PREDICTIVE PERFORMANCE OF BAYESIAN DIAGNOSES

A dissertation submitted to

THE GRADUATE SCHOOL OF ENGINEERING SCIENCE

OSAKA UNIVERSITY

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN ENGINEERING

BY
NAOKI ISOGAWA

MARCH 2012

謝辞

本稿の作成にあたり，多くの方々にご指導・ご支援を頂戴いたしました．ここに，心よりお礼申し上げます．

指導教員の白旗慎吾先生には，本論文の全体を通して，多大なご教示をいただき，著者の研究に対して貴重なご意見や鋭いご指摘を幾度となく賜りました．大阪大学基礎工学部数理科学コース 4 年生のときに白旗研究室に配属されてからの 6 年間，大変にお世話になりました．心よりお礼を申し上げるとともに，今後もますますのご高配のほど，よろしくお願い申し上げます．大阪大学教授（大学院基礎工学研究科統計数理講座）の狩野 裕先生には，本論文の初稿を査読いただき，公聴会では貴重なご指摘を賜りました．大阪大学教授（大学院基礎工学研究科数理計量ファイナンス講座）の内田雅之先生には，本論文の初稿を査読いただき，公聴会では貴重なご意見を頂戴いたしました。大阪大学准教授の坂本 亘先生には，白旗慎吾先生と同じく，著者が大阪大学基礎工学部数理科学コース 4 年生のときに白旗研究室に配属されてからの 6 年間，大変にお世話になり，著者の研究を熱心に指導していただきました．著者の拙い質問に対しても丁寧に答えていただき，研究に行き詰ったときにも時間を割いて解決策を一緒に考えてくださいました．

本稿の指導をしていただきました特定非営利活動法人 医学統計研究会 (Biostatistical Research Association :BRA) 理事長の後藤昌司先生には，著者が大阪大学基礎工学部数理化学コース 4 回生の卒業間近に BRA の門を叩いてから今までの 5 年間，未熟で至らない学生であった著者に対して，一人の人間として真正面から向き合っていただき，ひとかたならぬご厚情を賜りました．後藤先生には，学問の分野だけでなく，日常の生活からこれからの人生で歩むべき道筋に至るまで大変に貴重なご指導をいただきました．とくに，後藤先生の門下の先輩方の心に刻み込まれている「掃除・勤行・学問」という座右の銘は著者にとって最も重みをもった言葉となっております．日々の「掃除・勤行」を疎かにしていては「学問」だけできていても意味がない，この順序を大切にして人を惹きつけるような「人間的魅力の醸成」に注力しなさい，という後藤先生からのお言葉はこれまでの著者の考え方を一新させるものであり，大変に納得させられました．しかしながら，修士論文の提出時には，自分に対する甘えから優先順位を間違えた著者を厳しく叱責していただき，順序を守ることの大切さをご教示いただきました．また，著者が学生のときから BRA の定例会や夏季・秋季セミナーをはじめ，大分統計談話会，日本計算機統計学会，国際計算機統計学会

（International Association of Statistical Computing :IASC）などの学会や，さまざまな研究会・シンポジウムでの発表の機会を与えていただき，大変に良い経験をさせていただきました．また，その後の「遊」にも何度か同行させていただき,「遊学一如」の精神の真髄の一端を学ぶことができました．社会人になってからも，後藤先生は，著者の怠惰な性格を見抜いておられ，折に触れて，ご自身の「哲学」に基づく物事の核心をついた貴重なお話を通して，著者を叱咤激励していただきました．これらの親身に優るお世話に心より感謝申し上げます．また，本稿の作成にあたって，ご自身の業務日程が過密であるにもかかわらず，丁寧にご校閲・ご指導いただきました．後藤門下生の名に恥じぬよう，今後も日々精進して参ります．本当にありがとうございました．

鹿児島高等専門学校教授の藤崎恒晏先生には，お会いするたびに気さくに声をかけていただきました．とくに秋季セミナー 2007 の遊学では大変にお世話になり，くろず情報館「壺畑」や桜島といった鹿児島県の観光名所を自家用車で案内してくださいました．また，Latex に関する質問をうけた際に，著者の知識不足から拙い説明しかできなかったにも関わらず，高価なお礼の品をいただきました．長崎大学教授の柴田義貞先生には，サマー・フォーラム 2007 で著者の発表に対して鋭いご指摘をいただくとともに著者の研究主題に関する統計的背景を詳しくご教示いただきました．また，大分統計談話会・第 36 回大会にて著者が発表させていただいた内容に関する資料としてご自身の貴重な講義資料を送ってくださいました．大分大学教授の越智義道先生には，大分統計談話会での発表の機会を与えていただき，その度に貴重なご助言を賜りました．弘前大学准教授の杉本知之先生には，シンポジウムなどでのご講演を伺うことで，研究に向かう姿勢など多くのことを学ばせていただきました兵庫県立医科大学講師の大門貴志先生には，直接お会いする機会はあまりありませんでしたが，本稿の作成にあたって大門先生の研究を大いに参考にさせていただきました．山梨大学准教授の下川敏雄先生には，日頃より著者のことを気にかけていただき，研究に関する指導だけでなく最新の文献やアプリケーション・ソフトに関する情報もご提供いただきました．夏季セミナー 2007 の際には，車で山梨県の観光名所を案内してくださり，とても思い出に残る 2 日間となりました．大分県地域成人病検診センター 池邉淑子先生には，健康診断のデータを提供していただき，それらのデータに基づく保健指導に関する研究についてご指導いただきました．また，池邉淑子先生が同センターで実施されている「特定健診・保健指導」に関する詳しい情報をご教示いただきました．上記の諸先生方のご指導・ご意見・励ましのお言葉に心より感謝いたします．

