

Hellenic Complex Systems Laboratory

# A Bayesian Inference Based Computational Tool for Parametric and Nonparametric Medical Diagnosis

Technical Report XXV

Theodora Chatzimichail and Aristides T. Hatjimihail  
2023

# A Bayesian Inference Based Computational Tool for Parametric and Nonparametric Medical Diagnosis

Theodora Chatzimichail, MRCS <sup>a</sup>, Aristides T. Hatjimihail, MD, PhD <sup>a</sup>

<sup>a</sup> Hellenic Complex Systems Laboratory

**Search Terms:** Bayesian Diagnosis; Bayesian Inference; Prior Probability; Posterior Probability; Likelihood; Parametric Distribution; Nonparametric Distribution; Copula Distribution; Kernel Density Estimator; Probability Density Function; Diabetes mellitus

## Abstract

Medical diagnosis is the basis for treatment and management decisions in healthcare. Conventional methods for medical diagnosis commonly employ established clinical criteria and fixed numerical thresholds. Such a rigid approach may fail to capture the intricate interactions between diagnostic tests and the varying prevalence of diseases. To explore this further, we have developed a freely available specialized computational tool that employs Bayesian inference to calculate the posterior probability of disease diagnosis. This novel software comprises of three distinct modules, each designed to allow users to define and compare parametric and nonparametric distributions effectively. The tool is equipped to analyze datasets generated from two separate diagnostic tests, each performed on both diseased and nondiseased populations. We demonstrate the utility of this software by analyzing fasting plasma glucose and glycated hemoglobin A1c data from the National Health and Nutrition Examination Survey. Our results are validated using the oral glucose tolerance test as a reference standard, and we explore both parametric and nonparametric distribution models for the Bayesian diagnosis of diabetes mellitus.

## Keywords

Bayesian Diagnosis; Bayesian Inference; Prior Probability; Posterior Probability; Likelihood; Parametric Distribution; Nonparametric Distribution; Copula Distribution; Kernel Density Estimator; Probability Density Function; Diabetes mellitus

## Introduction

Medical diagnosis is a critical process of accurately identifying pathological conditions in patients. The term "diagnosis" has its etymological origins in the ancient Greek word "διάγνωσις," signifying 'discernment' (Weiner, Simpson, and Oxford University Press 1989). Traditionally, diagnostic tests are used to divide individuals into two principal categories: those who are afflicted with a specific disease and those who are not. Notably, the probability distributions associated with quantitative diagnostic test outcomes often demonstrate some overlap between the diseased and nondiseased groups. To address this, numerical diagnostic thresholds or cut-off points have been formulated to provide a binary classification of these test outcomes (Zweig and Campbell 1993). Nevertheless, this introduces a certain measure of uncertainty into the diagnostic accuracy (Chatzimichail and Hatjimihail 2021). This dichotomous method represents a significant shift in medical decision-making by linking a continuum of evidence, which spans across a spectrum of reliability, to binary clinical decisions such as to treat or not to treat (Djulbegovic et al. 2015).

Despite the evident efficiency of traditional diagnostic methods, they sometimes fail to capture the complexity and heterogeneity of disease presentations across diverse patient populations (Choi, Johnson, and Thurmond 2006). To address these limitations, our research focuses on implementing Bayesian inference to calculate the posterior probabilities associated with disease diagnosis (Viana and Ramakrishnan 1992; Gelman et al. 2013; van de Schoot et al. 2021). Within this Bayesian paradigm, prior probabilities of disease are integrated with distributions of diagnostic measurands in both diseased and nondiseased populations. This approach enables the evaluation of the information conveyed by diagnostic measurements and combination of data from multiple diagnostic tests, which may improve diagnostic accuracy and precision while introducing flexibility, adaptability, and versatility into the diagnostic process (Carlin and Louis 2008).

A considerable challenge in integrating Bayesian inference into medical diagnosis is the limited availability of literature detailing the statistical distributions of diagnostic variables in both pathological and non-pathological states (Dawid 1984).

The ubiquitous application of the normal distribution in clinical laboratory indicators is due, in part, to its mathematical simplicity, the foundational Central Limit Theorem, and a rich collection of statistical methods designed for Gaussian data (Lehmann and Romano 2008). However, the universal applicability of the normal distribution is subject to critique, especially when dealing with clinical measurands that exhibit skewness, bimodality, or multimodality (G. E. P. Box and Cox 1964). Hence, while the normal distribution remains invaluable in statistical modeling, critical evaluation of its appropriateness for specific diagnostic measurands is necessary. This evaluation should be accompanied by a willingness to employ alternative statistical distributions when circumstances demand (D'Agostino and Pearson 1973).

This foundational data is crucial for Bayesian inference, establishing the essential context against which new diagnostic measurements can be compared. The absence of such normative data could potentially compromise the reliability and validity of Bayesian diagnostic methods.

To address the complex issues related to Bayesian diagnosis and the selection of appropriate statistical distributions for diagnostic variables, we have developed an interactive software tool programmed in the Wolfram Language. This tool features three modules that allow users to explore and compare both parametric and nonparametric distributions to calculate posterior probabilities for disease. It is designed to analyze datasets of measurements of two measurands of two distinct diagnostic tests, conducted on both diseased and nondiseased populations.

## Methods

### Computational Methods

#### Bayesian Diagnostic Approach

The Bayesian diagnostic approach is a cornerstone in statistical inference and particularly useful in medical diagnosis (Viana and Ramakrishnan 1992; Velanovich 1994; Wilkes 2022). The approach relies on Bayes' theorem (Gelman et al. 2013). For effective implementation of the Bayesian diagnostic method, knowledge concerning the statistical distributions of the measurements of the diagnostic tests is essential (Lehmann and Romano 2008).

Bayes Theorem is presented in Appendix I.

#### Parametric Distributions

Parametric statistics assume that sample data comes from a population that can be adequately modeled by a probability distribution that has a fixed set of parameters (Geisser and Johnson 2006). The parametric distributions provided by the program are the following:

1. *Normal Distribution*
  - 1.1. Univariate
  - 1.2. Bivariate
2. *Lognormal Distribution*
  - 2.1. Univariate
  - 2.2. Bivariate
3. *Gamma Distribution*
  - 3.1. Univariate
  - 3.2. Bivariate
4. *Copula Distributions*

The copula distributions of the program are bivariate, with a bivariate normal distribution with correlation  $\rho$  as kernel, and univariate normal, lognormal and gamma marginals.

The Probability Density Functions (PDF) of the parametric distributions are mathematically defined in Appendix I.

#### Nonparametric Distributions

Conversely, nonparametric models were also employed, which do not make a priori assumptions about the distribution's mathematical form (Spiegelhalter, Abrams, and Myles 2004). These are particularly useful for exploratory data analysis and are implemented as shown in Appendix I.

#### Histograms

A histogram is the graphical representation of the distribution of a dataset as a series of bins.

The program plots histograms of the provided datasets.

### Kernel Density Estimators

In contrast to histograms, a kernel density estimator (KDE) generates a continuous and smooth estimate of the underlying PDF by summing the contributions of kernel functions centered at each data point.

The KDE offers a flexible nonparametric approach to density estimation, allowing for a better representation of the data's underlying distribution.

The program provides univariate and bivariate Gaussian KDE. The bivariate KDE use radial-type kernels.

## The Program

To facilitate the calculation of Bayesian probabilities of disease diagnosis, *Bayesian Diagnosis* was developed in Wolfram Language, using Wolfram Mathematica® Ver. 13.3<sup>1</sup>. This interactive program consists of three primary modules with eighteen submodules. It allows the calculation, plotting and comparison of Bayesian posterior probability of disease for two diagnostic tests, assuming two sets of alternative parametric and nonparametric distributions of the measurements of those tests in diseased and nondiseased populations.

It is freely available as a Wolfram Mathematica Notebook (.nb) (Supplementary File: BayesianDiagnosis.nb). It can be run on Wolfram Player® or Wolfram Mathematica® (see Appendix II).

### Interface of the Program

The program is designed with an intuitive user interface, constructed to allow users to input and modify various prior probability and measurement parameters and to select parametric distributions and Kernel Density Estimators (KDE) related to medical diagnosis (see Appendix III and Supplementary File: BayesianDiagnosisInterface.pdf).

### Input Parameters

#### Prior Probability

The user initiates the diagnostic evaluation by specifying the prior probability of disease occurrence in the population under study. This serves as a foundational metric for subsequent analyses.

#### Parametric Distributions

To facilitate a diagnostic model, the program allows for the definition of various parametric distributions for both the diseased and nondiseased populations across two diagnostic tests.

1. *Distribution Selection*: The user selects the type of distribution from a predefined list:
  - 1.1. Normal Distribution
  - 1.2. Lognormal Distribution
  - 1.3. Gamma Distribution
2. *Statistical Parameters*: For each chosen distribution, the user defines the mean  $\mu$  and standard deviation  $\sigma$  of the respective population.
3. *Correlation Coefficients*: Users specify the correlation coefficients  $\rho$  between the measurements of the first and second diagnostic tests for both diseased and nondiseased populations.

---

<sup>1</sup> Wolfram Research, Inc., Mathematica, Version 13.3, Champaign, IL (2023).

## Kernel Density Estimators

Alternatively, the user can opt to define the KDE for the measurements in both diseased and nondiseased populations across the two tests:

1. *Bandwidth Parameter*: For each KDE, the user defines the bandwidth parameter  $h$ .
2. *Correlation Coefficients*: As with parametric distributions, the user defines correlation coefficients  $\rho$  between the measurements of the two diagnostic tests.

## Output Specifications

### Visualizations

The program generates a series of plots designed to elucidate various diagnostic metrics and statistics:

1. *Posterior Probability of Disease*: Plots are generated to show the posterior probability of disease for each test and their combination.
2. *PDF*: Univariate PDF for each test and the bivariate PDF of their combination are plotted. An option to overlay histograms on these plots is also provided.
3. *Quantile-Quantile (Q-Q) Plots*: These plots are produced for each measurand to examine the distributional characteristics (Wilk and Gnanadesikan 1968).
4. *Probability-Probability (P-P) Plots*: Similar to Q-Q plots, P-P plots are generated for further assessment of the distribution of each measurand (Wilk and Gnanadesikan 1968).

### Tables

1. *Population Statistics*: The program tabulates key statistical metrics such as mean, median, standard deviation, skewness, and kurtosis for each user-defined distribution and dataset. For each bivariate distribution of the first and second test in diseased and nondiseased populations, the correlation coefficients are calculated and displayed.
2. *Posterior Disease Probabilities*: For a user-defined pair of test measurement values, the program computes and presents the posterior probabilities for disease for each test and their combination.

By providing this comprehensive set of input parameters and output specifications, the program offers a robust platform for exploring the Bayesian diagnosis of disease using either parametric distributions or KDE of medical diagnostic measurements.

## Datasets

Although the program includes four datasets of measurements, one for each diagnostic test, applied to a diseased and a nondiseased population, these can be replaced by other appropriate datasets selected by the user (see Appendix II). Therefore, it can be used for any combination of diagnostic tests and diseases.

## Illustrative Application

To demonstrate the application of the program, fasting plasma glucose (FPG) was used as the first measurand and glycated hemoglobin A1c (HbA1c) as the second measurand for Bayesian diagnosis of diabetes mellitus. The oral glucose tolerance test (OGTT) was used as the reference diagnostic method. A diagnosis of diabetes was confirmed if the plasma glucose value was equal to or exceeded 200 mg/dl, measured two hours after oral administration of 75 g of glucose, during an OGTT (ElSayed et al. 2023).

National Health and Nutrition Examination Survey (NHANES) data from participants was retrieved for the period from 2005 to 2016 (National Center for Health Statistics 2005-20016) ( $n = 60,936$ ).

NHANES is a series of studies designed to evaluate the health and nutritional status of adults and children in the United States.

The inclusion criteria for participants were:

1. Age  $\geq 18$  years (n = 36,287)
2. Valid fasting plasma glucose (FPG), glycated hemoglobin (HbA1c), and oral glucose tolerance test (OGTT) results (n=11,563)
3. A negative response to NHANES question DIQ010 regarding a diabetes diagnosis (National Center for Health Statistics 2005-20016) (n=11,210)
4. Non-pregnancy status (n=11,206)

Participants with a 2-hr OGTT measurement of  $\geq 200$  mg/dl were considered diabetic (n = 687).

Descriptive statistics, including the mean, median, and standard deviation, were computed for each dataset. Univariate distributions were employed to model the distributions of FPG and HbA1c and bivariate distributions to model the joint distribution of FPG and HbA1c. Likelihoods and posterior probabilities were estimated for FPG, HbA1c and their combinations.

The prior probability of diabetes was estimated as

$$v = \frac{687}{11206} = 0.061.$$

The statistics of the dataset are presented in Table 1.

