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 [a], Aristides T. Hatjimihail, MD, PhD [a]

[a] *Hellenic Complex Systems Laboratory*

Abstract
Medical diagnosis is the basis for treatment and management decisions in healthcare. Conventional methods for medical diagnosis commonly use established clinical criteria and fixed numerical thresholds. Such an approach is limited and may fail to capture the intricate relations 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 three distinct modules, each designed to define and compare parametric and nonparametric distributions effectively. The tool analyzes datasets generated from two separate diagnostic tests performed on 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 (NHANES). 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 divide individuals into two principal categories: those afflicted with a specific disease and those not. Notably, the probability distributions associated with quantitative diagnostic test outcomes often demonstrate some overlap between the diseased and nondiseased groups. 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 in the diagnostic accuracy of those tests (Chatzimichail and Hatjimihail 2021). This dichotomous method represents a significant shift in medical decision-making by linking a continuum of evidence to binary clinical decisions such as to treat or not to treat (Djulbegovic et al. 2015).

Despite the efficiency of traditional diagnostic methods, they sometimes fail to capture the complexity and heterogeneity of disease presentations across diverse 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; Bours 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 a 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). Furthermore, the Bayesian approach extends its utility beyond the medical field by offering a robust framework for quantifying uncertainty in various domains, enriching its applicability in diagnostic and prognostic contexts (Martin et al. 2023; Liu, Liu, and Wong 2013).

A considerable challenge in integrating Bayesian inference into medical diagnosis is the limited literature on diagnostic variables' statistical distributions in 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, critically evaluating its appropriateness for specific diagnostic measurands is necessary. This evaluation should be accompanied by an openness to adopt alternative statistical distributions when needed (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 *Bayesian Diagnosis*, an interactive software tool programmed in the Wolfram Language. This tool allows users to explore and compare parametric and nonparametric distributions to calculate posterior probabilities for disease. It is designed to analyze datasets of measurements of two distinct diagnostic tests performed on diseased and nondiseased populations.

Methods

The Program

Bayesian Diagnosis was developed using Wolfram Mathematica® Ver. 13.3[1]. 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).

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 diagnostic tests and diseases.

Computational Methods

Bayesian Diagnostic Approach

The Bayesian diagnostic approach is a cornerstone of statistical inference and is 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 dataset data comes from a population that can be adequately modeled by a probability distribution with 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

[1] Wolfram Research, Inc., Mathematica, Version 13.3, Champaign, IL (2023).

    3.2. Bivariate
4. *Copula Distributions*

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

The probability density functions (PDFs) 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 Appendix I shows.

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 a better representation of the data's underlying distribution.

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

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. Prior probability of disease 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 μ and standard deviation σ of the measurand in the respective population.

3. *Correlation Coefficients*: The user specifies the correlation coefficients ρ between the measurands of the first and second diagnostic tests for diseased and nondiseased populations.

Kernel Density Estimators

Alternatively, the user can opt to define the KDE for the measurands 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 ρ between the measurands 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 measurand and their combination.
2. *PDF*: Univariate PDF for each measurand and bivariate PDF for 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 its 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. The correlation coefficients are calculated and displayed for each bivariate distribution of the two measurands in diseased and nondiseased populations.
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 measurand 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 measurands.

Illustrative Application

To demonstrate the program's application, 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 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 (ElSayed et al. 2023) during an OGTT (2-h PG). Notably, the study population was confined to individuals aged between 40 and 60 years, a decision informed by the well-documented strong correlation between age and the prevalence of diabetes (Sun et al. 2022).

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 40 – 60 years (n = 11,782)
2. Valid fasting plasma glucose (FPG), glycated hemoglobin (HbA1c), and oral glucose tolerance test (OGTT) results (n=4,015)
3. A negative response to NHANES question DIQ010 regarding a diabetes diagnosis (National Center for Health Statistics 2005-20016) (n=3,854)
4. Non-pregnancy status (n=3,854)

Participants with a 2-h PG measurement of ≥ 200 mg/dl were considered diabetic (n = 211).

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 were used 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

$$v = \frac{211}{3{,}854} = 0.055.$$

The statistics of the dataset are presented in Table 1.

	Diabetic Patients		Nondiabetic Patients	
n	687		10519	
Measurand (Units)	FPG (mg/dl)	HbA1c (%)	FPG (mg/dl)	HbA1c (%)
Mean	141.3	6.67	99.9	5.47
Median	124.0	6.30	99.0	5.50
Standard Deviation	54.0	1.57	10.1	0.38
Skewness	2.375	2.201	0.576	- 0.058
Kurtosis	9.037	8.377	4.213	3.615
Correlation Coefficient	0.914		0.320	

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

Results

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

	Diabetic Patients		Nondiabetic Patients	
Measurand (Units)	FPG (mg/dl)	HbA1c (%)	FPG (mg/dl)	HbA1c (%)
Parametric Distribution	Lognormal	Lognormal	Lognormal	Lognormal
Parametric Distribution Mean	141.3	6.67	99.9	5.47
Parametric Distribution SD	54.0	1.57	10.1	0.38
KDE Smoothing Bandwidth (SD units)	0.32	0.34	0.34	0.35
Correlation Coefficient	0.914		0.320	

Table 2: The settings of the program for Figures 1-13.

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

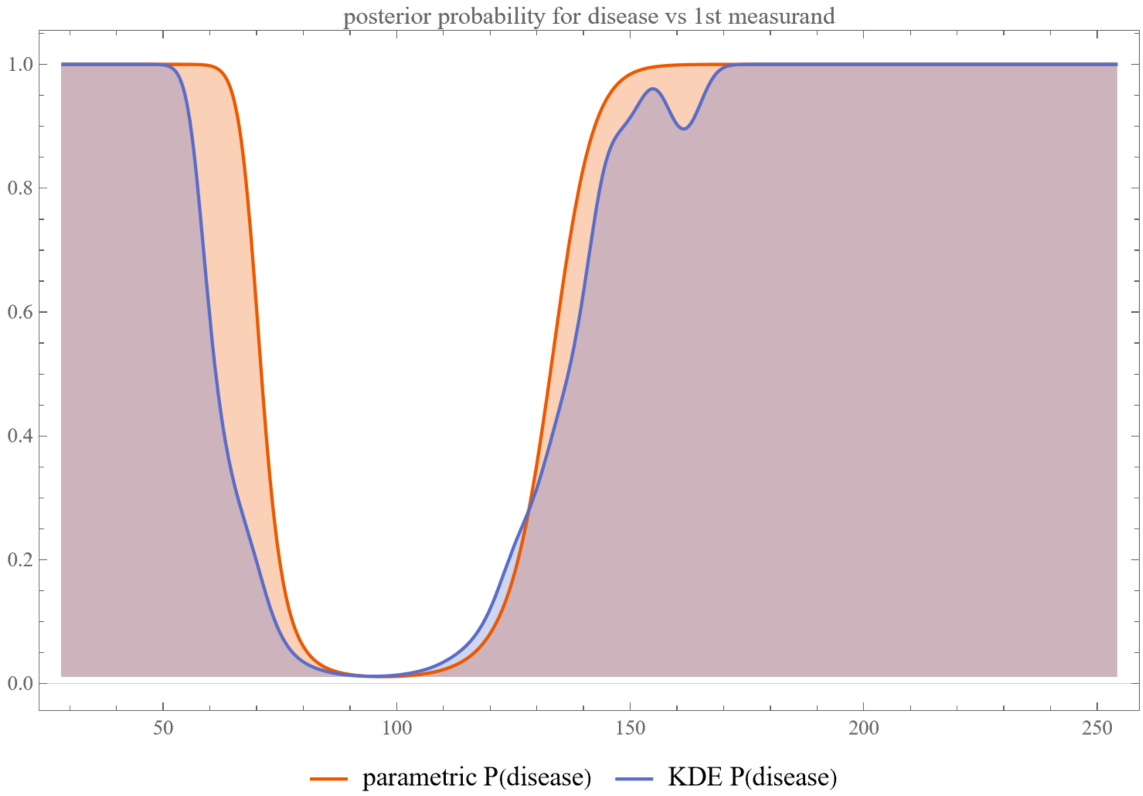

Figure 1: Posterior probability of disease (diabetes) versus the first measurand (FPG), assuming parametric and KDE distributions of the measurand, with the program's settings in Table 2.