本稿の作成にあたって BRA の皆さまには大変にお世話になりました．臨床研究情報センター（財団法人 先端医療振興財団）の松原義弘博士には BRA の諸会合の折にいつも激励していただきました．また，日本計算機統計学会・第 22 回シンポジウムでの発表の機会を与えていただき，貴

重な経験をさせていただきました．特定非営利活動法人 医学統計研究会 常務理事の魚井 徹さんには，BRA の諸会合でお会いするたびに気さくにお声をかけていただきました．また，発表に対するご意見をうかがうたびに医学統計の奥深さに気づかされるとともに自分の未熟さを痛感いたしました．株式会社 新日本科学の勘場 貢さんには，いつも気さくにお声をかけていただき，ご自身の経験に基づく貴重なご助言を賜りました．また，勘場さんの趣味に関する楽しい話もお聴かせいただきました．株式会社ベルシステム 24 の前田 博さんには，いつも励ましのお言葉をいただきました．また，社会人としての心構えなどについてご教示いただきました．株式会社ベルシステム 24 の後藤浩司さんには，学生のときから株式会社ベルシステム 24「統計科学研究会」にお誘いいただき，著者も発表の機会をいただきました．また，著者の発表に対して貴重なご意見をいただきました．株式会社 富士通大分ソフトウェアラボラトリの衛藤俊寿博士には，著者が大分統計談話会などで大分に訪れたときに大変にお世話になりました．とくに，大分統計談話会・第 39 回大会の後の「大愚の会：修了記念」旅行の折には，宇佐・中津の観光名所を案内していただき，行く先々で観光名所にまつわる楽しいお話を聴かせていただきました．株式会社フィールドワークスの木田義之さんには，BRA の諸会合でお会いするたびに気さくにお声をかけていただきました．先述の宇佐・中津への修了記念旅行の際には，観光名所をご案内いただきました．株式会社ソリューションラボの志賀 功さんには，池邉淑子先生と共同で保健指導に関する研究を行う機会を与えていただきました．池邉淑子先生との共同研究を通じて研究の面白さを実感することができ，多くのことを学びました．また，大分統計談話会・第 36 回大会では，発表後に学生であった著者らをご自身の自家用車で高崎山や別府海浜砂湯に連れて行っていただいたうえに大分名物「とり天」までご馳走になり，労をねぎらっていただきました．さらに，先述の修了記念旅行の際には，本来ならば著者らがすべき旅行行程の立案から別府の宿の手配までしていただき，学生生活の最後に忘れられない思い出を作ることができました．この修了記念旅行では，志賀 功さん，衛藤俊寿博士を始め，後藤昌司先生，藤崎恒晏先生，魚井 徹さん，柴田義貞先生ご夫妻，木田義之さん，河合統介博士，藤澤正樹博士，伊藤雅憲博士，丸尾和司博士，大江基樹さん，中村将俊さん，山口祐介さんの方々に大変にお世話になりました．ここに厚くお礼申し上げます．

大愚の会の先輩方には大変にお世話になりました．小野薬品工業株式会社の冨金原 悟博士には，BRA の諸会合の折にいつも激励していただきました．IASC2008 では，はじめての学会で緊張する著者を気遣い，温かい言葉をかけていただきました．IASC2008 の開催期間中に何度も食事に誘っていただき，豪華な焼肉や中華料理をご馳走になりました．あすか製薬株式会社の藤澤正樹博士には,「世のため，人のため」に働くことの大切さをご自身の身を以ってご教示いただきました．業務が多忙であるにも関わらず，日々の「勤行」を率先して行う姿に感銘を受けました．未熟