	Diabetics		Nondiabetics	
n	687		10519	
Measurand (Units)	FPG (mg/dl)	HbA1c (%)	FPG (mg/dl)	HbA1c (%)
Mean	133.6	6.19	98.7	5.36
Median	121.0	6.10	98.5	5.40
Standard Deviation	45.1	1.18	10.5	0.37
Skewness	2.804	2.674	0.649	0.025
Kurtosis	11.923	11.566	5.053	3.497
Correlation Coefficient	0.885		0.381	

**Table 1:** The descriptive statistics of FPG and HbA1c.

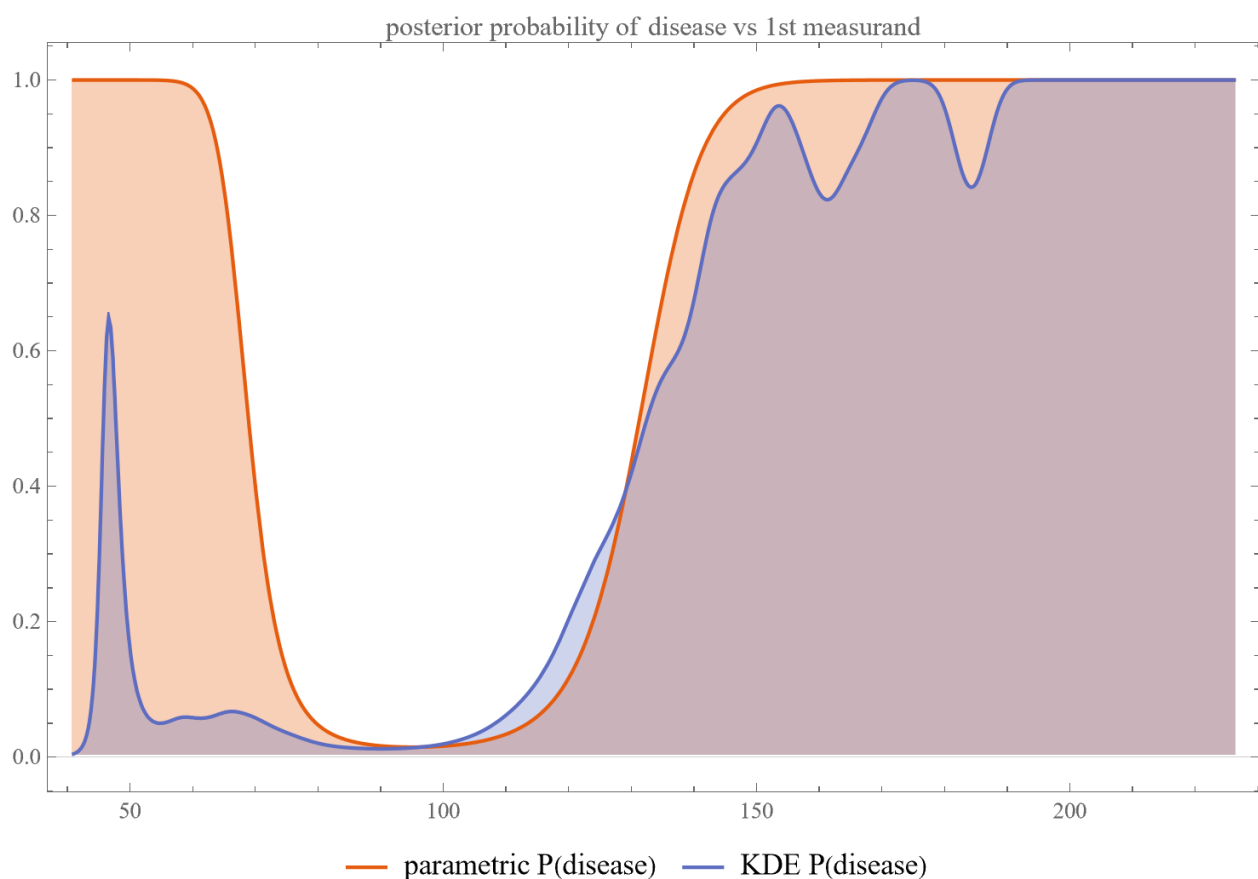
## Results

Using the settings of Table 2, the program generated the plots of Figures 1-11 and the tables of Figures 12-13.

Measurand (Units)	Diabetics		Nondiabetics	
	FPG (mg/dl)	HbA1c (%)	FPG (mg/dl)	HbA1c (%)
Parametric Distribution	Lognormal	Lognormal	Lognormal	Lognormal
Parametric Distribution Mean	133.6	6.19	98.7	5.36
Parametric Distribution SD	45.1	1.18	10.5	0.37
KDE Smoothing Bandwidth (SD units)	0.25	0.24	0.24	0.27
Correlation Coefficient	0.885		0.381	

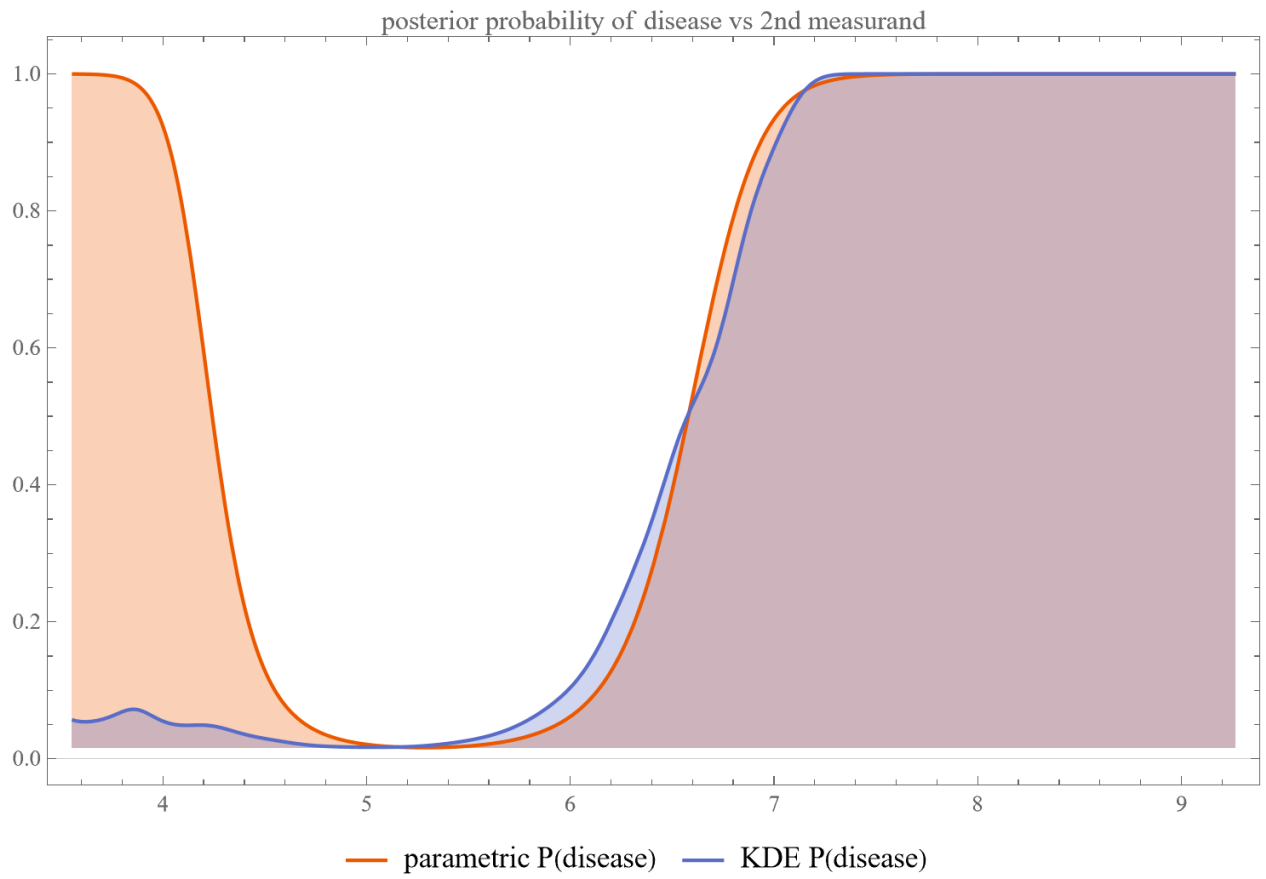
**Table 2:** The settings of the program for Figures 1-123.

The KDE smoothing bandwidth was set to double that given by Silverman's rule of thumb (Menke et al. 2014; Silverman 1986).



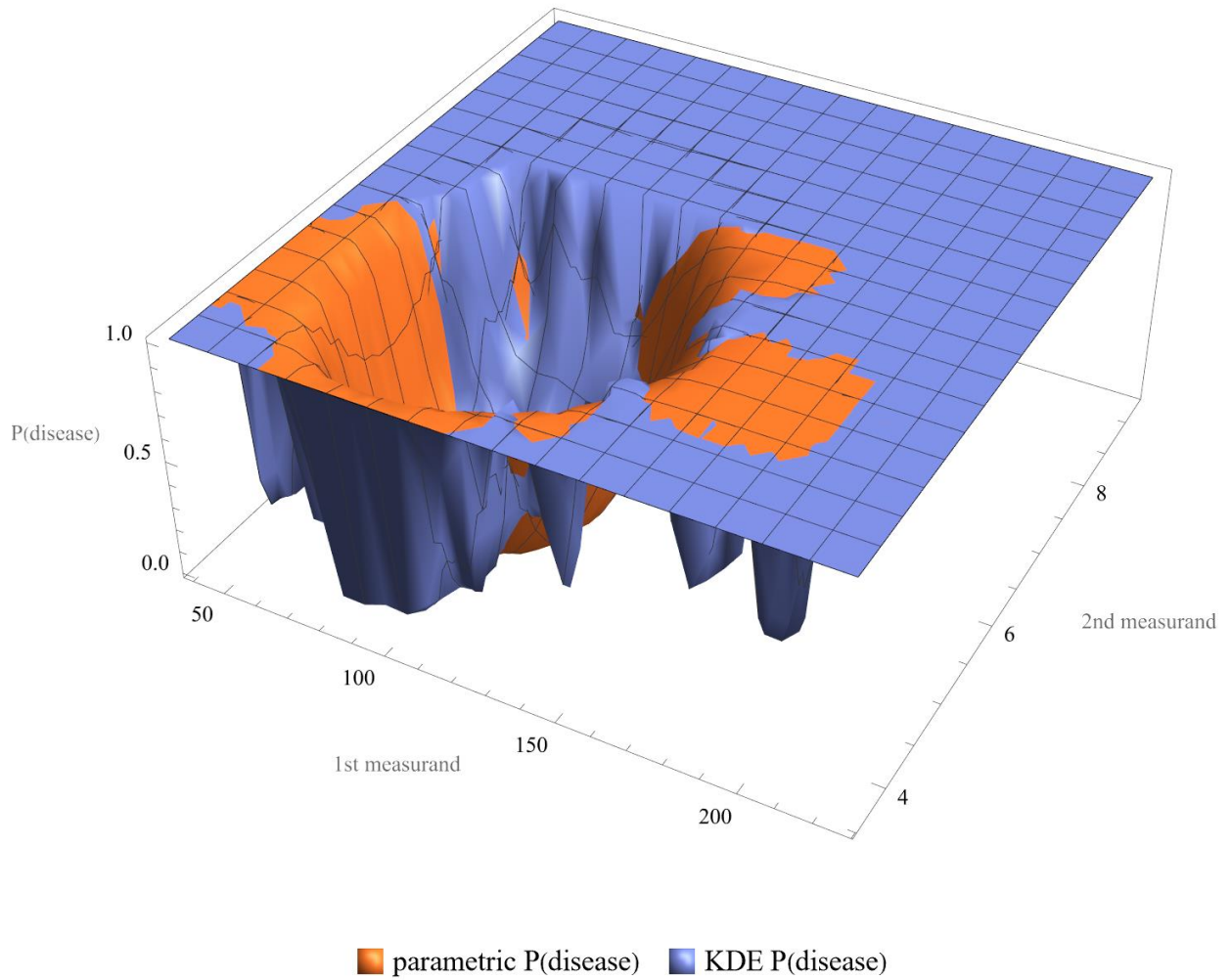
**Figure 1:** Posterior probability of disease (diabetes) versus the first measurand (FPG), assuming parametric and KDE distributions of the measurand, with the settings of the program in Table 2.



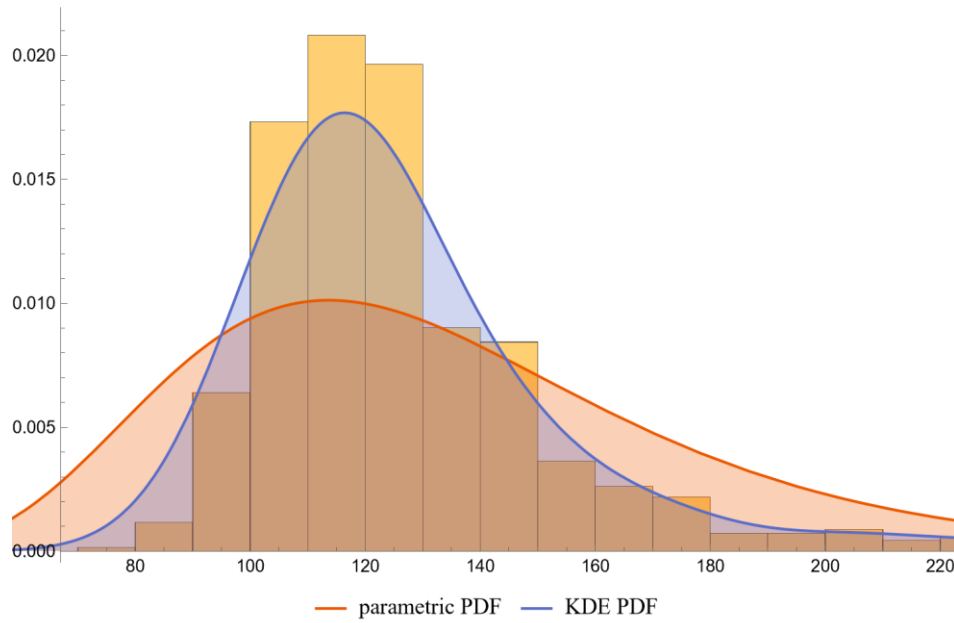


**Figure 2:** Posterior probability of disease (diabetes) versus the second measurand (HbA1c), assuming parametric and KDE distributions of the measurand, with the settings of the program in Table 2.