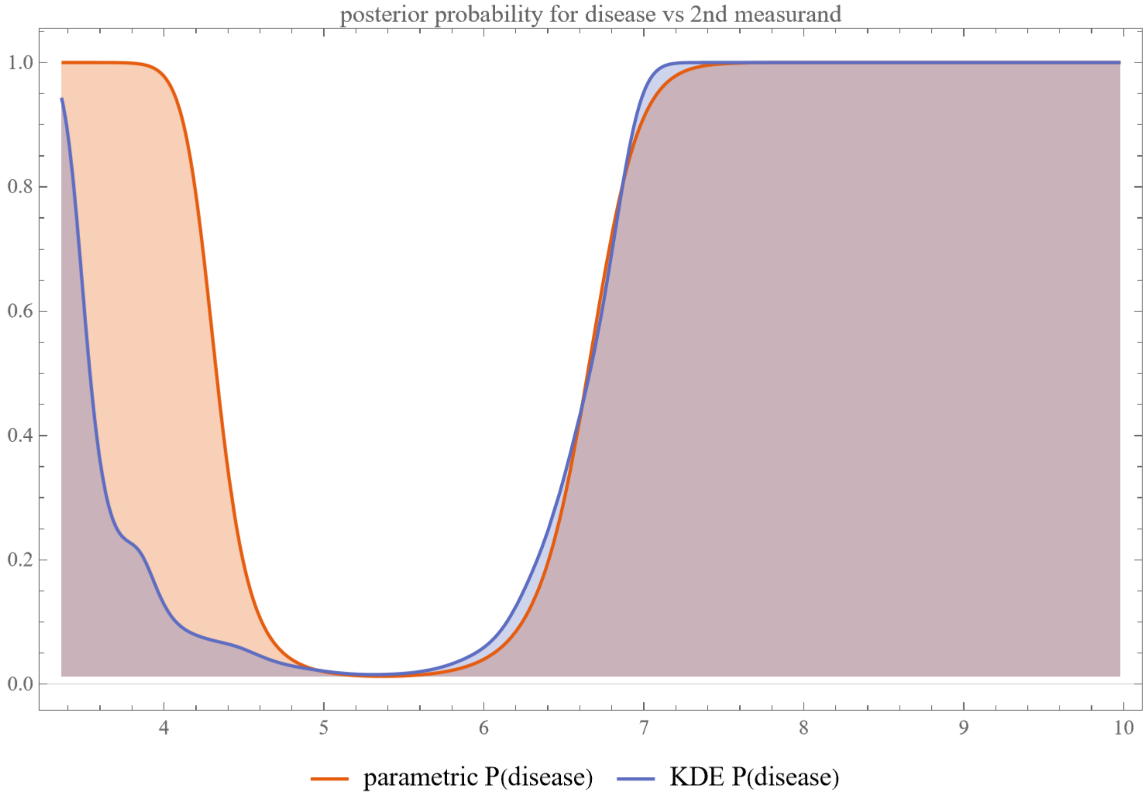

Figure 2: Posterior probability of disease (diabetes) versus the second measurand (HbA1c), assuming parametric and KDE distributions of the measurand, with the program's settings in Table 2.

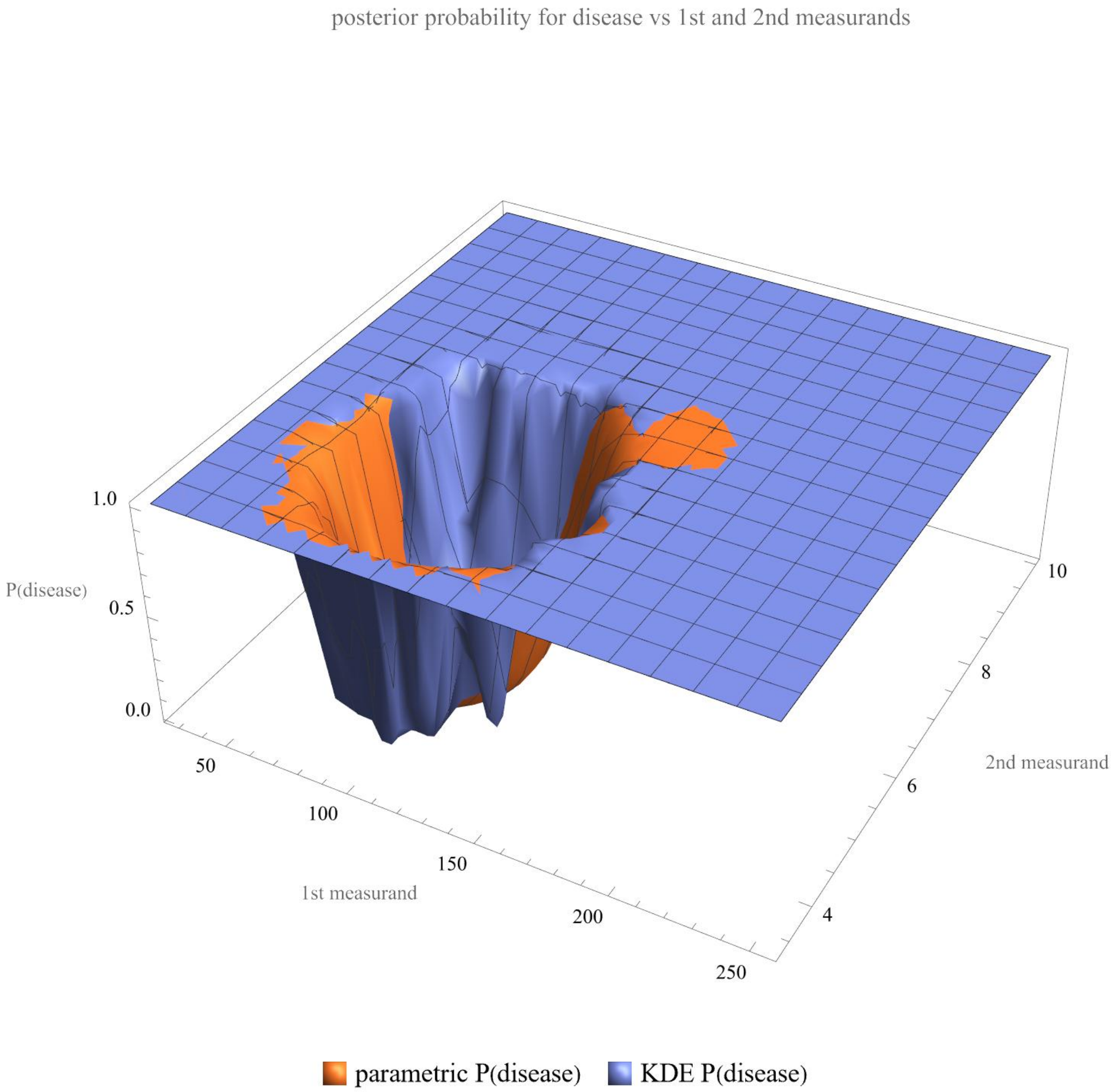

Figure 3: Posterior probability of disease (diabetes) versus both measurands (FPG and HbA1c), assuming parametric and KDE distributions of the measurands, with the program's settings in Table 2.