な著者を温かく見守っていただき，ときにはご自身の経験談を交えて叱咤激励していただきました．エーザイ株式会社の高瀬貴夫さんには，いつも気遣っていただき，優しい言葉で励ましていただきました。アステラス製薬株式会社の伊藤雅憲博士さんには，お会いするたびに気さくに声をかけていただきました．社会経験の乏しい著者にとって「遊」も「学」も全力で全うする伊藤さんの豪快さに強い魅力を感じました．ファイザー株式会社の山邉太陽さんには，会社の先輩でもあり，学位取得という同じ目標に向かって研究を続ける中で，著者が落ち込んでいたときにも励ましていただき，前向きな言葉で勇気を分けてくださいました．協和発酵キリン株式会社の古川泰伸さんには，学生時代に参加した日本計算機統計学会・第 22 回シンポジウムのときには食事をご馳走になり，発表を翌日に控えた著者を励ましていきただき，ご自身の学生時代の経験を教えていただきました．ノバルティスファーマ株式会社の池田公俊さんには，大分統計談話会・第 37 回大会で貴重な助言を賜りました．また，古川さんと同じく日本計算機統計学会・第 22 回シンポジウムのときには食事をご馳走になり，発表を翌日に控えた著者を励ましてくれました．ファイザー株式会社の弘 新太郎博士は，著者が学生のときから研究に対する姿勢をご教示いただき，入社後も統計に対する考えについて一緒に議論をさせていただく中で，社会人として自立することの大切さを教えていただきました．第一アスビオファーマ株式会社の永久保太志博士は，著者と年齢が比較的近いこともあり，ざっくばらんな話を通じて温かく接していただきました．株式会社ベルシステム 24 の金 水龍さんには，学生時代に株式会社ベルシステム 24「統計科学研究会」にお誘いいただき，著者にも発表の機会を与えていただきました．その後の懇親会にもご一緒させていただき，おいしい食事をご馳走になりました．グリー株式会社の元垣内広毅さんには，BRA に入るきっかけをつくってくださり，大変にお世話になりました．著者が学部 4 回生のときに白旗研のセミナーに参加していただき，著者ら学生を BRA に勧誘くれました．そして，著者が BRA に興味を示したときには焼肉を馳走していただき，BRA に関する話や BRA で得た経験についてこと細かに語ってくれました．また，著者に「時間を守ることの大切さ」をご教示いただきました．興和株式会社の丸尾和司博士には，常日頃より著者のことを気遣っていただき，さまざまなご支援をいただきました．元垣内さんと一緒に白旗研のセミナーで著者ら学生を BRA に勧誘していただき，興味を示した著者に焼肉屋で，夜遅くまで BRA で学ばれたことについて語ってくれました．その後も折に触れて，著者の研究に対して励ましのお言葉をいただくとともに幾度となく貴重なご助言を賜りました．また，社会人になってからも，共同研究をさせていただいていることもあり，著者の拙い質問に対しても丁寧にご対応いただき，大変に勉強になりました．一番身近な先輩として著者を支えてくださったことに心より感謝いたしております．

株式会社 大塚製薬工場の大江基貴さんには，白旗研究室に配属されてから，BRA の一員とし

てともに過ごした大学院の 2 年間を含めて大変にお世話になりました．年長者として年下である著者の相談に何度ものっていただき，温かい言葉で励ましていただきました．お互いに支え合って，刺激し合ってきたからこそ，自分も少しは成長できたと思います．本当にありがとうございました．大日本住友製薬株式会社の中村将俊さんには，BRA の一員としてともに過ごした大学院の 2 年間を含めて大変にお世話になりました．いつも明るく著者を温かい言葉で勇気づけてくれたおかげで充実した学生時代の 2 年間を過ごすことができました．また，何があっても前向きな気持ちで突き進む姿に多くのことを学びました．大江さん・中村さんとともに後藤先生のもとで学ぶことができた経験は私の貴重な財産です．本当にありがとうございました．大阪大学大学院博士後期課程 2 年の山口祐介さんには，本稿の作成にあたり，文献を送っていただいたり，図表の修正等を手伝っていただいたりと大変にお世話になりました．また，学生時代から勤行を率先して引き受けて下さり，色々と助けていただきました．大阪大学大学院博士前期課程 2 年の大山秀輔さん，横山隼人さん，吉川隆範さんには，著者が帰阪した際にお世話いただきました。これらの方々に心より感謝いたします．

後藤昌司先生の奥様の後藤 孚様には，いつも著者らの健康を気遣っていただき，学生時代には一人暮らしの学生の昼食とは考えられないような豪華で栄養バランスのとれた美味しいお弁当を幾度となく頂戴いたしました．オフィスで頂くお弁当の時間をいつも楽しみにしていました．また，学生時代から社会人になってからも，新年早々に後藤先生のマンションで開かれる，大愚の会の新年会にもお招きに預かり，大きなタラバガニや鮭の丸焼き，といったこれ以上にない贅沢で豪華絢爛な手料理をご馳走になりました．BRA 書記の亀山日名子さんには，BRA のオフィスに訪問した際には，いつも温かいコーヒーやお菓子を準備していただきました．上記の方々に重ねてお礼を申し上げます．