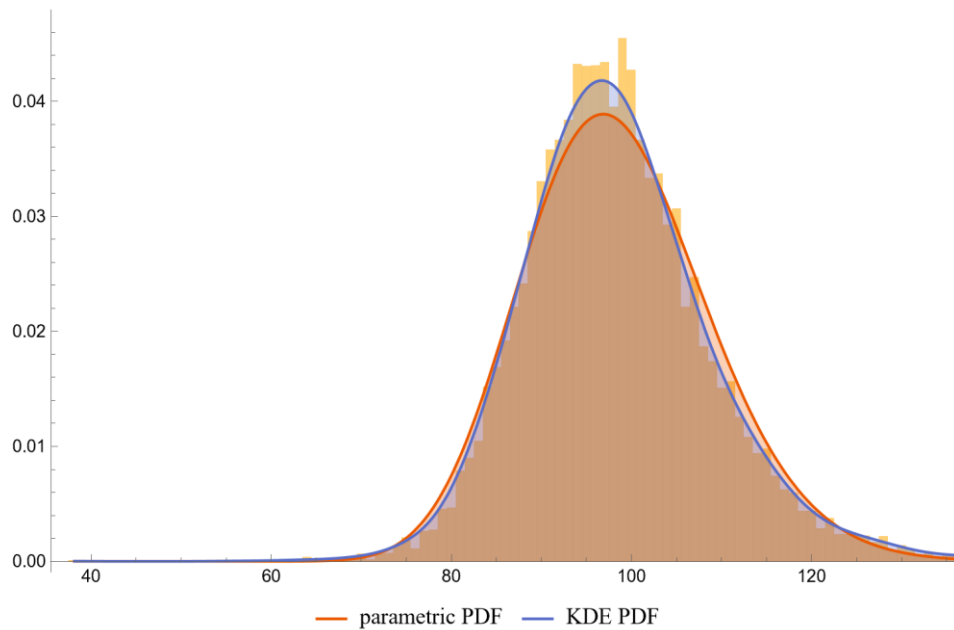
posterior probability of disease vs 1st and 2nd measurands



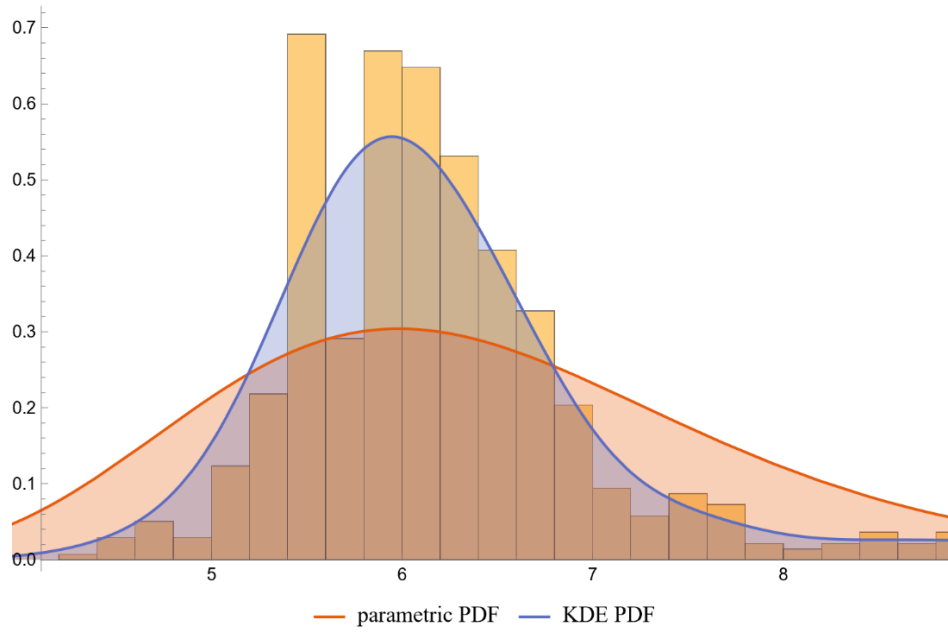
**Figure 3:** Posterior probability of disease (diabetes) versus both measurands (FPG and HbA1c), assuming parametric and KDE distributions of the measurands, with the settings of the program in Table 2.



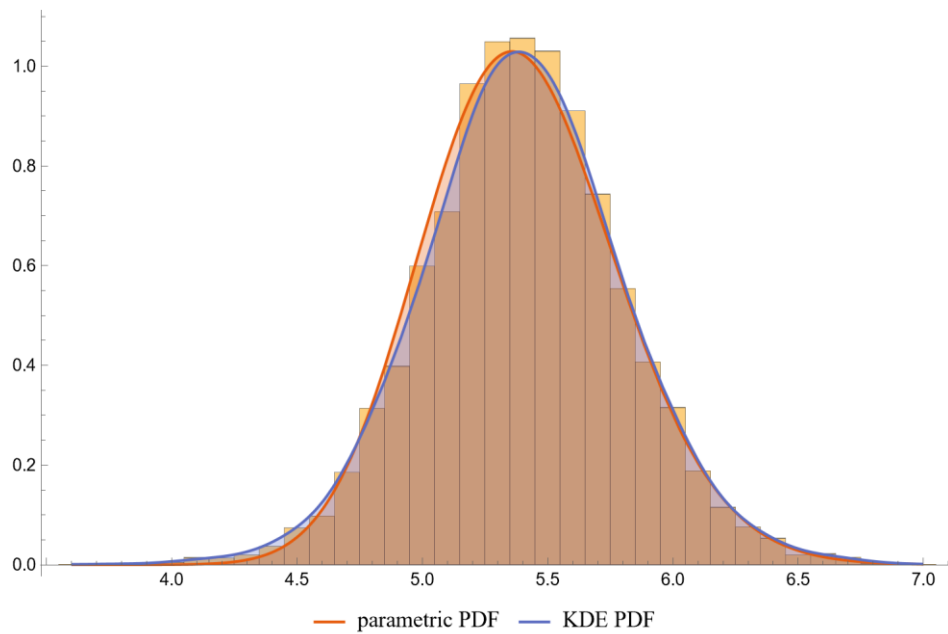
**Figure 4:** The PDF of the first measurand (FPG) in diseased (diabetics), assuming parametric and KDE distributions of the measurand, and the histogram of the respective sample (NHANES dataset, with the settings of the program in Table 2.



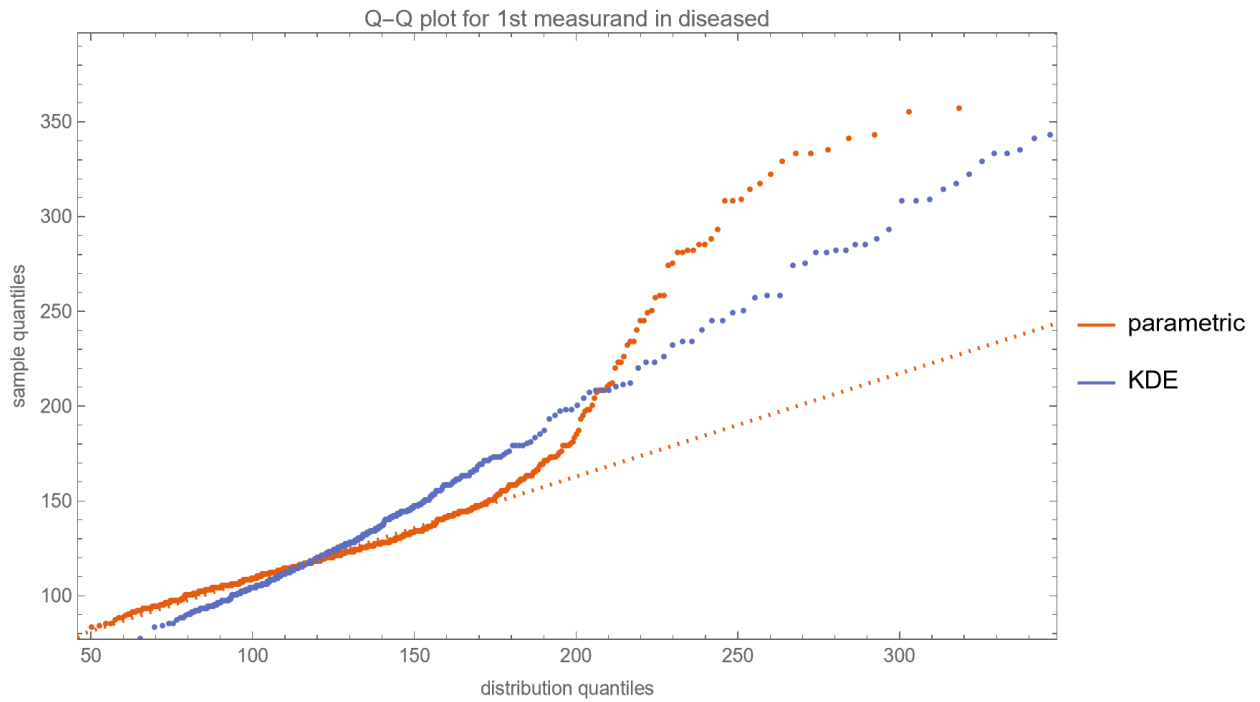
**Figure 5:** The PDF of the first measurand (FPG) in nondiseased (nondiabetics), assuming parametric and KDE distributions of the measurand, and the histogram of the respective sample (NHANES dataset), with the settings of the program in Table 2.



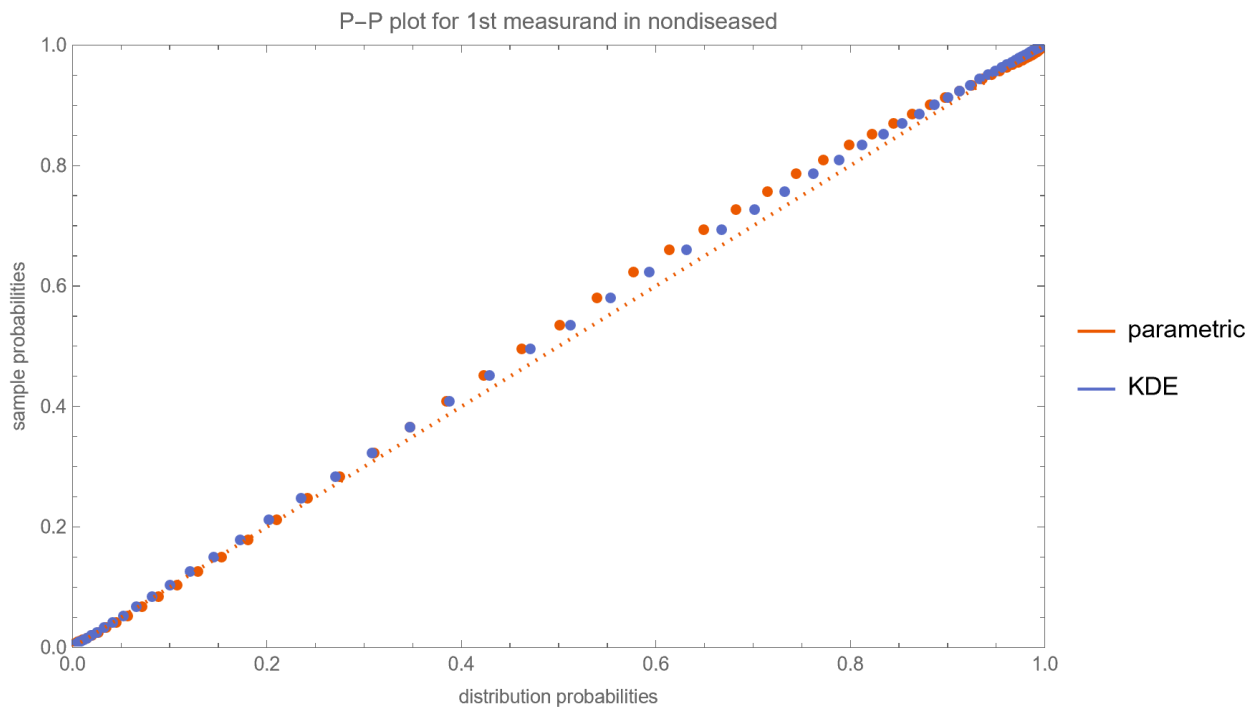
**Figure 6:** The PDF of the second measurand (HbA1c) in diseased (diabetics), assuming parametric and KDE distributions of the measurand, and the histogram of the respective sample (NHANES dataset), with the settings of the program in Table 2.