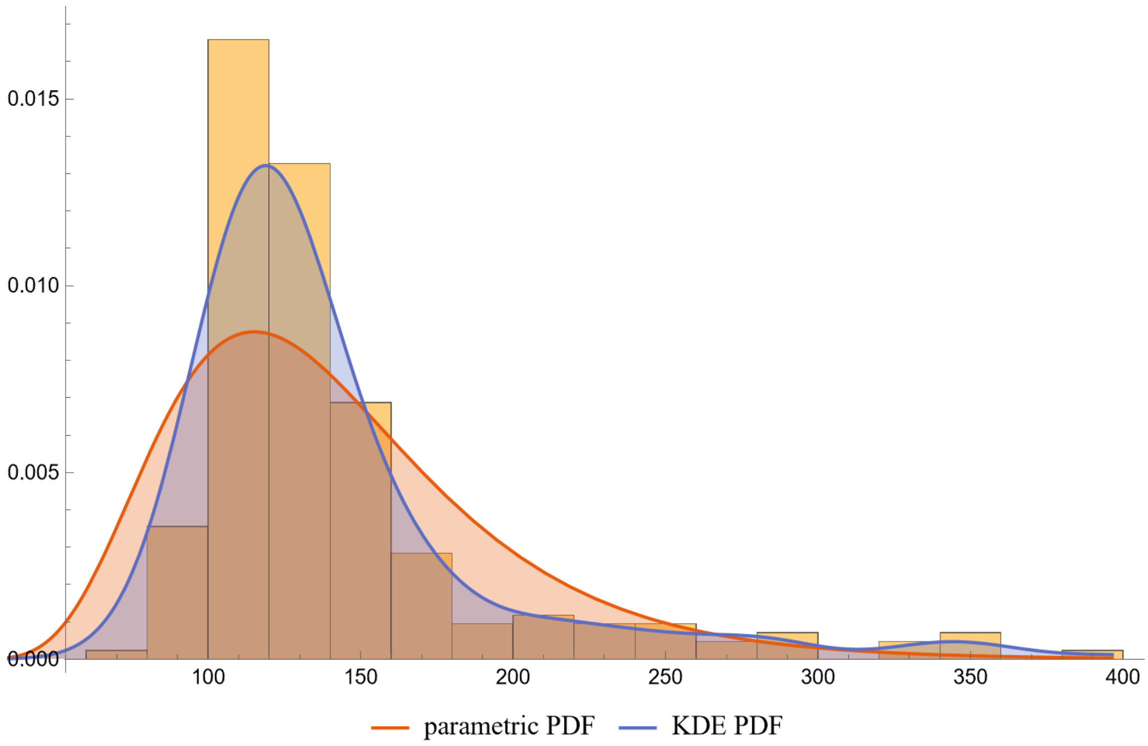

Figure 4: The PDF of the first measurand (FPG) in diseased (diabetic patients), assuming parametric and KDE distributions of the measurand, and the histogram of the respective dataset (NHANES dataset), with the settings of the program in Table 2.

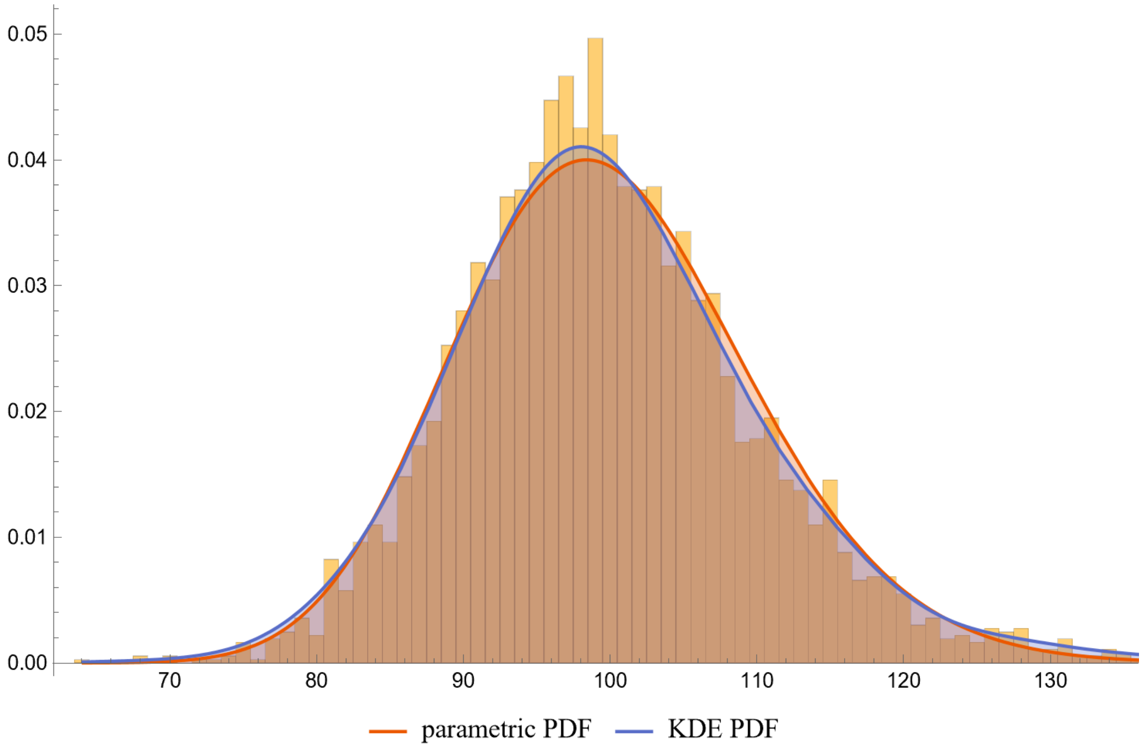

Figure 5: The PDF of the first measurand (FPG) in nondiseased (nondiabetic patients), assuming parametric and KDE distributions of the measurand, and the histogram of the respective dataset (NHANES dataset), with the settings of the program in Table 2.