ファイザー株式会社 臨床統計部部長の河合統介博士には，著者が初めて BRA の会合に参加させていただいたスプリング・フォーラム 2007 のときに初めてお会いし，BRA の諸先生方・諸先輩方ばかりで緊張状態の著者を BRA の一員として温かく輪の中に迎えていただきました．また，研究に向かう心構えから社会人としてのマナーに至るまで多くのことをご教示いただきました．就職活動の際には大変にお世話になり，入社後も著者が後期課程に入学を希望したときには多大なご助力をいただきました．常に目標を提示していただき，機会がある度に叱咤激励をしていただきました．ファイザー株式会社 臨床統計部課長の丸山奈美博士には，業務と研究との両立について諸種のご配慮をいただき，また，国際学会発表などの沢山の機会を与えてくださいました．何事も前向きにチャレンジすることの大切さを教えていただきました．また，臨床統計部の皆さまには折に触れて，温かい言葉で励ましていただきました．本当にありがとうございました．著者

は，博士後期課程（社会人コース）の約3年間にわたり，ファイザー株式会社から経済的な援助を受けました．ここにその援助に対し深謝いたします．

最後に，絶えず筆者の身を案じてくれた親戚一同の皆さま，そして，両親，弟に心より感謝いたします．

Abstract

In this paper, we focus on three different topics. They are "Predictive performance of Bayesian diagnoses", "A preliminary evaluation about health guidance" and "The impact of the shape of the underlying distribution of observations on test results". The main results of this study are as follows:

Predictive performance of Bayesian diagnoses. In a framework of Bayesian approach, though we have an advantage which we select various prior distributions according to the situation, the number of the model which we have to evaluate is very large. When we make model diagnoses, previously we need to confirm whether the model diagnoses meet our intended purpose of model selection. We are often interested to data which will be gained in future. So we consider two diagnostic methods that focus on prediction: Bayesian predictive information criterion (Ando, 2007), prior and posterior predictive checking approach (Box, 1980; Rubin, 1984; Gelman, Meng and Stern, 1996; Daimon and Goto, 2007). We try to clarify the characteristics of these approaches and express the situations of effective diagnosis. As the result, models with strong prior information gave lower BPIC than models with weak prior information totally. It means that BPIC prefer to models with strong prior information. Conversely, in the framework of predictive checking approach, models with weak prior information gave higher predictive checking probability than models with strong prior information. It means that predictive checking probability prefer to models with weak prior information. In our simulations, these findings were unaffected by whether prior mean was true or not. So we have a concern that it has possibilities of selecting not models with true prior mean but models with no true prior mean in several situations. Therefore, to select model appropriately, it is important to clarify the characteristics of these predictive model diagnoses in application situation and consider how to find the operational characteristics of the diagnoses (including combination) before model evaluation.

A preliminary evaluation about Health Guidance Since April 2008, Ministry of Health, Labor and Welfare of Japan has carried out Health Checkups and Healthcare Advice with a particular focus on the Metabolic Syndrome which make it obligatory for person aged 40 through 74 to reduce medical expenses and prevent lifestyle-related diseases. However, Kondo (2004) indicates a lack of foundation for health checkup. We also wonder about effect of making health checkup compulsory. The Health Checkup that aims to prevent disease was carried out in April 2004 and a doctor classified subjects into uncontrolled, directed (teaching of better living) and clinical group (includes medicine), based on their results. In this paper, we explore foundation about the doctor's judgment, especially classification of the directed group, attempting to figure the doctor's character, and further evaluate directed effect for the directed group. As a result, we confirmed that the doctor classified subjects from their body types such as weight and BMI, and that it reduced weight and BMI as the directed effect, but it gave increase of TG and decrease of HDL which are likely to develop abnormal lipid metabolism. So, we found that adequate evaluation about effect of health care advice leads to suggestion of scientific foundation for health checkups and health advice.

The impact of the shape of the underlying distribution of observations on test results In clinical research, we consider difference between pre- and post-treatment observations as an evaluation indicator for treatment effect. Then, though we generally focus on a normality of the difference, the relation between distributions of these treatment observations and the difference is not discussed in detail. In this paper, when it is assumed that pre- and post- treatment observations follow bivariate power-normal distribution, we clarify the relation between the distribution of these treatment observations and the distribution of the difference comprehensively and quantitatively, and evaluate the impact of the distribution of the difference on a paired and two-samples t-test which require the normal assumptions. As a result, the skewness of the difference of the distribution were very small compared to the distribution of these treatment observations and approached to symmetry. Moreover, we gained certain findings that the power in these tests remained high even if the normal assumption was violated a little, though the power in a paired and two-samples t-test decreased as the potential distribution was right-skewed. Thus we found that it is useful for the interpretation of the test results to focus on not only the distribution of the difference but also the potential distribution which these treatment

observations follow.

