$.

Based on the result above, it is time to evaluate all possible decisions one can take towards H_0^1 and what limitation each of them entails on the evaluation of H_0^2. By setting $\alpha_1, \alpha_2 \in (0,1)$, it can be concluded that

- Rejecting H_0^1 implies that $\varepsilon < q_{\alpha_1}\left[\inf_{h\in H_0^1} d(h,\mathcal{H}(\boldsymbol{x}))\right] \leq q_{\alpha_1}\left[\inf_{h\in H_0^2} d(h,\mathcal{H}(\boldsymbol{x}))\right]$, thus one must reject H_0^2 as well;
- Accepting H_0^1 implies that $\varepsilon \geq q_{\alpha_2}\left[\inf_{h\in H_0^1} d(h,\mathcal{H}(\boldsymbol{x}))\right]$. This situation tells nothing about the relationship between ε, $q_{\alpha_1}\left[\inf_{h\in H_0^2} d(h,\mathcal{H}(\boldsymbol{x}))\right]$ and $q_{\alpha_2}\left[\inf_{h\in H_0^2} d(h,\mathcal{H}(\boldsymbol{x}))\right]$, so any conclusion about H_0^2 is theoretically feasible;
- Remaining agnostic about H_0^1 implies that

  $$q_{\alpha_2}\left[\inf_{h\in H_0^2} d(h,\mathcal{H}(\boldsymbol{x}))\right] \geq q_{\alpha_2}\left[\inf_{h\in H_0^1} d(h,\mathcal{H}(\boldsymbol{x}))\right] > \varepsilon \geq q_{\alpha_1}\left[\inf_{h\in H_0^1} d(h,\mathcal{H}(\boldsymbol{x}))\right],$$

  which leads to the non-acceptance of H_0^2, but is still not informative enough to tell if one should reject H_0^2 or remain agnostic about it.

Thus, the results derived are exactly the same as those presented in Table 3, confirming the monotonicity property of the test. □


ICMC
USP
SÃO CARLOS