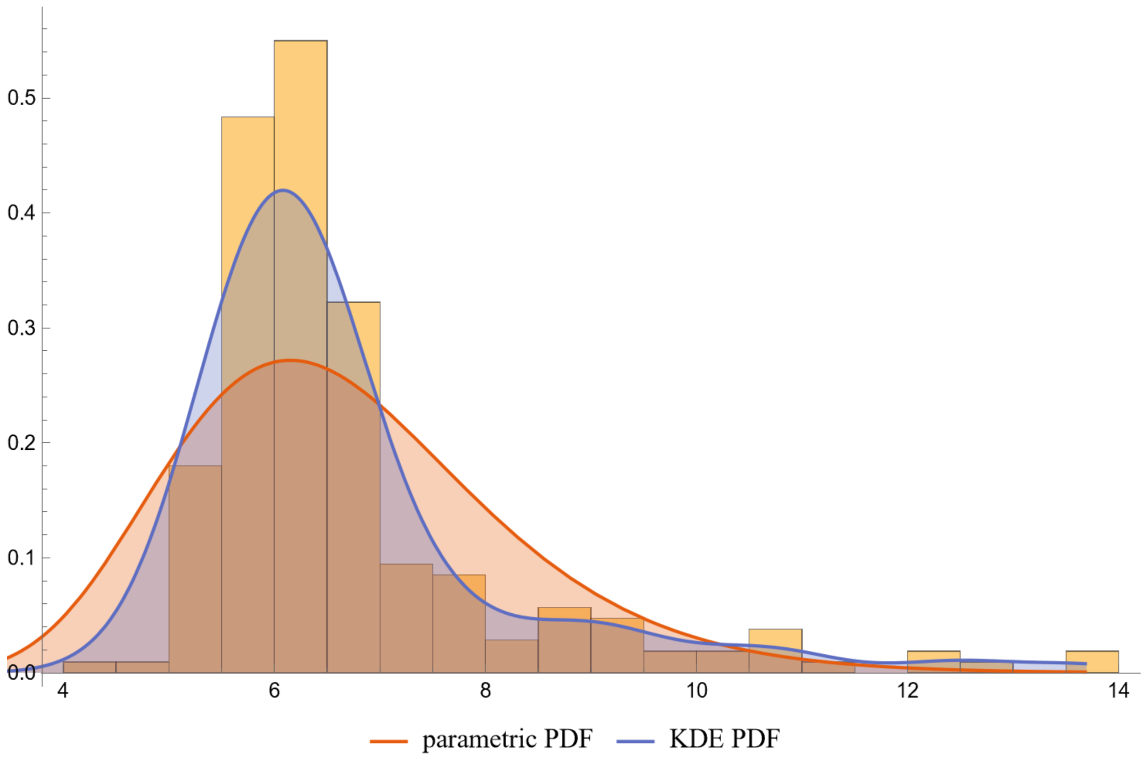

Figure 6: The PDF of the second measurand (HbA1c) in diseased (diabetic patients), assuming parametric and KDE distributions of the measurand, and the histogram of the respective dataset (NHANES dataset), with the settings of the program in Table 2.

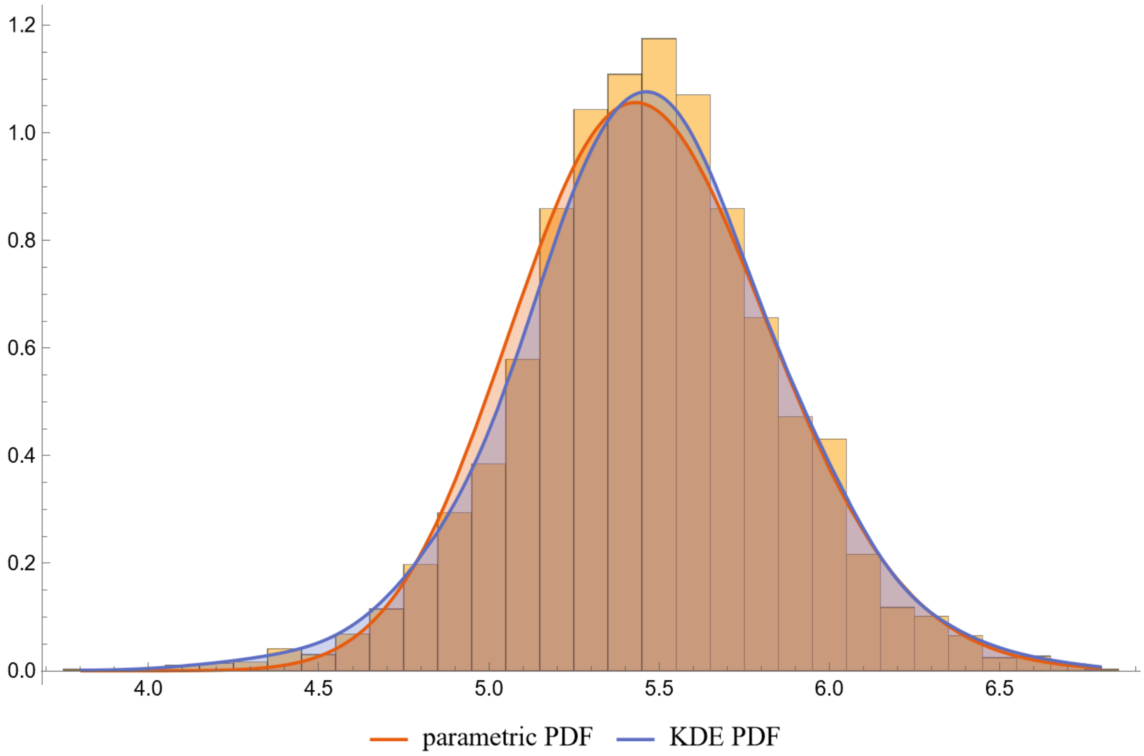

Figure 7: The PDF of the second measurand (HbA1c) in nondiseased (nondiabetic patients), assuming parametric and KDE distributions of the measurand, and the histogram of the respective dataset (NHANES dataset), with the settings of the program in Table 2.

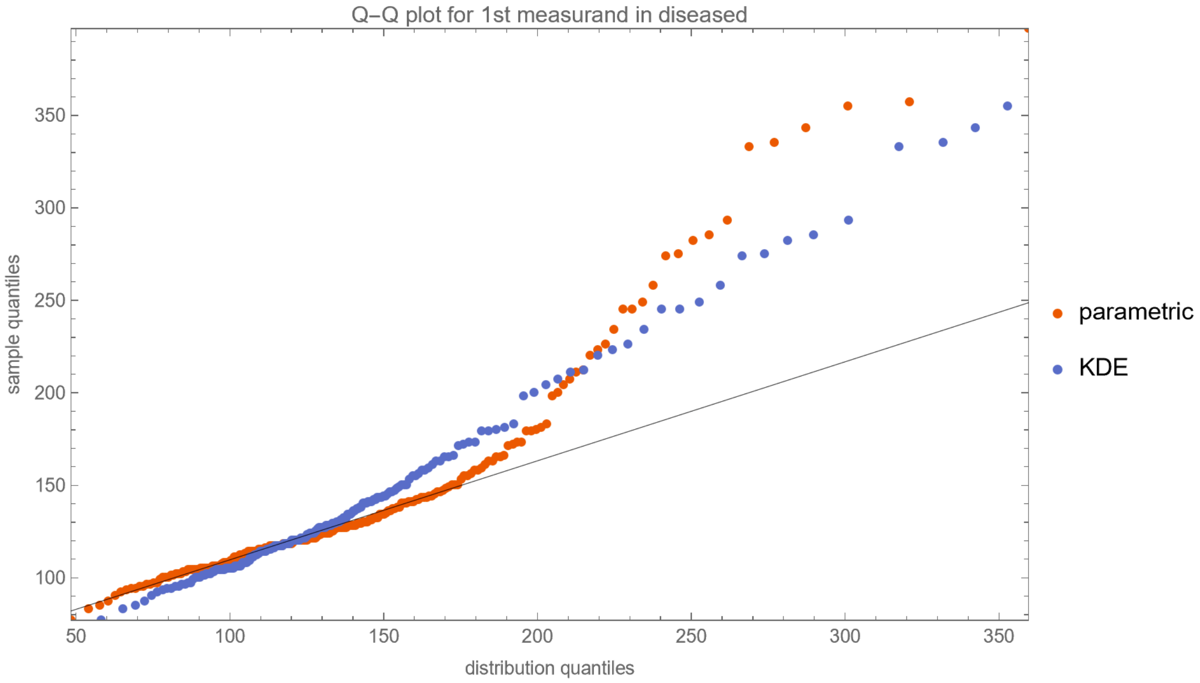

Figure 8: The Q-Q plot of the first measurand (FPG) in diseased (diabetic patients) versus the respective dataset (NHANES dataset), assuming parametric and KDE distributions of the measurand, with the settings of the program in Table 2.