Acknowledgment

The author would like to express his deep and sincere gratitude to many people who gave much suggestions, helps, and encouragement throughout the preparation of his dissertation.

The author wishes to thank Professor Shingo Shirahata of Osaka University for providing helpful and useful comments. Professor Yutaka Kano of Osaka University provided the helpful comments. Professor Masayuki Uchida of Osaka University provided the useful suggestions. Associate professor Wataru Sakamoto of Osaka University reviewed this thesis carefully, and provided the helpful advice.

The author would especially like to thank Dr. Masashi Goto of Biostatistics Research Association (BRA), NPO, who led the author to the theme of this research and provided adequate and useful suggestions throughout writing this thesis.

The author wishes to thank all of members of BRA. In the meeting of BRA, he received valuable comments; especially, he would like to thank Dr. Satoru Fukinbara, Dr. Masaki Fujisawa, Dr. Tomoyuki Sugimoto, Dr. Toshio Shimokawa, Mr. Takao Takase, Mr. Takaharu Yamabe, Mr. Yasunobu Furukawa, Dr. Taro Amagasaki, Dr. Masanori Ito, Mr. Kimitoshi Ikeda, Dr. Takashi Nagakubo, Mr. Hiroki Motogaito, Dr. Kazushi Maruo, Mr. Motoki Ohe, Mr. Masatoshi Nakamura, Mr. Yusuke Yamaguchi.

The author has been able to devote so much time to his thesis because of the cooperation of his superiors, Department Manager Dr. Norisuke Kawai and Section Chief Dr. Nami Maruyama, in Pfizer Japan Inc. And the author would like to thank all the members of Clinical Statistics at Pfizer Japan Inc. for their continuing kindness and substantial support.

Finally, the author is grateful to his parents, brother for their support and encouragement.

Notations

notation	definition
Chapter 2	
θ	parameter
p	number of parameter
n	sample size
y,$\tilde{y}$	data
y_d,$\tilde{y}_d$	observed data
μ	mean
σ^2	variance
μ_0	prior mean
σ_0^2	prior variance
n_0	prior sample size
$p(\theta)$	prior probability
$p(y\|\theta)$	likelihood function of sampling distribution
$p(\theta\|y)$	posterior probability
$p(y,\theta)$	joint probability of parameter θ and data y
$p(\tilde{y},\theta\|y)$	Given y, joint probability of parameter θ and data $\tilde{y}$
$g(\cdot)$	predictive checking function
Ω	sample space
F	any events
E_i	measurable event
$\mathrm{P}r(E_i)$	generated probability
$\mathrm{P}r(E_i\|F)$	conditional generated probability
N	normal distribution

notation	definition
Chapter 3,4	
X_B,X_T	pre- and post-treatment observation
Δ	clinical effect
e_B,e_T	error term of pre- and post-treatment observation
X,x	probability variable and observed value on the original scale
$X^{(\lambda)}, x^{(\lambda)}$	probability variable and observed value on the transformed
ϕ	probability density function of standard normal distribution
Φ	cumulative density function of standard normal distribution
λ	shape parameter
μ	location parameter
σ	scale parameter
ϵ_p	100p percent point
N	normal distribution
PN	power-normal distribution
MPN	multivariate power-normal distribution

Contents

1. Introduction

1.1 Background and motivation

In this paper, we focus on three different topics, they are "Predictive performance of Bayesian diagnoses", "A preliminary evaluation about health guidance" and "The impact of the shape of the underlying distribution of observations on test results". In this paper, we introduce their backgrounds and motivations separately.

Predictive performance of Bayesian diagnoses In a process of Bayes inference which formulates iterative procedure of scientific research, we select a prior distribution based on cumulative experiences, experiments and knowledge, and compose a probability model under the prior distribution. Then "Criticism" and "Estimation" which Box(1980) refer to are repeated. After a model is composed from known data, it shows the necessary of data analysis for the model and more data (predictive part), as the result of the analysis, a revised model is obtained (posterior part). If the model is correct, we can make proper inferences about parameter using a posterior distribution which is combination of prior information and data information. However, because the posterior distribution is composed using only a pair of data that has actually occurred, it is important to make diagnosis/checking for the model. Then, we can diagnose the model in the following three terms at least: (1) Sensitivity analysis for variation of prior distribution and likelihood, (2) Appropriateness of posterior inference for the model in the context of the actual application, (3) Fitness of the model to the data. In this paper, we notice on a model diagnosis in terms of (3). In the framework of traditional model selection such as Bayesian information criteria (BIC) (Schwarz, 1978) and Bayes factor (BF), a model with highest posterior model probability is selected. However, in fact, we are often interested to data which will be gained in the future. Therefore, we consider two diagnostic methods that focus on prediction: Bayesian predictive information criterion (BPIC) (Ando, 2007), prior and posterior predictive checking approach (Prior- and Post-PCA) (Box, 1980; Rubin, 1984; Gelman, Meng and Stern,

1996; Daimon and Goto, 2007).