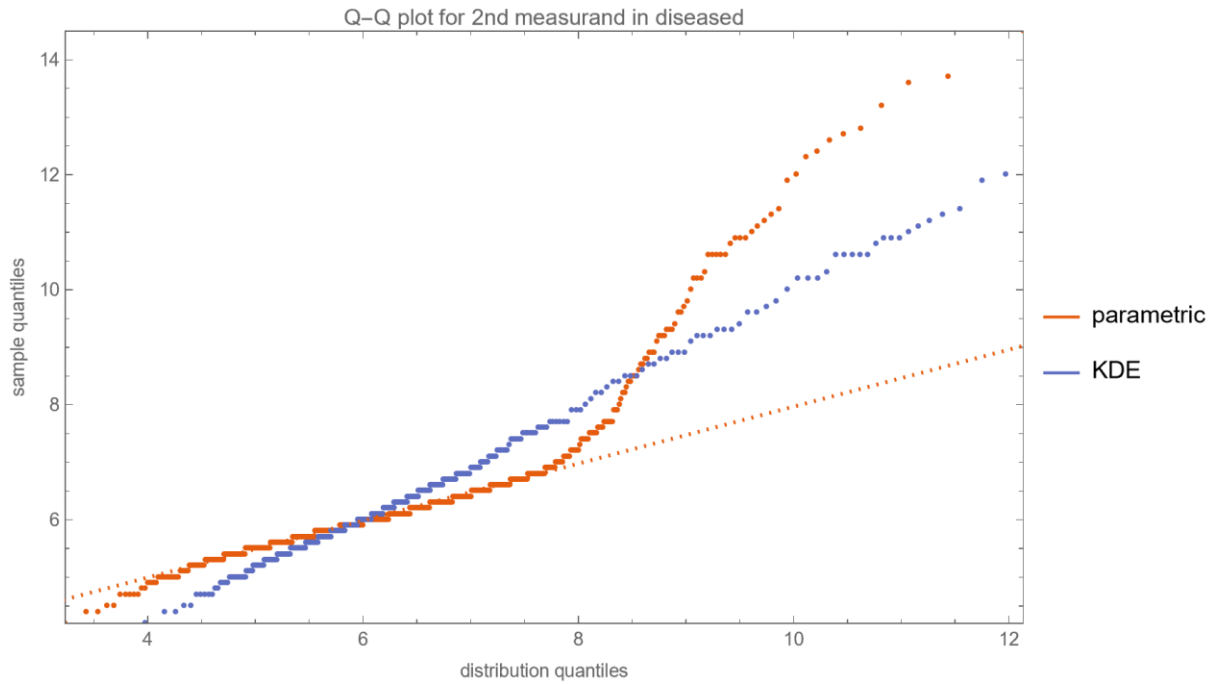
**Figure 7:** The PDF of the second measurand (HbA1c) in nondiseased (nondiabetics), assuming parametric and KDE distributions of the measurand, and the histogram of the respective sample (NHANES dataset), with the settings of the program in Table 2.



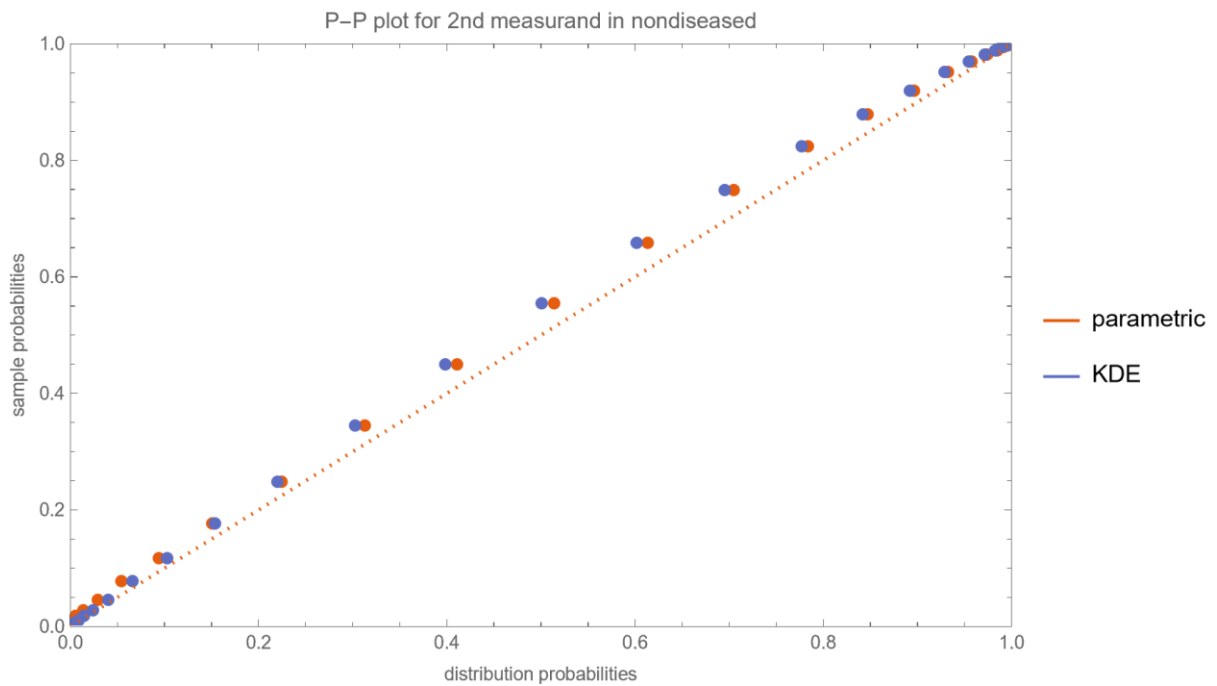
**Figure 8:** The Q-Q plot of the first measurand (FPG) in diseased (diabetics) vs the respective sample (NHANES dataset), assuming parametric and KDE distributions of the measurand, with the settings of the program in Table 2.



**Figure 9:** The P-P plot of the first measurand (FPG) in nondiseased (nondiabetics) vs the respective sample (NHANES dataset), assuming parametric and KDE distributions of the measurand, with the settings of the program in Table 2.



**Figure 10:** The Q-Q plot of the second measurand (HbA1c) in diseased (diabetics) vs the respective sample (NHANES dataset), assuming parametric and KDE distributions of the measurand, with the settings of the program in Table 2.



**Figure 11:** The P-P plot of the second measurand (HbA1c) in nondiseased (nondiabetics) vs the respective sample (NHANES dataset), assuming parametric and KDE distributions of the measurand, with the settings of the program in Table 2.

measurement statistics							
		diseased			nondiseased		
	parameters	parametric	KDE	dataset	parametric	KDE	dataset
1st measurand	mean	133.6000	133.5983	133.5983	98.5000	98.4764	98.4764
	median	126.5821	122.4251	121.0000	97.9555	97.6842	98.0000
	sd	45.1000	46.4389	45.0795	10.4000	10.7048	10.4069
	skewness	1.0512	2.5597	2.8045	0.3179	0.5959	0.6486
	kurtosis	5.0271	10.9001	11.9230	3.1802	4.8336	5.0533
	log-likelihood	-3417.5010	-3231.9360		-39372.7400	-39203.7100	
	correlation coefficient	0.8850	0.8241	0.8853	0.3810	0.3300	0.3805
2nd measurand	mean	6.4100	6.4102	6.4102	5.4000	5.4024	5.4024
	median	6.2644	6.0947	6.1000	5.3860	5.3991	5.4000
	sd	1.3900	1.4285	1.3899	0.3900	0.4089	0.3947
	skewness	0.6607	2.4573	2.6737	0.2170	0.0222	0.0246
	kurtosis	3.7862	10.6549	11.5664	3.0839	3.4319	3.4973
	log-likelihood	-1079.1250	-880.2985		-5199.0250	-5110.5450	

**Figure 12:** The descriptive statistics of both measurands (FPG and HbA1c) in diseased (diabetics) and nondiseased (nondiabetics), assuming parametric and KDE distributions, and of the respective samples (NHANES datasets), with the settings of the program in Table 2.

prior probability of disease			
prevalence		0.0610	
posterior probability of disease			
	parametric	KDE	difference
1st measurand = 126	0.266384	0.324109	0.0577251
2nd measurand = 6.5	0.391503	0.440304	0.0488016
1st measurand = 126 & 2nd measurand = 6.5	0.541984	0.462648	-0.0793361

**Figure 13:** The prior and posterior probabilities of disease (diabetes) for values of the first measurand (FPG) equal to 126 mg/dl and of the second measurand (HbA1c) equal to 6.5 %, assuming parametric and KDE distributions, with the settings of the program in Table 2.

Figures 1-2 show the plots of the posterior probability of diabetes vs FPG and HbA1c respectively. The curves of the parametric distributions are smooth double sigmoidal, while the curves of the nonparametric distributions are multimodal.

Figure 3 shows the plot of the posterior probability of diabetes vs FPG and HbA1c combined. The surface of the parametric distribution is smooth, while the surface of the nonparametric distribution is multimodal.

Figures 4-7 show the PDF of FPG and HbA1c in diabetics and nondiabetics and the histograms of the respective NHANES datasets. It is visually evident that the nonparametric distributions fit the datasets better, especially in diabetic patients.

Figures 8-11 show Q-Q and P-P plots of the PDF of FPG and HbA1c in diabetics and nondiabetics vs the respective NHANES datasets. The plots show clearly that the nonparametric distributions fit the datasets better, especially in diabetic patients.

Figure 12 shows a table with the descriptive statistics of FPG and HbA1c in diabetics and nondiabetics, assuming parametric and KDE distributions, and of the respective NHANES datasets. The data, including the loglikelihood values, support the hypothesis that the nonparametric distributions fit the datasets better, especially in diabetic patients.

Figure 13 shows a table of prior and posterior probabilities for diabetes for values of FPG equal to 126 mg/dl and of HbA1c equal to 6.5 %, the established thresholds of the two measurands for the diagnosis of diabetes (ElSayed et al. 2023), assuming parametric and KDE distributions.

## Discussion

### Reevaluation of Traditional Diagnostic Methods

The findings of the present study highlight the importance of considering to incorporate Bayesian methods in medical diagnosis and management. Conventional approaches based on rigid diagnostic criteria, are often unable to account for the intricate relationships between disease pathology and diagnostic procedures and therefore offer a personalized patient approach (James et al., 2013). In stark contrast, Bayesian methodologies offer a framework that enhances diagnostic precision through a more comprehensive probabilistic assessment (Choi, Johnson, and Thurmond 2006). This Bayesian foundation, therefore, serves as an enabler for tailored medical interventions, echoing similar arguments in existing literature advocating for individualized medicine (Spiegelhalter et al., 2011).

Even though the KDE from our illustrative application, as parameterized in Table 2, provide only an approximate fit to the NHANES datasets for FPG and HbA1c measurements, the posterior probabilities for diabetes delineated in Figure 13 suggest a limited concordance between the classifications of diabetes derived from the OGTT, HbA1c, and FPG tests, as found previously in existing literature (Tucker 2020).

### Challenges and Considerations in Bayesian Analysis for Disease Diagnosis

Despite the evident merits of Bayesian analytics in medical diagnostics, it is paramount to address the intrinsic challenges associated with this methodological shift. One such issue resides in the limited availability of scholarly publications that provide a comprehensive statistical exploration of the measurands in both the diseased and nondiseased populations (Smith and Gelfand 1992).

#### Ramifications of Incomplete Information:

1. *Over-dependence on Prior Probabilities:* The scarcity of empirically derived distributions amplifies reliance on prior probabilities, thereby inducing distortions in the calculation of posterior probabilities. This can result in suboptimal clinical judgments and potentially inaccurate diagnoses (O'Hagan et al. 2006).
2. *Elevated Uncertainty* Insufficient data contributes to broader confidence intervals in the computed posterior probabilities, which, in turn, exacerbates clinical indecisiveness (Berger 1985).



3. *Risk of Bias*: The introduction of systemic bias due to unrepresentative data sets can compromise the fidelity of Bayesian calculations (Gelman et al. 2013).
4. *Imperative for Collaborative Research*: More coordinated research is needed, including multi-center studies, meta-analyses, and open-access databases—to accumulate and disseminate data essential for effective Bayesian diagnostics (McGrayne 2011).
5. *Exploration of Alternative Methodologies*: Given the lack of comprehensive data, the utility of combining Bayesian methods with other statistical techniques or diagnostic modalities becomes increasingly pertinent (George E. P. Box and Tiao 2011).

### Parametric Versus Nonparametric Bayesian Models

In the context of diagnosing diabetes mellitus through FPG and HbA1c levels, our computational tool revealed that nonparametric Bayesian models typically produce a better fit to data distributions, corroborating existing literature that emphasizes the robustness of nonparametric techniques in capturing complex data distributions (Menke et al. 2014; Wasserman 2006).