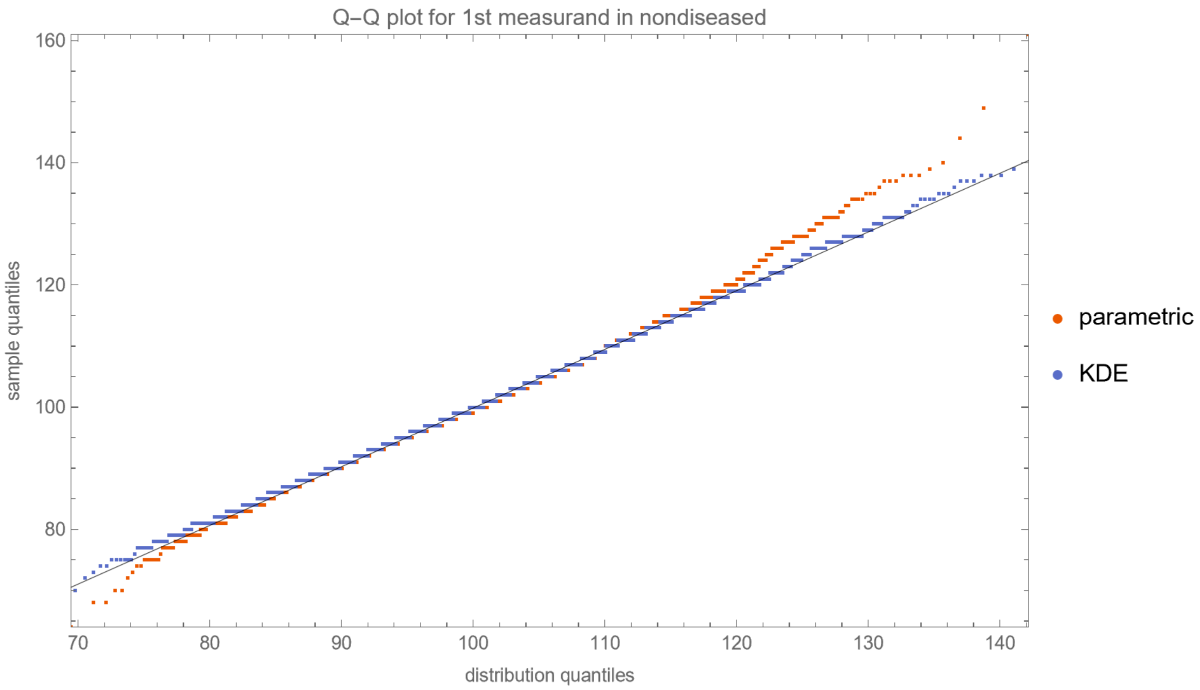

Figure 9: The Q-Q plot of the first measurand (FPG) in nondiseased (nondiabetic patients) versus the respective dataset (NHANES dataset), assuming parametric and KDE distributions of the measurand, with the settings of the program in Table 2.

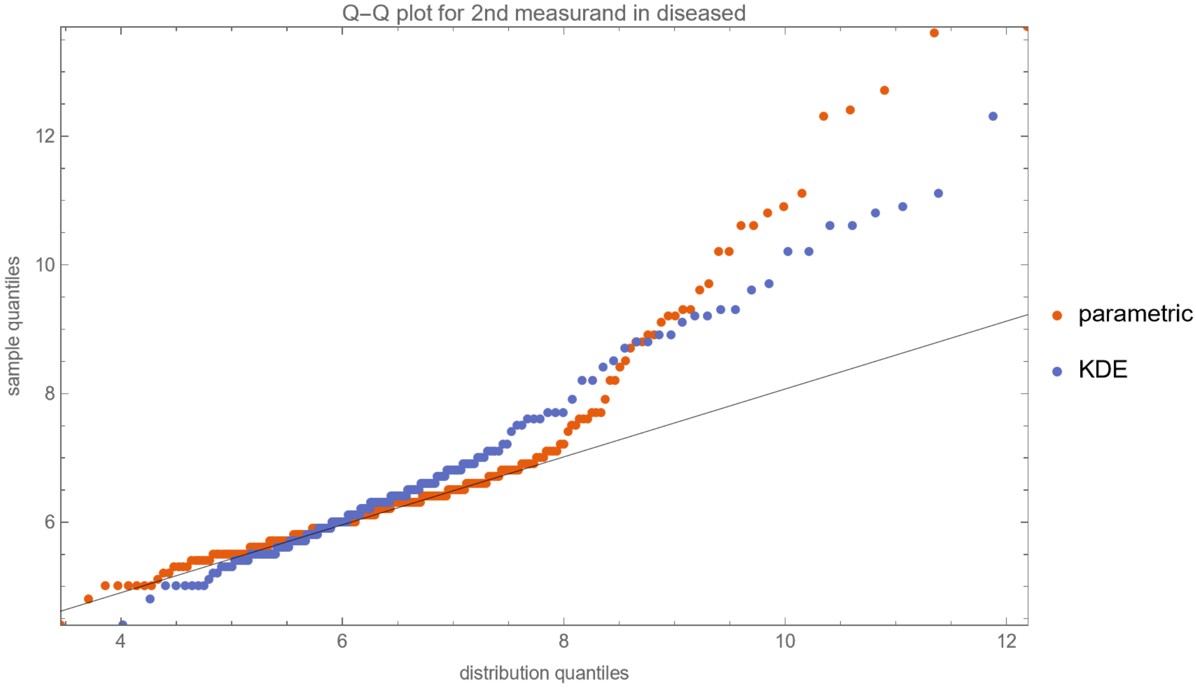

Figure 10: The Q-Q plot of the second measurand (HbA1c) in diseased (diabetic patients) versus the respective dataset (NHANES dataset), assuming parametric and KDE distributions of the measurand, with the settings of the program in Table 2.