BIC which is most familiar information criterion in Bayesian approach is a criterion of model evaluation based on a posterior probability and select best model with the highest posterior probability among several model candidates. Bayes factor, extended Bayesian information criteria (Konishi, Ando and Imoto, 2004) are well-known as other model evaluation criteria in the same position. Recently, BPIC has been proposed as new diagnosis method which evaluates model fitness from a position of the prediction. BPIC selects a model with the highest expected log likelihood.

By integrating model consisted of prior distribution of the parameter and joint probability distribution of data in the parameter, we can get a predictive distribution (refer to it as "prior predictive distribution"). Box (1980) proposed prior predictive checking approach which compares the prior predictive distribution of future observations to the data that have actually occurred and judge an appropriateness of the model. Then, we can consider whether data is included in the prior predictive distribution, and can check the compatibility between prior information and data information. However, in actual situation, it is desirable to develop the diagnosis which focuses on selecting the model for meeting our intended purpose rather than whether model is true (Tiao and Xu, 1993). So it is often necessary to assess not only the model itself but also interesting indices such as sample mean, sample variance or on which the decision making is based. In the framework of the prior predictive checking approach, it is also possible to evaluate the interesting situation by setting an appropriate predictive checking function and referring predictive probability of the predictive checking function obtained from the data that have actually occurred to predictive distribution of the predictive checking function.

Rubin (1984) proposed Post-PCA as an alternative method of the prior predictive checking approach. This approach focuses on compatibility between posterior information and data information. An initial paper which defined the idea about this posterior predictive checking approach is Guttman (1967) and Dempster (1971). After that, Gelman *et al.* (1996) extended this approach and proposed a method which conducts a posterior prediction by numerical calculation as the diagnosis of fitness of the single model for directly measuring the diremption between data and an assumed model. The characteristic of the model diagnosis is that a parameter of model is not treated as a point estimator but is generated from a posterior distribution. This point is different from a classical model diagnosis. Therefore, it is possible to diagnose a model taking

into consideration uncertainly of parameter. Also, just like prior predictive checking approach, we can calculate posterior predictive checking probability for interesting indices. For example, even in many model diagnoses such as test for a outlier, residual plot and normal plot, it is interpretable to measure the diremption between expected results under an assumed model and actual data (Gelman, Carlin, Stern & Rubin, 2004).

In prior and posterior predictive checking approach, we can conduct the diagnoses for these model from the two viewpoints which are "Exploratory data analysis" and "Confirmatory data analysis". From the viewpoint of "Exploratory data analysis", it is possible to find the shape of the predictive distribution of data and predictive checking function visually by showing data and the value of predictive checking function. Also, from the viewpoint of "Confirmatory data analysis", it is possible to measure the significance of model as the prior and posterior predictive checking probability which show the diremption between the model and data. Moreover, in the case of that we use both prior and posterior predictive checking approach simultaneously, we can get the following findings. For example, if it is suspicious for a model or an interesting index in prior predictive checking approach, it doubts about the appropriateness of prior distribution. So the posterior predictive checking approach based on the prior distribution become meaningless. However, if it is suspicious for a model or an interesting index in not prior predictive checking approach but posterior predictive checking approach, it means that the assumed model for sampling distribution is unworthy of belief (Daimon & Goto, 2007).

Though the above BPIC and predictive checking approach are only a few diagnoses focused on the prediction, the predictive performance remained unclear. So in chapter 2, we focus on the above BPIC and predictive checking approach and evaluate the predictive performance under various situations.

A preliminary evaluation about health guidance A prevalence and reserves of "lifestyle-related disease (adult disease)" increase as the lifestyle habit changes and the number of elderly people grows. "lifestyle-related disease" is a collective term of some diseases involving lifestyle such as a smoking, diet, drinking, exercise and sleep. And the incidence of cerebral stroke and ischemic cardiac disease increases as the risk factor of lifestyle-related disease such as hypercholesteremia pile up.

Since April 2008, Ministry of Health, Labor and Welfare of Japan has carried out "Health

Checkups and Healthcare Advice" with a particular focus on the Metabolic Syndrome which make it obligatory for person aged 40 through 74 to reduce medical expenses and prevent lifestyle-related diseases (Health Service Bureau of Health, Labour and Welfare, 2007). Though the aim of "Health Checkups and Healthcare Advice" is "Reduction of medical cost" and "Prevention of the lifestyle-related disease", it is deeply concerned about the appropriateness of "Practice criteria" and no evidence for "Prevention"（Ohgushi, 2006: 2007）. Also as Kondo(2004) indicates a lack of foundation for health checkup, we wonder about making health checkup compulsory, too.