### Multimodal Versus Double Sigmoidal Bayesian Probability of Disease Curve

The nonparametric Bayesian probabilities for disease exhibited multimodal patterns, in contrast to the bimodal, double sigmoidal curves generated by parametric models.

#### Multimodal Curve

##### Potential Causes:

1. *Complex Pathophysiology*: Multiple etiological pathways may influence the same measurand in divergent ranges, adding layers of complexity to diagnostic processes (Dawid 1984).
2. *Diagnostic Confounders*: External variables affecting the measurand could compromise its efficacy as a standalone diagnostic criterion (Pearl 1994).
3. *Population Subgroups*: The existence of demographically or genetically distinct subgroups within the studied population could also account for the observed multimodality (Heckerman et al., 1995).
4. *Statistical Artifacts*: Demographically or genetically distinct subgroups may be a factor contributing to observed multimodal distributions (Heckerman, Geiger, and Chickering 1995).

##### Theoretical Implications:

Multimodal distributions present a clinical conundrum, compelling healthcare providers to potentially employ additional diagnostic tests or even alternative methodologies (Dawid 1984).

#### Double Sigmoidal Curve

A curve composed of two mirrored sigmoid functions, one delineating the probability behavior for lower measurand values and the other for higher values—offers a fascinating nuance in the realm of diagnostic statistics and medical decision-making.

#### Interpretation

1. *Two Zones of Risk*: Such a curve suggests that the risk of the disease is heightened both at low and high extremes of the measurand but reduced in a middle "safe zone."
2. *Multifactorial Etiology*: This might reflect a situation where both deficiency and excess of a particular biological factor contribute to disease risk. For example, both low and high levels of certain hormones could be problematic.

#### Clinical and Diagnostic Implications

1. *Threshold Decision-making*: Unlike a single sigmoid curve, where one threshold may be adequate for diagnosis, the double-sigmoid may necessitate multiple thresholds, defining a "safe zone" for the measurand.

2. *Treatment Strategies*: Clinicians must be cautious when intervening based on such a measurand, as moving the measurand too far in either direction could heighten risk.
3. *Population Stratification*: This curve shape might imply that different sub-populations or disease subtypes could be better distinguished by additional tests or measurements.

## Shortcomings of this study

The main shortcomings of this study were the following:

1. The OGTT was used as reference diagnostic method for diabetes mellitus. The diagnostic threshold for 2-h plasma glucose (PG) was established in relation to the risk of diabetic retinopathy, a microvascular complication of diabetes mellitus (American Diabetes Association, 2021). However, glucose tolerance is influenced by complex interactions of factors, both physiological and environmental, which pose significant implications for clinical diagnosis and research. The considerations that could affect glucose tolerance and, therefore, the interpretation of the 2-h PG measurement, include the following:
  - 1.1. *Age and Gender*  
Age and gender are significant variables in glucose tolerance. Insulin sensitivity often decreases with age, resulting in higher PG levels (Meneilly and Elliott 1999). Gender differences, particularly related to hormonal changes in females, could also affect glucose metabolism (Geer and Shen 2009).
  - 1.2. *Diurnal Variability*  
Glucose tolerance is subject to diurnal variation, which could affect the 2-h PG test outcomes. Insulin sensitivity is generally higher in the morning than in the evening (Van Cauter, Polonsky, and Scheen 1997).
  - 1.3. *Physical Activity*  
Exercise improves insulin sensitivity and therefore could affect glucose tolerance tests. The timing and intensity of physical activity could have a direct influence on the 2-h PG results (Colberg et al. 2010).
  - 1.4. *Dietary Patterns*  
Short-term and long-term dietary habits, including the macronutrient composition of the diet, may alter the body's glucose and insulin response (Salmerón et al. 1997).
  - 1.5. *Stress and Emotional States*  
The acute stress response includes a transient rise in glucose levels as a result of catecholamine release, potentially affecting the 2-h PG test (Surwit et al. 2002).
  - 1.6. *Medications*  
Certain medications like corticosteroids, antipsychotics, and diuretics affect glucose metabolism, thereby influencing 2-h PG test outcomes (Pandit et al. 1993).
  - 1.7. *Genetic Factors*  
Genetic predispositions influence glucose tolerance, and not accounting for this introduce variability in the 2-h PG test (Dupuis et al. 2010).
2. The lognormal distributions and the KDE, as parameterized in Table 2, fitted only approximately to the NHANES datasets for FPG and HbA1c measurements. It is well known that biological measurands, such as FPG and HbA1c, do not always follow textbook statistical distributions like normal or lognormal distributions. Numerous papers have noted the skewness or kurtosis in the distribution of metabolic variables, urging the use of flexible statistical models (Haeckel, Wosniok, and Arzideh 2007; Arzideh et al. 2007).

## Conclusion and Future Directions

The intricacies of the double-sigmoid curve and multimodal distributions introduce a new frontier in personalizing healthcare provision. While smoother relationships between measurements and Bayesian probability facilitate clinical interpretability, multimodal distributions might serve as sentinel indicators of underlying complexities or methodological shortcomings, thus providing a useful tool in the field of medical diagnosis.

As a pivotal next step, future research should aim to validate the utility and reliability of the Bayesian inference based method applied in this study through real-world clinical trials, in addition to extending its application to include more diagnostic modalities. The ultimate aim is to combine this approach with existing clinical protocols, thereby optimizing the diagnostic precision and consequently improving patient outcomes.

In addition to its potential for clinical applications, the computational tool developed for this study could hold considerable promise as an educational and research adjunct. By facilitating the analysis of Bayesian probabilities in disease diagnosis, it serves as an invaluable resource for both medical practitioners in training and experienced researchers in the field. Its modular design and user-friendly interface make it easily adaptable to various research settings and educational curricula, thereby accelerating the adoption and dissemination of Bayesian approaches in medical statistics and diagnostics.

## Acknowledgements

The authors extend sincere gratitude to OpenAI for ChatGPT's contributions to the manuscript drafting, editing and data interpretation, to Wolfram Research for their Mathematica software and computational plugin, which were critical for the symbolic and numerical computations, and to Google Scholar for its invaluable support in literature review.

## Supplementary Material

The program is available online as a Wolfram Mathematica Notebook, at <https://www.hcsl.com/Tools/BayesianDiagnosis/BayesianDiagnosis.nb>

## Author Contributions

Conceptualization: T.C.; methodology: T.C. and A.H.; software: T.C. and A.H.; validation: T.C.; formal analysis: T.C. and A.H.; investigation: T.C.; resources: A.H.; data curation: T.C.; writing—original draft preparation: T.C.; writing—review and editing: A.H.; visualization: T.C.; supervision: A.H.; project administration: T.C.

All authors have read and agreed to the published version of the manuscript.

## Funding

This research received no external funding.

## Institutional Review Board Statement

Data collection was carried out following the rules of the Declaration of Helsinki. The Ethics Review Board of the National Center for Health Statistics approved data collection and posting the data online for public use (National Center for Health Statistics 2022).

## Informed Consent Statement

Written consent was obtained from each subject participating in the survey.

## Data Availability Statement

The data presented in this study are available at <https://wwwn.cdc.gov/nchs/nhanes/default.aspx> (Accessed at 25/07/2023).

## Conflicts of Interest

The authors declare no conflict of interest.

## Appendix I

### Formalisms and Notation

#### Abbreviations

PDF: probability density function(s)

CDF: cumulative density function(s)

KDE: kernel density estimator(s)

OGTT: oral glucose tolerance test

PG: plasma glucose

FPG: fasting plasma glucose

HbA1c: glycated hemoglobin A1c

NHANES: National Health and Nutrition Examination Survey

#### Datasets

$\mathbf{x}$ : a dataset  $x_1, x_2, \dots, x_n$  of measurements

#### Parameters

$\nu$ : prevalence of disease

$\mu, m$ : mean

$\sigma, s$ : standard deviation

$\rho, r$ : correlation coefficient

$k$ : shape parameter

$\vartheta$ : scale parameter

$h$ : nonparametric kernel density bandwidth

## Functions

$f^{-1}$ : the inverse of the function  $f$

$|H|$ : determinant of the matrix  $H$

$P(A)$ : probability of the event  $A$

$P(A|B)$ : conditional probability of the event  $A$  given the event  $B$

$cov(X, Y)$ : covariance of two jointly distributed random variables  $X$  and  $Y$

$\mathbb{E}[Z]$ : expected value of a random variable  $Z$

$\ln(x)$  : natural logarithm

$\mathcal{L}(\theta|z)$ : likelihood function

$l(\theta|z)$ : loglikelihood function

$p(x)$ : probability mass function

$P_Q(k; q)$ : the  $k$ -th  $q$ -quantile of a random variable

$erf(z)$ : error function

$erfc(z)$ : complementary error function

$\Gamma(z)$ : gamma function

$\gamma(z, x)$ : incomplete gamma function

$Q(a, z)$ : regularized incomplete gamma function

$\gamma(z, x_0, x_1)$ : generalized incomplete gamma function

$Q(z, x_0, x_1)$ : regularized generalized incomplete gamma function

$K(u)$ : kernel function

$f(x)$ : univariate PDF

$f(x|\theta)$ : univariate PDF with parameters  $\theta$

$f(x, y)$ : bivariate PDF

$f(x, y|\theta)$ : bivariate PDF with parameters  $\theta$

$F(x)$ : univariate cumulative PDF

$F(x|\theta)$ : univariate cumulative PDF with parameters  $\theta$

$F(x, y)$ : bivariate cumulative PDF

$F(x, y|\theta)$ : bivariate cumulative PDF with parameters  $\theta$

## Definitions of Functions

### Inverse Function

The inverse function  $f^{-1}$  of a function  $f$  (also called the inverse of  $f$ ) is a function that undoes the operation of  $f$ . Therefore:

$$f^{-1}(f(x)) = x$$

and

$$f(f^{-1}(y)) = y$$

## Natural Logarithm

$$\ln(x) = \int_1^x \frac{1}{t} dt$$

## Error Function

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

## Complementary Error Function

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x)$$

## Gamma Function

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$$

for all complex numbers  $z$ , except the non-positive integers.

## Incomplete Gamma Function

$$\gamma(z, x) = \int_x^\infty t^{z-1} e^{-t} dt$$

## Regularized Incomplete Gamma Function

$$Q(z, x) = \frac{\gamma(z, x)}{\Gamma(z)}$$

## Generalized Incomplete Gamma Function

$$\gamma(z, x_0, x_1) = \int_{x_0}^{x_1} t^{z-1} e^{-t} dt$$

## Regularized Generalized Incomplete Gamma Function

$$Q(z, x_0, x_1) = \frac{\gamma(z, x_0, x_1)}{\Gamma(z)}$$

## Probability Density Function

### Univariate

The probability density function (PDF) is a statistical function that describes the likelihood of a continuous random variable taking on a particular value.

For a continuous random variable  $X$ , the PDF, denoted by  $f(x)$ , is defined as:

$$f(x) = \lim_{\Delta x \rightarrow 0} \frac{P(x \leq X < x + \Delta x)}{\Delta x}$$

where  $P(x \leq X < x + \Delta x)$  is the probability that the random variable  $X$  falls within the interval  $[x, x + \Delta x)$

## Bivariate

The bivariate PDF is a statistical measure that describes the likelihood of two continuous random variables  $X$  and  $Y$ , taking on particular values  $x$  and  $y$ . It is denoted as  $f_{X,Y}(x, y)$  and defined as:

$$f_{X,Y}(x, y) = \lim_{\Delta x, \Delta y \rightarrow 0} \frac{P(x \leq X < x + \Delta x, y \leq Y < y + \Delta y)}{\Delta x \Delta y}$$

where  $P(x \leq X < x + \Delta x, y \leq Y < y + \Delta y)$  is the probability that the random variables  $X$  and  $Y$  fall within the intervals  $[x, x + \Delta x)$  and  $[y, y + \Delta y)$  respectively.

## Cumulative Distribution Function

### Univariate

The univariate cumulative distribution function (CDF) is closely related to the PDF and provides the cumulative probability for a random variable up to a specific value.

For a random variable  $X$ , the CDF, denoted by  $F(x)$ , is defined as:

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt$$

where  $f(t)$  is the PDF of the random variable.

The CDF is the integral of the PDF, and conversely, the PDF is the derivative of the CDF (when it exists):

$$f(x) = \frac{dF(x)}{dx}$$

### Bivariate

The bivariate CDF is a function that describes the probability that the random variables  $X$  and  $Y$  simultaneously take on values less than or equal to  $x$  and  $y$ , respectively. It is denoted as  $F_{X,Y}(x, y)$  and defined as:

$$F_{X,Y}(x, y) = \int_{-\infty}^x \int_{-\infty}^y f_{X,Y}(u, v) du dv$$

## Skewness

Skewness is a statistical measure that describes the asymmetry of a probability distribution about its mean. It quantifies the extent and direction of skew (departure from horizontal symmetry) in the data.

$$skewness(X) = \frac{\mathbb{E}[(X - \mu)^3]}{\sigma^3}$$

where  $X$  is a random variable and  $\mu$  and  $\sigma$  are the mean and the standard deviation of  $X$ .