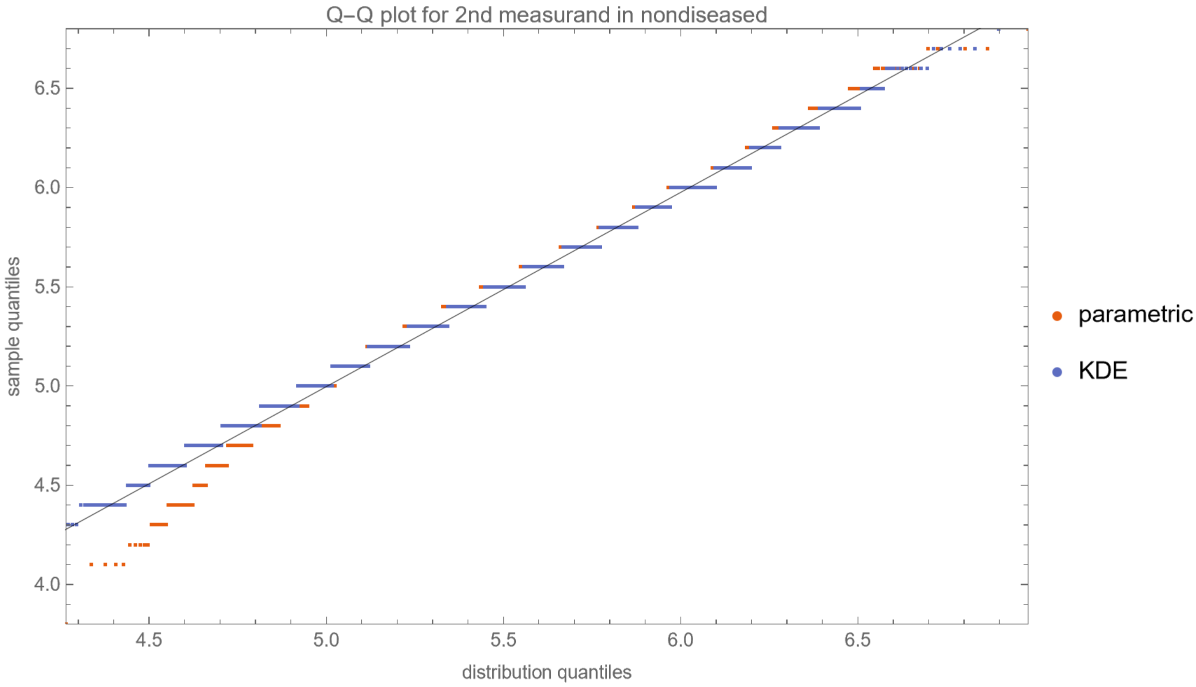

Figure 11: The Q-Q plot of the second measurand (HbA1c) in nondiseased (nondiabetic patients) versus the respective dataset (NHANES dataset), assuming parametric and KDE distributions of the measurand, with the settings of the program in Table 2.

measurements statistics							
		diseased			nondiseased		
	parameters	parametric	KDE	dataset	parametric	KDE	dataset
1st measurand	mean	141.3000	141.2512	141.2512	99.9000	99.8636	99.8636
	median	131.9898	127.2044	124.0000	99.3933	99.1482	99.0000
	sd	54.0000	56.5278	53.9507	10.1000	10.6961	10.1265
	skewness	1.2023	2.0498	2.3746	0.3043	0.4888	0.5763
	kurtosis	5.6759	7.9620	9.0374	3.1651	3.9743	4.2133
	log-likelihood	-1086.9440	-1045.4940		-13537.3200	-13518.6800	
	prior probability	0.055	0.055	0.055	0.945	0.945	0.945
	correlation coefficient	0.9140	0.8141	0.9139	0.3200	0.2791	0.3204
2nd measurand							
	mean	6.6700	6.6697	6.6697	5.4700	5.4705	5.4705
	median	6.4926	6.2863	6.3000	5.4568	5.4693	5.5000
	sd	1.5700	1.6552	1.5704	0.3800	0.3996	0.3772
	skewness	0.7192	1.8660	2.2009	0.2087	-0.0489	-0.0582
	kurtosis	3.9336	7.3146	8.3766	3.0776	3.4876	3.6147
	log-likelihood	-359.0682	-313.1941		-1643.1330	-1599.1410	
	prior probability	0.055	0.055	0.055	0.945	0.945	0.945

Figure 12: The descriptive statistics of the distributions of the measurands (FPG and HbA1c) in diseased (diabetic patients) and nondiseased (nondiabetic patients), assuming parametric and KDE distributions, and of the respective datasets (NHANES datasets), with the settings of the program in Table 2.

prior probability for disease			
prevalence	0.055		
posterior probability for disease			
	parametric	KDE	difference
1st measurand = 126	0.200471	0.238737	0.0382659
2nd measurand = 6.5	0.294458	0.331717	0.0372596
1st measurand = 126 & 2nd measurand = 6.5	0.480279	0.344877	-0.135402

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 versus 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 versus 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 diabetic and nondiabetic patients 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 the Q-Q plots of the PDF of FPG and HbA1c in diabetic and nondiabetic patients versus 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 diabetic and nondiabetic patients, assuming parametric and KDE distributions and 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 present study highlights the importance of incorporating Bayesian methods in medical diagnosis and management. Conventional approaches based on rigid diagnostic criteria often cannot account for the intricate relationships between disease pathology and diagnostic procedures and, therefore, offer a personalized patient approach (Obermeyer and Emanuel 2016). 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 (Topol 2014).

Even though the KDEs 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 classification criteria 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, addressing the intrinsic challenges associated with this methodological shift is paramount. One such issue resides in the limited availability of scholarly publications comprehensively exploring the measurands in 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 calculating 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*: Introducing 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 diagnosis (McGrayne 2011).
5. *Exploration of Alternative Methodologies*: Given the lack of comprehensive data, the utility of combining Bayesian methods with other statistical and computational techniques or diagnostic modalities becomes increasingly pertinent (George E. P. Box and Tiao 2011; Tamrakar, Choubey, and Choubey 2023).

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*: 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 contribute 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 diagnostic statistics and medical decision-making.

Interpretation

1. *Two Zones of Risk*: Such a curve suggests that the risk of the disease is heightened 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 elevated levels of 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 additional tests or measurements could distinguish different sub-populations or disease subtypes better.

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 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 physiological and environmental factors, 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 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 directly influence 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 due to 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, influencing 2-h PG test outcomes (Pandit et al. 1993).
    1.7. Genetic Factors
    Genetic predispositions influence glucose tolerance, introducing 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 of 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).

Related Statistical Software

All major general or Bayesian statistical software packages (BUGS, JASP®, Matlab®, NCSS®, R, SAS®, SPSS®, Stan, and Stata®) include routines for Bayesian inference. The program presented in this work provides 29 different types of parametric and nonparametric plots. None of the above programs provide this range of plots without advanced statistical programming.

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 helpful tool in 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 and extend its application to include more diagnostic modalities. The aim is to combine this approach with existing clinical protocols, optimizing diagnostic precision and 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 is an invaluable resource for 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, accelerating the adoption and dissemination of Bayesian approaches in medical statistics and diagnostics.

Acknowledgments

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 Files

1. BayesianDiagnosis.nb: The program as a Wolfram Mathematica Notebook. Available at https://www.hcsl.com/Tools/BayesianDiagnosis/BayesianDiagnosis.nb
2. BayesianDiagnosisInterface.pdf: A brief description of the program interface. Available at: https://www.hcsl.com/Documents/BayesianDiagnosisInterface.pdf

Author Contributions

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

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 on 25/07/2023).

Conflicts of Interest

The authors declare no conflict of interest.