The health checkup that aims to prevent disease was carried out in April 2004 and a doctor classified subjects into uncontrolled, directed (teaching of better living) and clinical group (includes medicine), based on their results. After that, the teaching of better living or treatment was conducted for the directed and clinical groups and how the clinical test results improve was examined. Here the definition for the clinical group was based on the constant criterion value and the definition for the directed group was based on the judgment of the doctor. By using this data, we explore foundation about the doctor's judgment, especially classification of the directed group, attempting to figure the doctor's character, and further evaluate directed effect for the directed group.

The impact of the shape of the underlying distribution In clinical research, we consider the difference between pre- and post-treatment observations as an evaluation indicator for treatment effect. When we examine whether the treatment effect exists or not, it is often assumed that the observations follow normal distribution, and a paired t-test in a one-sample problem and two-samples t-test in a two-sample problem are applied for them. But a lot of endpoints exist in the actual clinical research and the endpoints do not always follow the normal distribution. When it is assumed that pre- and post-treatment observations follow various distributions, we evaluate the impact of them on tests which require the normal assumptions. Because we often conduct two-group comparison between actual group and placebo group in clinical research, we consider not only one-sample problem but also two-sample problem. We evaluate the performance of the paired t-test in one-sample problem and the two-samples t-test in two-sample problem, but also use the Wilcoxon signed rank test in one-sample and the Wilcoxon rank sum test as the comparison of the t-tests. To clarify the relationship and the structure between the distributions of pre- and post-observations and the distribution of the difference, we especially

focus on the following points.

(a) Relation between non-normality of distributions of pre- and post-observations and non-normality of the distribution of the difference.

(b) Influence of non-normality of distribution of the difference on power in above tests.

(c) Availability of interpreting the test results corresponding to distributions of pre- and post-treatment samples.

As an approach to (a), we assume that pre- and post-observations follow a bivariate power-normal distribution (BPND: Goto and Hamasaki, 2002) in order to consider the relationship between the distributions of pre- and post-observations and the distribution of the difference comprehensively and quantitatively. The bivariate power-normal distribution is the bivariate extended form of an univariate power-normal distribution (PND) which was proposed by Goto, Matsubara and Tsuchiya (1983). The univariate power-normal distribution is defined as the distribution which the observations before the power-transformation (Box and Cox, 1964) follow, and contains various distributions including well-known normal distribution and log-normal distribution, so can cover real situations to some extent and is useful to evaluate the discrepancies between ideal (model and hypothesis) and reality (data) (Goto, Uesaka and Inoue, 1979; Goto and Inoue, 1980; Goto, Matsubara and Tsuchiya, 1983). Moreover, because pre- and post-observations have the correlated relationship, the bivariate power-normal distribution including the correlation structure is suitable for assessing our problem. Because the PND express the features of the distribution which the data follow even if the distribution is not known previously, we notice on the PND in this paper. Additionally, to make clear the situation examined in this paper, we identify the distribution of pre- and post-observations by using a shape parameter (power-parameter) which expresses a skewness of the distribution and an indicator which express a variation of the distribution defined later And we derive the distributions of the difference from numerical integral in several situations and inquire the properties about the distributions of the difference. As an approach to (b) and (c), we examine the impact of the shape of the potential distribution on the results of the t-tests.

1.2 Components of this paper

In chapter 2, we explain about BPIC and the predictive checking approach, and describe the results and new findings obtained from the simulation to make clear the the predictive performance. In chapter 3, we conduct a preliminary evaluation about health guidance for data of 1,141 subjects who had the health checkup that was carried out in April 2004. And we summarize the results of the data analysis. In chapter 4, we examine the impact of the shape of the underlying distribution of observations on test results and specifically present occasions where t-test works well. In chapter 5, we contain our concluding remarks about the findings obtained from chapter 2,3 and 4.

2. Predictive performance of Bayesian diagnoses

2.1 Introduction

In a process of Bayes inference which formulates iterative procedure of scientific research, we select a prior distribution based on cumulative experiences, experiments and knowledge, and compose a probability model under the prior distribution. Then "Criticism" and "Estimation" which Box (1980) refer to are repeated. After a model is composed from known data, it shows the necessary of data analysis and more data (predictive part), as the result of the analysis, a revised model is obtained (posterior part). If the model is correct, we can make proper inferences about parameter using a posterior distribution which is combination of prior information and data information. However, because the posterior distribution is composed using only a pair of data that has actually occurred, it is important to make diagnosis/checking for the model. Then, we can diagnose the model in the following three terms at least: (1) Sensitivity analysis for changes of prior distribution and likelihood, (2) Appropriateness of posterior inference for the model in the context of the actual situation, (3) Fitness of the model to the data. In this paper, we notice on a model diagnosis in terms of (3). In the framework of traditional model selection such as Bayesian information criteria (BIC) (Schwarz, 1978) and Bayes factor (BF), a model with highest posterior model probability is selected. However, in fact, we are often interested to data which will be gained in the future. Therefore, we consider two diagnostic methods that focus on prediction: Bayesian predictive information criterion (BPIC) (Ando, 2007) and prior and posterior predictive checking approach (Prior- and Post-PCA) (Box, 1980; Rubin, 1984; Gelman, Meng and Stern, 1996; Daimon and Goto, 2007).