If  $skewness(X) < 0$ , the distribution is said to be left-skewed. If  $skewness(X) > 0$ , is said to be right-skewed. If  $skewness(X) = 0$ , the distribution is symmetric.

## Kurtosis

Kurtosis is a statistical measure that quantifies how heavy the tails of a distribution are compared to a normal distribution.

$$kurtosis(X) = \frac{\mathbb{E}[(X - \mu)^4]}{\sigma^4}$$

where  $X$  is a random variable and  $\mu$  and  $\sigma$  are the mean and the standard deviation of  $X$ .

If  $kurtosis(X) = 3$ , the distribution has the same kurtosis as the normal distribution (mesokurtic).

If  $kurtosis(X) < 3$ , the distribution is platykurtic (light tails).

If  $kurtosis(X) > 3$ , the distribution is leptokurtic (heavy tails).

## Correlation Coefficient

The correlation coefficient  $\rho_{X,Y}$  of two random variables  $X$  and  $Y$ , with means  $\mu_X$  and  $\mu_Y$ , is defined as:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

where

$$cov(X,Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

Given two samples of independent and identically distributed (i.i.d.) data points  $x_1, x_2, \dots, x_n$ , and  $y_1, y_2, \dots, y_n$ , with means  $\mu_X$  and  $\mu_Y$ , their correlation coefficient  $\rho_{X,Y}$  is defined as:

$$\rho_{X,Y} = \frac{\sum_{i=1}^n (x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sum_{i=1}^n (x_i - \mu_X)^2} \sqrt{\sum_{i=1}^n (y_i - \mu_Y)^2}}$$

The correlation coefficient quantifies the strength and direction of the linear relationship between  $X$  and  $Y$ . We have  $-1 \leq \rho_{X,Y} \leq 1$ . If  $\rho_{X,Y} = 0$  it is implied that there is no linear dependency between the respective variables. If  $\rho_{X,Y} = 1$  it signifies a perfect linear relationship between the variables. If  $\rho_{X,Y} = -1$  it signifies a perfect negative linear relationship.

## Loglikelihood Function

The likelihood function of a random variable  $X$  is defined as:

$$\mathcal{L}(\theta|x) = f(x|\theta)$$

where  $f(x|\theta)$  is the PDF of  $X$  given a parameter set  $\theta$ .

The likelihood and loglikelihood functions of a parameter set  $\theta$ , given a dataset  $\mathbf{x} = x_1, x_2, \dots, x_n$  of independent and identically distributed (i.i.d.) data points of a random variable  $X$ , are defined as:

$$\mathcal{L}(\theta|\mathbf{x}) = \prod_{i=1}^n f(x_i|\theta)$$

$$l(\theta|\mathbf{x}) = \sum_{i=1}^n \ln(f(x_i|\theta))$$

where  $f(x_i; \theta)$  is the PDF of  $X$ .

## Quantiles

A quantile is a statistical term that refers to dividing a probability distribution into continuous intervals with equal probabilities or dividing a dataset into subsets with the same probability mass  $p_X(x)$ , where  $p_X(x)$  is a function that gives the probability that a discrete random variable is exactly equal to some value:

$$p_X(x) = P(X = x)$$



Specifically, the  $k$ -th  $q$ -quantile of a probability distribution or a dataset is a numerical value that divides the data into  $q$  equal parts, such that exactly  $\frac{k}{q}$  of the data or probability distribution is less than or equal to that value.

The  $k$ -th  $q$ -quantile of a probability distribution with CDF  $F(x)$  is given by (Hyndman and Fan 1996):

$$P_Q(k; q) = F^{-1}\left(\frac{k}{q}\right)$$

where  $F^{-1}$  is the inverse of the CDF.

In the context of empirical data, the  $k$ -th  $q$ -quantile is a value that partitions the data into  $q$  equally probable subsets.

## Bayes Theorem

For the purposes of our study, Bayes theorem is formulated as follows:

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)} = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|\bar{D})(1 - P(D))}$$

where:

$P(D|T)$  represents the posterior probability of having the disease given a set of test results  $\mathbf{x}$ .

$P(T|D)$  denotes the likelihood of obtaining the set of test results given the presence of the disease.

$P(T|\bar{D})$  denotes the likelihood of obtaining the set of test results given the absence of the disease.

$P(D)$  is the prior probability or prevalence  $v$  of the disease.

$P(T)$  signifies the overall probability of the set of test results.

Therefore, for a parameter set  $\theta$ :

$$P(D|T) = \frac{\mathcal{L}_D(\theta|\mathbf{x})v}{\mathcal{L}_D(\mathbf{x}|\theta)v + \mathcal{L}_{\bar{D}}(\mathbf{x}|\theta)(1 - v)} = \frac{f_D(\mathbf{x}|\theta)v}{f_D(\mathbf{x}|\theta)v + f_{\bar{D}}(\mathbf{x}|\theta)(1 - v)}$$

where  $\mathcal{L}_D(\theta|\mathbf{x})$  and  $f_D(\mathbf{x}|\theta)$  denote the likelihood function and the PDF in the presence of the disease, while  $\mathcal{L}_{\bar{D}}(\mathbf{x}|\theta)$  and  $f_{\bar{D}}(\mathbf{x}|\theta)$  denote the respective functions in the absence of the disease.

## Q-Q plot

A  $Q$ - $Q$  plot is constructed by plotting the quantiles from a distribution and a dataset against each other. If the dataset comes from the theoretical distribution, the points in the  $Q$ - $Q$  plot will approximately lie on the line  $y = x$ .

## P-P plot

A  $P$ - $P$  plot is constructed by plotting the cumulative probabilities from a distribution and a dataset against each other. If the dataset comes from the theoretical distribution, the points in the  $P$ - $P$  plot will approximately lie on the line  $y = x$ .

## Parametric Distributions

### Normal Distribution

#### Univariate

The univariate normal distribution or Gaussian distribution is a continuous probability distribution of a real-valued random variable  $X$ . The general form of its PDF is:

$$f_N(x; \mu, \sigma) = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}$$

where the parameter  $\mu$  is the mean or expectation of  $X$ , while the parameter  $\sigma$  is its standard deviation (Forbes et al. 2011).

#### Bivariate

The bivariate normal distribution or Gaussian distribution is a continuous probability distribution of two normally distributed random variables  $X$  and  $Y$ . The general form of its PDF is:

$$f_N(x, y; \mu_X, \sigma_X, \mu_Y, \sigma_Y, \rho) = \frac{e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)}}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}$$

where the parameters  $\mu_X$  and  $\mu_Y$  are the means of  $X$  and  $Y$ ,  $\sigma_X$  and  $\sigma_Y$  are their standard deviations, and  $\rho$  their correlation coefficient (Forbes et al. 2011).

### Lognormal Distribution

#### Univariate

The univariate lognormal distribution is a continuous probability distribution of a random variable  $X$  whose logarithm is normally distributed. The general form of its PDF is:

$$f_L(x; m, s) = \frac{e^{-\frac{1}{2}\left(\frac{\ln(x)-m}{s}\right)^2}}{xs\sqrt{2\pi}}$$

where the parameter  $m$  is the mean and  $s$  the standard deviation of  $\ln(X)$  (Forbes et al. 2011).

If  $\mu$  and  $\sigma$  are the mean and the standard deviation of  $X$  we have:

$$\mu = e^{m + \frac{1}{2}s^2}$$
$$\sigma = \sqrt{e^{2m+2s^2}}$$

Therefore,

$$m = \ln\left(\frac{\mu^2}{\sqrt{\sigma^2 + \mu^2}}\right)$$
$$s = \ln\left(1 + \frac{\sigma^2}{\mu^2}\right)$$

$$f_L(x; \mu, \sigma) = \frac{e^{\left( \frac{1}{2} \left( \frac{\ln(x) - \ln\left(\frac{\mu^2}{\sqrt{\sigma^2 + \mu^2}}\right)}{\ln\left(1 + \frac{\sigma^2}{\mu^2}\right)} \right)^2 \right)}}{\sqrt{2\pi} x \ln\left(1 + \frac{\sigma^2}{\mu^2}\right)} = \frac{e^{\left( \frac{(2\ln(x) - 2\ln(\mu) + \ln(1 + \frac{\sigma^2}{\mu^2}))^2}{8\ln\left(1 + \frac{\sigma^2}{\mu^2}\right)} \right)}}{x \sqrt{2\pi \ln\left(1 + \frac{\sigma^2}{\mu^2}\right)}}$$

### Bivariate

The bivariate lognormal distribution is a continuous probability distribution of two lognormally distributed variables  $X$  and  $Y$ . If  $m_X$  and  $m_Y$  are the means of  $\ln(X)$  and  $\ln(Y)$ ,  $s_X$  and  $s_Y$  are their standard deviations, and  $r$  their correlation coefficient, the general form of its PDF is (Forbes et al. 2011):

$$f_L(x, y; m_X, s_X, m_Y, s_Y, r) = \frac{1}{d} e^a$$

where

$$a = \frac{1}{2} \left( \frac{-(\ln(y) - m_Y)b - (\ln(x) - m_X)c}{s_X^2 s_Y^2 - r^2 s_X^2 s_Y^2} \right)$$

$$b = (\ln(y) - m_Y)s_X^2 - r(\ln(x) - m_X)s_X s_Y$$

$$c = (\ln(x) - m_X)s_Y^2 - r(\ln(y) - m_Y)s_X s_Y$$

$$d = 2\pi xy \sqrt{s_X^2 s_Y^2 - r^2 s_X^2 s_Y^2}$$

We have

$$r = \frac{\mu_X \mu_Y}{\sigma_X \sigma_Y} \left( -1 + e^{\rho \sqrt{\ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right) \ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)}} \right) \sqrt{\ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right) \ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)}$$

where  $\mu_X$  and  $\mu_Y$  are the means of  $X$  and  $Y$ ,  $\sigma_X$  and  $\sigma_Y$  are their standard deviations and  $\rho$  their correlation coefficient.

Therefore,

$$f_L(x, y; \mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho) = \frac{e^{\frac{ab+c}{d}}}{g}$$

where

$$a = -2 \left( -1 + e^{\rho \sqrt{\ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right) \ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)}} \right) \left( \ln(x) - \ln\left(\frac{\mu_X^2}{\sqrt{\mu_X^2 + \sigma_X^2}}\right) \right)$$

$$b = m_X m_Y \sqrt{\ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right) \ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)} \left( \ln(y) - \ln\left(\frac{\mu_Y^2}{\sqrt{\mu_Y^2 + \sigma_Y^2}}\right) \right)$$

$$c = \left( \ln(y) - \ln\left(\frac{\mu_Y^2}{\sqrt{\mu_Y^2 + \sigma_Y^2}}\right) \right)^2 \sigma_X^2 + \left( \ln(x) - \ln\left(\frac{\mu_X^2}{\sqrt{\mu_X^2 + \sigma_X^2}}\right) \right)^2 \sigma_Y^2$$

$$d = 2 \left( \left( -1 + e^{\rho \sqrt{\ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right) \ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)}} \right)^2 \ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right) \ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right) \mu_X^2 \mu_Y^2 - \sigma_X^2 \sigma_Y^2 \right)$$

$$g = 2\pi xy \sqrt{- \left( -1 + e^{\rho \sqrt{\ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right) \ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)}} \right)^2 \ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right) \ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right) \mu_X^2 \mu_Y^2 + \sigma_X^2 \sigma_Y^2}$$

## Gamma Distribution

### Univariate

The univariate Gamma distribution is a continuous probability distribution of a random variable  $X$ . The general form of its PDF is:

$$f_G(x; k, \vartheta) = \frac{1}{\Gamma(k)\vartheta^k} x^{k-1} e^{-\frac{x}{\vartheta}}$$

where  $k$  is a shape parameter,  $\vartheta$  is a scale parameter and  $\Gamma(u)$  the gamma function (Forbes et al. 2011).

The mean  $\mu$  and the standard deviation  $\sigma$  of  $X$ , are calculated as following:

$$\mu = k\vartheta$$

$$\sigma = k\vartheta^2$$

Therefore,

$$k = \frac{\mu^2}{\sigma^2}$$

$$\vartheta = \frac{\sigma^2}{\mu}$$

and

$$f(x; \mu, \sigma) = \frac{1}{\Gamma\left(\frac{\mu^2}{\sigma^2}\right) \left(\frac{\sigma^2}{\mu}\right)^{\frac{\mu^2}{\sigma^2}}} x^{\left(\frac{\mu^2}{\sigma^2}-1\right)} e^{-\frac{x\mu}{\sigma^2}}$$

### Bivariate

The bivariate Gamma distribution is a continuous probability distribution of two variables  $X$  and  $Y$ . The copula version of its PDF is:

$$f(x, y; k_X, k_Y, \vartheta_X, \vartheta_Y, \rho) = \frac{ab}{c}$$

where

$$a = e^{\left( \operatorname{erfc}^{-1}\left(2Q\left(k_Y, 0, \frac{y}{\vartheta_Y}\right)\right)^2 + \frac{\left(-\rho \operatorname{erfc}^{-1}\left(2Q\left(k_X, 0, \frac{x}{\vartheta_X}\right)\right) + \operatorname{erfc}^{-1}\left(2Q\left(k_Y, 0, \frac{y}{\vartheta_Y}\right)\right)\right)^2}{-1+\rho^2} - \frac{y}{\vartheta_Y} - \frac{x}{\vartheta_X} \right)}$$

$$b = x^{-1+k_X} y^{-1+k_Y} \vartheta_Y^{-k_Y} \vartheta_X^{-k_X}$$

$$c = \sqrt{1-\rho^2} \Gamma(k_X) \Gamma(k_Y)$$

and where  $k_X, k_Y$  are shape parameters,  $\vartheta_X, \vartheta_Y$  are scale parameters and  $\rho$  the correlation coefficient of  $X$  and  $Y$ .