Appendix I

Formalisms and Notation

Abbreviations

PDF: probability density function

CDF: cumulative distribution function

KDE: kernel density estimator

OGTT: oral glucose tolerance test

PG: plasma glucose

2-h PG: plasma glucose, measured two hours after oral administration of 75 g of glucose, during an OGTT

FPG: fasting plasma glucose

HbA1c: glycated hemoglobin A1c

NHANES: National Health and Nutrition Examination Survey

Tuples

$\mathbf{x}$: an n-tuple $(x_1, x_2, \dots, x_n)$

Parameters

v: prevalence of disease

μ, m: mean

σ, s: standard deviation

ρ, r: correlation coefficient

k: shape parameter

ϑ : 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

$\mathcal{L}(\boldsymbol{\theta}|z)$: likelihood function of the parameter $\boldsymbol{\theta}$ given the observed value z of the random variable Z

$\mathcal{L}(\boldsymbol{\theta}|\mathbf{z})$: likelihood function of the parameter $\boldsymbol{\theta}$ given the observed values $\mathbf{z}$ of the random variable Z

$l(\boldsymbol{\theta}|z)$: loglikelihood function of the parameter $\boldsymbol{\theta}$ given the observed value z of the random variable Z

$l(\boldsymbol{\theta}|\mathbf{z})$: loglikelihood function of the parameter $\boldsymbol{\theta}$ given the observed values $\mathbf{z}$ of the random variable Z

$p(x)$: probability mass function of a discrete variable X

$P_Q(k; q)$: 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

$f(x|\boldsymbol{\theta})$: univariate PDF with parameter vector $\boldsymbol{\theta}$

$f(x, y)$: bivariate PDF

$f(x; \boldsymbol{\theta}), f(x|\boldsymbol{\theta})$: univariate PDF given the multivariate parameter $\boldsymbol{\theta}$

$f(x,y)$: bivariate PDF

$f(x,y;\boldsymbol{\theta}), f(x,y|\boldsymbol{\theta})$: bivariate PDF given the multivariate parameter $\boldsymbol{\theta}$

$F(x)$: univariate CDF

$F(x;\boldsymbol{\theta}), F(x|\boldsymbol{\theta})$: univariate CDF given the multivariate parameter $\boldsymbol{\theta}$

$F(x,y)$: bivariate CDF

$F(x,y;\boldsymbol{\theta}), F(x,y|\boldsymbol{\theta})$: bivariate CDF given the multivariate parameter $\boldsymbol{\theta}$

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:

$$f^{-1}\big(f(z)\big) = z$$

and

$$f\big(f^{-1}(z)\big) = z$$

Natural Logarithm

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

Error Function

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

Complementary Error Function

$$erfc(z) = 1 - erf(z)$$

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:

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

where $P(x \le X < x + \Delta x)$ 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 values x and y. It is denoted as $f_{X,Y}(x, y)$ and defined as:

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

where $P(x \le X < x + \Delta x, y \le Y < y + \Delta y)$ 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:

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

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 μ and σ 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 μ and σ 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 μ_X and μ_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 tuples $(x_1, x_2, \dots, x_n)$ and $(y_1, y_2, \dots, y_n)$, of independent and identically distributed observed values of two random variables X and Y, with means μ_X and μ_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 the possibly multivariate parameter $\boldsymbol{\theta}$ given the value x of the random variable X is defined as:

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

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

The likelihood and loglikelihood functions of a possibly multivariate parameter $\boldsymbol{\theta}$, given a tuple $\mathbf{x} = (x_1, x_2, \dots, x_n)$ of independent and identically distributed observed values of a random variable X, are defined as:

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

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

where $f(x_i;\boldsymbol{\theta})$ 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 tuple of observed values of a random variable 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 tuple of observed values is a numerical value that divides the data into q equal parts, such that exactly $\frac{k}{q}$ of the tuple of observed values or the underlying 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 $F^{-1}(x)$ is the inverse of $F(x)$.

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

Bayes Theorem

For our study, Bayes theorem is formulated as follows:

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

where:

$P(D|T)$ represents the posterior probability of having the disease given a tuple of test results $\boldsymbol{z}$.

$P(T|D)$ denotes the likelihood of obtaining the tuple of test results $\boldsymbol{z}$ given the presence of the disease.

$P(T|\overline{D})$ denotes the likelihood of obtaining the tuple of test results $\boldsymbol{z}$ 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 tuple of test results $\boldsymbol{z}$.

Therefore, for a parameter vector $\boldsymbol{\theta}$:

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

where $\mathcal{L}_D(\boldsymbol{\theta}|\boldsymbol{z})$ and $f_D(\boldsymbol{z}|\boldsymbol{\theta})$ denote the likelihood function and the PDF in the presence of the disease, while $\mathcal{L}_{\bar{D}}(\boldsymbol{z}|\boldsymbol{\theta})$ and $f_{\bar{D}}(\boldsymbol{z};\ \boldsymbol{\theta})$ 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 tuple against each other. If the tuple comes from the theoretical distribution, the points in the Q-Q plot will approximately lie on the reference line $y = x$.

P-P plot

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

Parametric Distributions

Normal Distribution

Univariate

The univariate normal distribution or Gaussian distribution is a continuous probability distribution of a 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 μ is the mean of X, while σ is its standard deviation, and $-\infty < x < \infty$ (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 μ_X and μ_Y are the means of X and Y, σ_X and σ_Y are their standard deviations, ρ their correlation coefficient, , $-\infty < x < \infty$, and $-\infty < y < \infty$ (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:

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

where m is the mean, s the standard deviation of $ln(X)$, and $0 < x < \infty$ (Forbes et al. 2011).

If μ and σ 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,

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

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

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

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 their standard deviations, and r their correlation coefficient, for $0 < x < \infty$, and $0 < y < \infty$, 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

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

$$b = (ln(y)-m_Y){s_X}^2 - r(ln(x)-m_X)s_Xs_Y$$

$$c = (ln(x)-m_X){s_Y}^2 - r(ln(y)-m_Y)s_Xs_Y$$

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

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 μ_X and μ_Y are the means of X and Y, σ_X and σ_Y are their standard deviations and ρ 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, ϑ a scale parameter , $\Gamma(u)$ the gamma function and $0 < x < \infty$ (Forbes et al. 2011).

The mean μ and the standard deviation σ of X, are calculated as follows:

$$\mu = k\vartheta$$

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

Therefore,

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

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

and

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

Bivariate

The bivariate Gamma distribution is a continuous probability distribution of two variables X and Y. For $0 < x < \infty$ and $0 < y < \infty$, the copula version of its PDF is:

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

where

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

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

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

and where k_X, k_Y are shape parameters, ϑ_X, ϑ_Y are scale parameters and ρ the correlation coefficient of X and Y.

If μ_X and μ_Y are the means of X and Y, σ_X and σ_Y their standard deviations, and ρ their correlation coefficient, it can be shown that:

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

where

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

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

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

Copulas

If μ_X and μ_Y are the means of the variables $X$and Y, σ_X and σ_Y their standard deviations, and ρ 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

For $-\infty < x < \infty$, and $< y < \infty$,

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

where

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

$$b = -\frac{\left(\frac{2ln(y) - 2ln(\mu_Y) + ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)}{2\sqrt{2}\sqrt{ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)}}\sqrt{ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)}\mu_Y + \rho\sigma_Y erfc^{-1}\left(2Q\left(\frac{\mu_X^2}{\sigma_X^2}, 0, \frac{x\mu_X}{\sigma_X^2}\right)\right)\right)^2}{ln\left(1 + \frac{\sigma_Y^2}{\mu_Y^2}\right)\mu_Y^2 - \rho^2 s\sigma_Y^2}$$

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

$$d = \left(\frac{xm_X}{\sigma_X^2}\right)^{\frac{\mu_X^2}{\sigma_X^2}}$$

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

X: Lognormally Distributed – Y: Normally Distributed

For $0 < x < \infty$, and $-\infty < y < \infty$,

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

where

$$a = erfc^{-1}\left(2Q\left(\frac{\mu_Y^2}{\sigma_Y^2}, 0, \frac{y\mu_Y}{\sigma_Y^2}\right)\right)^2 - \frac{\left(2ln(x) - 2ln(\mu_X) + ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right)\right)^2}{8ln\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right)}$$

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

$$c = a + b - \frac{y\mu_Y}{\sigma_Y^2}$$

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

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

X: Normally Distributed – Y: Gamma Distributed

For $-\infty < x < \infty$, and $0 < y < \infty$,

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

X: Gamma Distributed– Y: Normally Distributed

For $0 < x < \infty$, and $-\infty < y < \infty$,

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

X: Lognormally Distributed – Y: Gamma Distributed

For $0 < x < \infty$, and $0 < y < \infty$,

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

where

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

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

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

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

X: Gamma Distributed – Y: Lognormally Distributed

For $0 < x < \infty$, and $0 < y < \infty$,

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

where

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

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

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

Nonparametric Distributions

Histograms

A histogram is a graphical representation of the distribution of a tuple of observed values of a variable X. If X is a continuous random variable, the histogram is an estimate of the probability distribution X. To construct a histogram:

1. The range of the tuple of the variable's observed values is divided into a set of bins.
2. The variable's observed values are sorted into each bin.
3. The number of variable's observed values that fall into each bin are counted.

The height of each bar in the histogram corresponds to the count of the variable's observed values 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 a histogram's optimal number of bins. It calculates the optimal bin width by maximizing a likelihood function, considering the variable's observed values as independently and identically distributed.

Given a tuple $(x_2, \dots, x_n)$ of observed values of a variable X, we find the optimal bin edges $\mathbf{B} = (b_1, b_2, \dots, b_k)$, by maximizing the following likelihood function:

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

where n is the total number of observed values, k is the number of bins, n_i is the number of observed values 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 tuple of independent and identically distributed observed values $(x_1, x_2, \dots, x_n)$ of a random variable X, the univariate KDE $\hat{f}_K(x; n, h)$ is defined as (Gramacki 2017):

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

where:

1. n is the number of the observed values of the variable,
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. $K(u)$ is the kernel function, which satisfies the properties:
   3.1. $\int K(u)du = 1$
   3.2. $\int u^2 K(u)du < \infty$

Given two tuples of independent and identically distributed observed values $(x_1, x_2, \dots, x_n)$ and $(y_1, y_2, \dots, y_n)$ of two random variables X and Y, the bivariate KDE $\hat{f}(x, y; n, h_1, h_2)$ is defined as (Gramacki 2017):

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

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 ρ is the correlation coefficient of the two tuples of data points.

A kernel function $K(u)$ could be conceptualized as a weighting mechanism in the context of kernel density estimation. For every observed value 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 6 September 2023)

Operating systems: Microsoft Windows, Linux, Apple macOS and 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 user defines the numerical settings with menus or sliders. Sliders can be finely manipulated by holding down the alt key or opt key while dragging the mouse. They can 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; pressing the *ctrl*, *alt*, or *opt* keys zooms in or out.

Range of input parameters

ν: 0.010 – 0.500

μ: 0.01 – 10000.00

σ : 0.01 – 3000.00

ρ : –1.000 – 1.000

h : 0.01 – 2.00

x : 0.01 – 10000.00

y : 0.01 – 10000.00

Datasets

The software is preloaded with the following datasets:

d1: Quantitative measurements of the first measurand (FPG) from diseased individuals (diabetic patients), aged 40-60.

d2: Quantitative measurements of the second measurand (HbA1c) from diseased individuals (diabetic patients), aged 40-60.

nd1: Quantitative measurements of the first measurand (FPG) from nondiseased individuals (nondiabetic patients), aged 40-60.

nd2: Quantitative measurements of the second measurand (HbA1c) from nondiseased individuals (nondiabetic patients), aged 40-60.

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.

Bours, Martijn Jl. 2021. "Bayes' Rule in Diagnosis." *Journal of Clinical Epidemiology* 131 (March): 158–60.

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 $\sqrt{b_1}$." *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.

Djulbegovic, 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.

Liu, Jialin, Shu Jie Liu, and David Shan Hill Wong. 2013. "Process Fault Diagnosis Based on Bayesian Inference." In *Computer Aided Chemical Engineering*, edited by Andrzej Kraslawski and Ilkka Turunen, 32:751–56. Elsevier.

Martin, Gael M., David T. Frazier, Worapree Maneesoonthorn, Rubén Loaiza-Maya, Florian Huber, Gary Koop, John Maheu, Didier Nibbering, and Anastasios Panagiotelis. 2023. "Bayesian Forecasting in Economics and Finance: A Modern Review." *International Journal of Forecasting*, July. https://doi.org/10.1016/j.ijforecast.2023.05.002.

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 US 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. 25 August, 2022. https://www.cdc.gov/nchs/nhanes/irba98.htm.

———. 2005-20016. "National Health and Nutrition Examination Survey Data." Centers for Disease Control and Prevention. 2005-20016. https://wwwn.cdc.gov/nchs/nhanes/default.aspx.

———. 2005-20016. "National Health and Nutrition Examination Survey Questionnaire." Centers for Disease Control and Prevention. 2005-20016. https://wwwn.cdc.gov/nchs/nhanes/Search/variablelist.aspx?Component=Questionnaire.

Obermeyer, Ziad, and Ezekiel J. Emanuel. 2016. "Predicting the Future - Big Data, Machine Learning, and Clinical Medicine." *The New England Journal of Medicine* 375 (13): 1216–19.

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 Healthcare Evaluation*. Wiley & Sons Australia, Limited, John.

Sun, Hong, Pouya Saeedi, Suvi Karuranga, Moritz Pinkepank, Katherine Ogurtsova, Bruce B. Duncan, Caroline Stein, et al. 2022. "IDF Diabetes Atlas: Global, Regional and Country-Level Diabetes Prevalence Estimates for 2021 and Projections for 2045." *Diabetes Research and Clinical Practice* 183 (January): 109119.

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.

Tamrakar, Sitendra, Shruti Bhargava Choubey, and Abhishek Choubey. 2023. *Computational Intelligence in Medical Decision Making and Diagnosis: Techniques and Applications*. CRC Press.

Topol, Eric J. 2014. "Individualized Medicine from Prewomb to Tomb." *Cell* 157 (1): 241–53.

Tucker, Larry A. 2020. "Limited Agreement between Classifications of Diabetes and Prediabetes Resulting from the OGTT, Hemoglobin A1c, and Fasting Glucose Tests in 7412 US 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 2004. *The Oxford English Dictionary*. Oxford, 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. A Bayesian Inference Based Computational Tool for Parametric and Nonparametric Medical Diagnosis. Technical Report XXV. Drama: Hellenic Complex Systems Laboratory, 2024. https://www.hcsl.com/TR/hcsltr25/hcsltr25.pdf

License

First Published: September 4, 2023

Revised: November 9, 2024