If  $\mu_X$  and  $\mu_Y$  are the means of  $X$  and  $Y$ ,  $\sigma_X$  and  $\sigma_Y$  are their standard deviations, and  $\rho$  their correlation coefficient, it can be shown that:

$$f(x, y; \mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho) = \frac{ab}{c}$$

where

$$a = e^{\left( \operatorname{erfc}^{-1}\left(2Q\left(\frac{\mu_Y^2}{\sigma_Y^2}, 0, \frac{y\mu_Y}{\sigma_Y^2}\right)\right)^2 + \frac{\left(-\rho \operatorname{erfc}^{-1}\left(2Q\left(\frac{\mu_X^2}{\sigma_X^2}, 0, \frac{x\mu_X}{\sigma_X^2}\right)\right) + \operatorname{erfc}^{-1}\left(2Q\left(\frac{\mu_Y^2}{\sigma_Y^2}, 0, \frac{y\mu_Y}{\sigma_Y^2}\right)\right)\right)^2}{-1+\rho^2} - \frac{x\mu_X}{\sigma_X^2} - \frac{y\mu_Y}{\sigma_Y^2} \right)}$$

$$b = x^{\left(-1+\frac{\mu_X^2}{\sigma_X^2}\right)} y^{\left(-1+\frac{\mu_Y^2}{\sigma_Y^2}\right)} \sigma_X^{-\frac{2\mu_X^2}{\sigma_X^2}} \mu_X^{\frac{\mu_X^2}{\sigma_X^2}} \mu_Y^{\frac{\mu_Y^2}{\sigma_Y^2}} \sigma_Y^{-\frac{2\mu_Y^2}{\sigma_Y^2}}$$

$$c = \sqrt{1-\rho^2} \Gamma\left(\frac{\mu_X^2}{\sigma_X^2}\right) \Gamma\left(\frac{\mu_Y^2}{\sigma_Y^2}\right)$$

## Copulas

If  $\mu_X$  and  $\mu_Y$  are the means of the variables  $X$  and  $Y$ ,  $\sigma_X$  and  $\sigma_Y$  are their standard deviations, and  $\rho$  their correlation coefficient, it can be shown that the bivariate PDF of the other copulas of the program are defined as follows:

*X: Normally Distributed – Y: Lognormally Distributed*

$$f_{NL}(x, y; \mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho) = \frac{e^c d}{g}$$

where

$$\begin{aligned}
a &= -\frac{2\ln(y) - 2\ln(\mu_Y) + \ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)^2}{2\sqrt{2\ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)}} - \frac{\left(2\ln(y) - 2\ln(\mu_Y) + \ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)\right)^2}{8\ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)} \\
b &= -\frac{\left(\frac{2\ln(y) - 2\ln(\mu_Y) + \ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)}{2\sqrt{2}\sqrt{\ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)}}\sqrt{\ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)}\mu_Y + \rho\sigma_Y\operatorname{erfc}^{-1}\left(2Q\left(\frac{\mu_X^2}{\sigma_X^2}, 0, \frac{x\mu_X}{\sigma_X^2}\right)\right)\right)^2}{\ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)\mu_Y^2 - \rho^2\sigma_Y^2} \\
c &= a + b - \frac{x\mu_X}{\sigma_X^2} \\
d &= \left(\frac{x\mu_X}{\sigma_X^2}\right)^{\frac{\mu_X^2}{\sigma_X^2}}
\end{aligned}$$

$$g = xy\Gamma\left(\frac{\mu_X^2}{\sigma_X^2}\right)\sqrt{2\pi\ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)\left(1 - \frac{\rho^2\sigma_Y^2}{\ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)\mu_Y^2}\right)}$$

*X: Lognormally Distributed – Y: Normally Distributed*

$$f_{LN}(x, y; \mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho) = \frac{e^c d}{g}$$

where

$$\begin{aligned}
a &= \operatorname{erfc}^{-1}\left(2Q\left(\frac{\mu_Y^2}{\sigma_Y^2}, 0, \frac{y\mu_Y}{\sigma_Y^2}\right)\right)^2 - \frac{\left(2\ln(x) - 2\ln(\mu_X) + \ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right)\right)^2}{8\ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right)} \\
b &= -\frac{\left(\operatorname{erfc}^{-1}\left(2Q\left(\frac{\mu_Y^2}{\sigma_Y^2}, 0, \frac{y\mu_Y}{\sigma_Y^2}\right)\right)\sqrt{\ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right)}\mu_X + \rho\sigma_X\left(\frac{2\ln(x) - 2\ln(\mu_X) + \ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right)}{2\sqrt{2}\sqrt{\ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right)}}\right)\right)^2}{\ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right)\mu_X^2 - \rho^2\sigma_X^2} \\
c &= a + b - \frac{y\mu_Y}{\sigma_Y^2}
\end{aligned}$$

$$d = \left( \frac{y\mu_Y}{\sigma_Y^2} \right)^{\frac{m_Y^2}{s_Y^2}}$$

$$g = \left( \sqrt{2\pi}xy\Gamma\left(\frac{\mu_Y^2}{\sigma_Y^2}\right) \sqrt{\ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right)} \sqrt{1 - \frac{\rho^2\sigma_X^2}{\ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right)\mu_X^2}} \right)$$

*X: Normally Distributed – Y: Gamma Distributed*

$$f_{NG}(x, y; \mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho) = \frac{e^{\left( \operatorname{erfc}^{-1}\left(2Q\left(\frac{\mu_Y^2}{\sigma_Y^2}, 0, \frac{y\mu_Y}{\sigma_Y^2}\right)\right)^2 - \frac{(x-\mu_X)^2}{2\sigma_X^2} + \frac{\left(x\rho - \rho\mu_X + \sqrt{2}\sigma_X\operatorname{erfc}^{-1}\left(2Q\left(\frac{\mu_Y^2}{\sigma_Y^2}, 0, \frac{y\mu_Y}{\sigma_Y^2}\right)\right)\right)^2}{2(-1+\rho^2)\sigma_X^2} - \frac{y\mu_Y}{\sigma_Y^2} \right) \left(\frac{y\mu_Y}{\sigma_Y^2}\right)^{\frac{\mu_Y^2}{\sigma_Y^2}}}}{y\sigma_X\sqrt{2\pi(1-\rho^2)}\Gamma\left(\frac{y\mu_Y}{\sigma_Y^2}\right)}$$

*X: Gamma Distributed– Y: Normally Distributed*

$$f_{GN}(x, y; \mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho) = \frac{e^{\left( \frac{x\mu_X}{s_X^2} + \frac{\left(y - m_Y + \sqrt{2}\rho\operatorname{erfc}^{-1}\left(2Q\left(\frac{\mu_X^2}{\sigma_X^2}, 0, \frac{x\mu_X}{\sigma_X^2}\right)\right)\sigma_Y\right)^2}{2(-1+\rho^2)\sigma_Y^2} \right) \left(\frac{x\mu_X}{\sigma_X^2}\right)^{\frac{\mu_X^2}{\sigma_X^2}}}}{x\sigma_Y\sqrt{2\pi(1-\rho^2)}\Gamma\left(\frac{\mu_X^2}{s\sigma_X^2}\right)}$$

*X: Lognormally Distributed – Y: Gamma Distributed*

$$f_{LG}(x, y; \mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho) = \frac{e^c}{d}$$

where

$$a = \left( \frac{2\ln(y) - 2\ln(\mu_Y) + \ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)}{2\sqrt{2\ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)}} \right)^2 - \frac{\left(2\ln(y) - 2\ln(\mu_Y) + \ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)\right)^2}{8\ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)}$$

$$b = - \frac{\left( \left( -\ln(y) + \ln(\mu_Y) - \frac{\ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)}{2} \right) m_Y\sigma_X + \rho(x - \mu_X)\sigma_Y \right)^2}{2\sigma_X^2 \left( \ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right) \mu_Y^2 - \rho^2\sigma_Y^2 \right)}$$

$$c = a + b - \frac{(x - \mu_X)^2}{2\sigma_X^2}$$

$$d = 2\pi y \sigma_X \sqrt{\ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right) \left(1 - \frac{\rho^2 \sigma_Y^2}{\ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right) \mu_Y^2}\right)}$$

*X: Gamma Distributed – Y: Lognormally Distributed*

$$f_{GL}(x, y; \mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho) = \frac{e^{a+b}}{c}$$

where

$$a = - \frac{\left(2\ln(x) - 2\ln(\mu_X) + \ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right)\right)^2}{8\ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right)}$$

$$b = - \frac{\left(\sqrt{\ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right)} \mu_X (y - \mu_Y) - \rho \sigma_X \sigma_Y \left(\frac{2\ln(x) - 2\ln(\mu_X) + \ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right)}{2\sqrt{\ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right)}}\right)\right)^2}{2\sigma_Y^2 \left(\ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right) m_X^2 - \rho^2 \sigma_X^2\right)}$$

$$c = 2\pi x \sigma_Y \sqrt{\ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right) \left(1 - \frac{\rho^2 \sigma_X^2}{\ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right) m_X^2}\right)}$$

## Nonparametric Distributions

### Histograms

A histogram is a graphical representation of the distribution of a dataset. It is an estimate of the probability distribution of a continuous variable. To construct a histogram:

1. The data range is divided into a set of bins.
2. The data points are sorted into each bin.
3. The number of data points that fall into each bin are counted.

The height of each bar in the histogram corresponds to the count of data points in bin. The width of each bar corresponds to the width of the bin.

The Knuth method(Knuth 2019) is a Bayesian approach to determining the optimal number of bins for a histogram. It calculates the optimal bin width by maximizing a likelihood function, considering the data as independently and identically distributed (i.i.d.).

Given a dataset  $X = \{x_1, x_2, \dots, x_n\}$ , we find the optimal bin edges  $B = \{b_1, b_2, \dots, b_k\}$ , by maximizing the following likelihood function:



$$\mathcal{L}(B|X) = n! \left( \prod_{i=1}^k \frac{1}{n_i!} \right) \frac{1}{k^n} \frac{1}{(b_k - b_0)^n}$$

where  $n$  is the total number of observations,  $k$  is the number of bins,  $n_i$  is the number of observations in the  $i$ -th bin, and  $b_0$  and  $b_k$  are the minimum and maximum bin edges, respectively.

There are variations of histograms where the height of bars represents relative frequencies (proportions or probabilities) instead of raw counts. In such cases, the area under the histogram integrates to 1.

### Kernel Density Estimators

Given a set of independent and identically distributed (i.i.d.) data points  $\{x_1, x_2, \dots, x_n\}$ , the univariate KDE  $\hat{f}_K(x; n, h)$  is defined as (Gramacki 2017):

$$\hat{f}_K(x; n, h) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$

Given two sets of independent and identically distributed (i.i.d.) data points  $\{x_1, x_2, \dots, x_n\}$ , and  $\{y_1, y_2, \dots, y_n\}$ , the bivariate KDE  $\hat{f}(x, y; n, h_1, h_2)$  is defined as (Gramacki 2017):

$$\hat{f}(x, y; n, h_1, h_2) = \frac{1}{n|H|^{\frac{1}{2}}} \sum_{i=1}^n K((z - z_i)^T H^{-1} (z - z_i))$$

where

$$z = \begin{bmatrix} x \\ y \end{bmatrix}$$

$$z_i = \begin{bmatrix} x_i \\ y_i \end{bmatrix}$$

$$H = \begin{bmatrix} h_1^2 & \rho h_1 h_2 \\ \rho h_1 h_2 & h_2^2 \end{bmatrix}$$

and

1.  $n$  is the number of datapoints,
2.  $h$  is the bandwidth, a positive scalar that determines the width and smoothness of the kernel. If  $h$  is too small, the estimate could be overly sensitive to noise in the data, leading to a "noisy" multimodal estimate. Conversely, if  $h$  is too large, the estimate could be overly smooth, potentially obscuring meaningful features in the data.
3.  $\rho$  is the correlation coefficient of the two sets of datapoints.
4.  $K(u)$  is the kernel function, which satisfies the properties:
  - 4.1.  $\int K(u) du = 1$
  - 4.2.  $\int u^2 K(u) du < \infty$

A kernel function  $K(u)$  could be conceptualized as a weighting mechanism in the context of kernel density estimation. For every observed data point  $u_i$  the kernel function  $K(u)$  superimposes a localized influence or "perturbation" centered at  $u_i$ . The magnitude and dispersion of this perturbation are governed by the properties of  $K(u)$  and the bandwidth parameter  $h$ , respectively. Specifically, the amplitude of the perturbation at  $u_i$  is contingent upon the value of  $K(u_i)$ , while the scale or spread of this influence is modulated by  $h$ . This ensures that each data point contributes to

the overall density estimate in a manner that is both localized and smooth, with the degree of localization and smoothness being adjustable via the choice of  $K(u)$  and  $h$ .

The program uses the Gaussian kernel function:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

### Univariate Kernel Density Estimator

$$\hat{f}(x; n, h) = \frac{1}{nh} \sum_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{\left(\frac{x-x_i}{h}\right)^2}{2}}$$

### Bivariate Kernel Density Estimator

$$\hat{f}(x, y; n, h_1, h_2) = \frac{1}{2\pi n |H|^{\frac{1}{2}}} \sum_{i=1}^n e^{-\frac{1}{2}(z-z_i)^T H^{-1}(z-z_i)}$$

where

$$z = \begin{bmatrix} x \\ y \end{bmatrix}$$

$$z_i = \begin{bmatrix} x_i \\ y_i \end{bmatrix}$$

$$H = \begin{bmatrix} h_1^2 & \rho h_1 h_2 \\ \rho h_1 h_2 & h_2^2 \end{bmatrix}$$

## Appendix II

### Software Availability and requirements

**Program name:** *Bayesian Diagnosis*

**Project home page:** <https://www.hcsl.com/Tools/BayesianDiagnosis/> (accessed 31 August 2023)

**Operating systems:** Microsoft Windows, Linux, Apple iOS

**Programming language:** Wolfram Language

**Other software requirements:**

*For running the program:* Wolfram Player®, freely available at: <https://www.wolfram.com/player/> (accessed 31 August 2023) or Wolfram Mathematica®.

*For editing the datasets:* Wolfram Mathematica®.

**System requirements:** Intel® i9™ or equivalent CPU and 32 GB of RAM

**License:** Attribution—Noncommercial—ShareAlike 4.0 International Creative Commons License

## Appendix III

### A Note about the Program

#### About the program controls

The numerical settings are defined by the user with menus or sliders. Sliders can be finely manipulated by holding down the alt key or opt key while dragging the mouse. They be even more finely manipulated by also holding the *shift* and/or *ctrl* keys.

Dragging with the mouse rotates the three-dimensional plots, while dragging with the mouse while pressing the *ctrl*, *alt*, or *opt* keys zooms in or out.

#### Range of input parameters

$\nu$ : 0.010 – 0.500

$\mu$ : 0.01 – 10000.00

$\sigma$ : 0.01 – 3000.00

$\rho$ : -1.000 – 1.000

$h$ : 0.01 – 2.00

$x$ : 0.01 – 10000.00

$y$ : 0.01 – 10000.00

#### Datasets

d1: measurements of the first measurand in diseased

d2: measurements of the second measurand in diseased

nd1: measurements of the first measurand in nondiseased

nd2: measurements of the second measurand in nondiseased

### References

- Arzideh, Farhad, Werner Wosniok, Eberhard Gurr, Wilhelm Hinsch, Gerhard Schumann, Nicodemo Weinstock, and Rainer Haeckel. 2007. "A Plea for Intra-Laboratory Reference Limits. Part 2. A Bimodal Retrospective Concept for Determining Reference Limits from Intra-Laboratory Databases Demonstrated by Catalytic Activity Concentrations of Enzymes." *Clinical Chemistry and Laboratory Medicine: CCLM / FESCC* 45 (8): 1043–57.
- Berger, James O. 1985. *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media.
- Box, G. E. P., and D. R. Cox. 1964. "An Analysis of Transformations." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 26 (2): 211–43.
- Box, George E. P., and George C. Tiao. 2011. *Bayesian Inference in Statistical Analysis*. Wiley & Sons, Incorporated, John.
- Carlin, Bradley P., and Thomas A. Louis. 2008. *Bayesian Methods for Data Analysis*. CRC Press.
- Chatzimichail, Theodora, and Aristides T. Hatjimihail. 2021. "A Software Tool for Calculating the Uncertainty of Diagnostic Accuracy Measures." *Diagnostics (Basel, Switzerland)* 11 (3). <https://doi.org/10.3390/diagnostics11030406>.
- Choi, Young-Ku, Wesley O. Johnson, and Mark C. Thurmond. 2006. "Diagnosis Using Predictive Probabilities without Cut-Offs." *Statistics in Medicine* 25 (4): 699–717.

- Colberg, Sheri R., Ronald J. Sigal, Bo Fernhall, Judith G. Regensteiner, Bryan J. Blissmer, Richard R. Rubin, Lisa Chasan-Taber, et al. 2010. "Exercise and Type 2 Diabetes: The American College of Sports Medicine and the American Diabetes Association: Joint Position Statement." *Diabetes Care* 33 (12): e147-67.
- D'Agostino, Ralph, and E. S. Pearson. 1973. "Tests for Departure from Normality. Empirical Results for the Distributions of  $b_2$  and  $V b_1$ ." *Biometrika* 60 (3): 613-22.
- Dawid, A. P. 1984. "Present Position and Potential Developments: Some Personal Views: Statistical Theory: The Prequential Approach." *Journal of the Royal Statistical Society. Series A* 147 (2): 278-92.
- Djulgovic, Benjamin, Jef van den Ende, Robert M. Hamm, Thomas Mayrhofer, Iztok Hozo, Stephen G. Pauker, and International Threshold Working Group (ITWG). 2015. "When Is Rational to Order a Diagnostic Test, or Prescribe Treatment: The Threshold Model as an Explanation of Practice Variation." *European Journal of Clinical Investigation* 45 (5): 485-93.
- Dupuis, Josée, Claudia Langenberg, Inga Prokopenko, Richa Saxena, Nicole Soranzo, Anne U. Jackson, Eleanor Wheeler, et al. 2010. "New Genetic Loci Implicated in Fasting Glucose Homeostasis and Their Impact on Type 2 Diabetes Risk." *Nature Genetics* 42 (2): 105-16.
- ElSayed, Nuha A., Grazia Aleppo, Vanita R. Aroda, Raveendhara R. Bannuru, Florence M. Brown, Dennis Bruemmer, Billy S. Collins, et al. 2023. "2. Classification and Diagnosis of Diabetes: Standards of Care in Diabetes-2023." *Diabetes Care* 46 (Suppl 1): S19-40.
- Forbes, Catherine, Merran Evans, Nicholas Hastings, and Brian Peacock. 2011. *Statistical Distributions*. John Wiley & Sons.
- Geer, Eliza B., and Wei Shen. 2009. "Gender Differences in Insulin Resistance, Body Composition, and Energy Balance." *Gender Medicine* 6 Suppl 1 (Suppl 1): 60-75.
- Geisser, Seymour, and Wesley O. Johnson. 2006. *Modes of Parametric Statistical Inference*. John Wiley & Sons.
- Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis*. CRC Press.
- Gramacki, Artur. 2017. *Nonparametric Kernel Density Estimation and Its Computational Aspects*. Springer.
- Haeckel, Rainer, Werner Wosniok, and Farhad Arzideh. 2007. "A Plea for Intra-Laboratory Reference Limits. Part 1. General Considerations and Concepts for Determination." *Clinical Chemistry and Laboratory Medicine: CCLM / FESCC* 45 (8): 1033-42.
- Heckerman, David, Dan Geiger, and David M. Chickering. 1995. "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data." *Machine Learning* 20 (3): 197-243.
- Hyndman, Rob J., and Yanan Fan. 1996. "Sample Quantiles in Statistical Packages." *The American Statistician* 50 (4): 361-65.
- Knuth, Kevin H. 2019. "Optimal Data-Based Binning for Histograms and Histogram-Based Probability Density Models." *Digital Signal Processing* 95 (December): 102581.
- Lehmann, Erich L., and Joseph P. Romano. 2008. *Testing Statistical Hypotheses*. Springer New York.
- McGrayne, Sharon Bertsch. 2011. *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, & Emerged Triumphant from Two Centuries of C*. Yale University Press.
- Meneilly, G. S., and T. Elliott. 1999. "Metabolic Alterations in Middle-Aged and Elderly Obese Patients with Type 2 Diabetes." *Diabetes Care* 22 (1): 112-18.
- Menke, Andy, Keith F. Rust, Peter J. Savage, and Catherine C. Cowie. 2014. "Hemoglobin A1c, Fasting Plasma Glucose, and 2-Hour Plasma Glucose Distributions in U.S. Population Subgroups: NHANES 2005-2010." *Annals of Epidemiology* 24 (2): 83-89.
- National Center for Health Statistics. 2022. "NHANES - NCHS Research Ethics Review Board Approval." Centers for Disease Control and Prevention. August 25, 2022. <https://www.cdc.gov/nchs/nhanes/irba98.htm>.

- . 2005-2016. "National Health and Nutrition Examination Survey Data." Centers for Disease Control and Prevention. <https://wwwn.cdc.gov/nchs/nhanes/default.aspx>.
- . 2005-2016. "National Health and Nutrition Examination Survey Questionnaire." Centers for Disease Control and Prevention. <https://wwwn.cdc.gov/nchs/nhanes/Search/variablelist.aspx?Component=Questionnaire>.
- O'Hagan, Anthony, Caitlin E. Buck, Alireza Daneshkhah, J. Richard Eiser, Paul H. Garthwaite, David J. Jenkinson, Jeremy E. Oakley, and Tim Rakow. 2006. *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons.
- Pandit, M. K., J. Burke, A. B. Gustafson, A. Minocha, and A. N. Peiris. 1993. "Drug-Induced Disorders of Glucose Tolerance." *Annals of Internal Medicine* 118 (7): 529–39.
- Pearl, Judea. 1994. "A Probabilistic Calculus of Actions." In *Uncertainty Proceedings 1994*, edited by Ramon Lopez de Mantaras and David Poole, 454–62. San Francisco (CA): Morgan Kaufmann.
- Salmerón, J., J. E. Manson, M. J. Stampfer, G. A. Colditz, A. L. Wing, and W. C. Willett. 1997. "Dietary Fiber, Glycemic Load, and Risk of Non-Insulin-Dependent Diabetes Mellitus in Women." *JAMA: The Journal of the American Medical Association* 277 (6): 472–77.
- Schoot, Rens van de, Sarah Depaoli, Ruth King, Bianca Kramer, Kaspar Märtens, Mahlet G. Tadesse, Marina Vannucci, et al. 2021. "Bayesian Statistics and Modelling." *Nature Reviews Methods Primers* 1 (1): 1–26.
- Silverman, Bernard W. 1986. *Density Estimation for Statistics and Data Analysis*. CRC Press.
- Smith, A. F. M., and A. E. Gelfand. 1992. "Bayesian Statistics without Tears: A Sampling-Resampling Perspective." *The American Statistician* 46 (2): 84–88.
- Spiegelhalter, David J., Keith R. Abrams, and Jonathan P. Myles. 2004. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley & Sons Australia, Limited, John.
- Surwit, Richard S., Miranda A. L. van Tilburg, Nancy Zucker, Cynthia C. McCaskill, Priti Parekh, Mark N. Feinglos, Christopher L. Edwards, Paula Williams, and James D. Lane. 2002. "Stress Management Improves Long-Term Glycemic Control in Type 2 Diabetes." *Diabetes Care* 25 (1): 30–34.
- Tucker, Larry A. 2020. "Limited Agreement between Classifications of Diabetes and Prediabetes Resulting from the OGTT, Hemoglobin A1c, and Fasting Glucose Tests in 7412 U.S. Adults." *Journal of Clinical Medicine Research* 9 (7). <https://doi.org/10.3390/jcm9072207>.
- Van Cauter, E., K. S. Polonsky, and A. J. Scheen. 1997. "Roles of Circadian Rhythmicity and Sleep in Human Glucose Regulation." *Endocrine Reviews* 18 (5): 716–38.
- Velanovich, V. 1994. "Bayesian Analysis in the Diagnostic Process." *American Journal of Medical Quality: The Official Journal of the American College of Medical Quality* 9 (4): 158–61.
- Viana, M. A. G., and V. Ramakrishnan. 1992. "Bayesian Estimates of Predictive Value and Related Parameters of a Diagnostic Test." *The Canadian Journal of Statistics = Revue Canadienne de Statistique* 20 (3): 311–21.
- Wasserman, Larry. 2006. *All of Nonparametric Statistics*. Springer Science & Business Media.
- Weiner, E. S. C., J. A. Simpson, and Oxford University Press. 1989. *The Oxford English Dictionary*. Oxford: Clarendon Press ; Melbourne.
- Wilk, M. B., and R. Gnanadesikan. 1968. "Probability Plotting Methods for the Analysis of Data." *Biometrika* 55 (1): 1–17.
- Wilkes, Edmund H. 2022. "A Practical Guide to Bayesian Statistics in Laboratory Medicine." *Clinical Chemistry* 68 (7): 893–905.
- Zweig, M. H., and G. Campbell. 1993. "Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine." *Clinical Chemistry* 39 (4): 561–77.

## Permanent Citation:

Chatzimichail T, Hatjimihail AT. Quality Control Using Convolutional Neural Networks Applied to Samples of Very Small Size. Technical Report XXV. Drama: Hellenic Complex Systems Laboratory, 2023. Available at: <https://www.hcsl.com/Documents/hcsltr25.pdf>

## License

[Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.](https://creativecommons.org/licenses/by-nc-sa/4.0/)

First Published: September 5, 2023