BIC which is most familiar information criterion in Bayesian approach is a criterion of model evaluation based on a posterior probability and select best model with the highest posterior probability among several model candidates. Bayes factor and extended Bayesian information criteria (Konishi et al., 2004) are well-known as other model evaluation criteria in the same

position. Recently, BPIC has been proposed as new diagnosis method which evaluates model fitness from a position of the prediction. BPIC selects a model with the highest expected log likelihood.

Prior-PCA provides checking models or indices by comparing data to the prior predictive distribution. This approach contrasts the prior information and the data information, and checks their compatibility. Post-PCA replaces the role of the prior distribution in Prior-PCA with it of the posterior distribution. Main feature of Prior- and Post-PCA is to be able to check not only a model itself but also interesting indices or statistics by setting proper predictive checking functions. Therefore, we can judge whether the model is suitable for the specific occasion or not. It is considered that this feature is quite effective because we do not always have to focus on the model itself and can select a proper model which meets the purpose of the research.

Though Bayesian approach has the advantage that it is possible to select a prior distribution according to an individual situation, there exists many models which should be evaluated. So we consider that BPIC and PCA have a specified situation suitable for each model diagnosis. But the profiles about BPIC and PCA have not been clarified enough yet. In this paper, our purpose is to make clear the properties of BPIC and PCA and propose the effective diagnosis situations.

In section 2.2, we explain BPIC and PCA. In section 2.3, we apply Bayesian predictive diagnoses which were introduced in section 2.2 to data of triglyceride concentration in the plasma, and evaluate the appropriate of several models. Several simulations are conducted to evaluate the two diagnosis methods and some productive findings are summarized in section 2.4. Finally, section 2.5 contains our concluding remarks.

2.2 Bayesian predictive diagnosis

2.2.1 Bayesian predictive information criterion

As the model diagnosis, Bayesian predictive information criterion (BPIC) is proposed by Ando (2007). BPIC is defined as an estimator of the posterior mean of the expected loglikelihood of the predictive distribution. In this criterion, we can evaluate the predictive distributions of hierarchical and empirical Bayes model even when the assumed family of probability distributions does not always contain the true model.

Akaike's information criterion (AIC; Akaike, 1973) and Generalized information criterion (GIC; Konishi and Kitagawa, 1996) known well as information criterion selects the maximum model of the expected loglikelihood using Kullback-Leibler information as the indicator for measuring a distance between assumed statistical model and true model. However, BPIC evaluate the statistical model composed of the posterior expected loglikelihood.

As notations, $p(\cdot)$ shows probability density function, where $p(\theta)$ is a prior probability which represents the degree of confidence for θ before getting data y, $p(y|\theta)$ is the likelihood function of the sampling distribution which data y (which generates from a parametric distribution) follow and $\int p(y|\theta)p(\theta)d\theta$ is a normalized constant. The probability distribution of the posterior probability $p(\theta|y)$ is posterior distribution and the probability distribution of the prior probability $p(\theta)$ is prior distribution.

Then, the posterior expected loglikelihood is given by

$$\eta(G) = \int \left\{ \int \log p(\tilde{y}|\theta)p(\theta|y)d\theta \right\} dG(\tilde{y}),$$

where $G(\tilde{y})$ is true model, $\tilde{y}$ is future observation and y is observation. Though the posterior expected loglikelihood is calculated from true model, the true model is actually unknown. So we have to calculate the estimator of the posterior expected loglikelihood $\eta(G)$.

By using the empirical distribution function as the nature estimator of the posterior loglikelihood, the following posterior loglikelihood is obtained.

$$\eta(\hat{G}) = \frac{1}{n} \int \log p(y|\theta)p(\theta|y)d\theta$$

However, the posterior loglikelihood $\eta(\hat{G})$ is calculated from both the Bayes estimator and the empirical distribution function, so the bias exists as the estimator of the posterior expected log likelihood. Therefore we have to reduce the bias

$$b(G) = \int \left\{ \eta(\hat{G}) - \eta(G) \right\} dG(y)$$

The estimator of the bias of the posterior loglikelihood $\eta(\hat{G})$ is expressed by

$$\begin{aligned} \hat{b}(G) \quad \approx \quad & \frac{1}{n} \int \left[\int \log\{p(y|\theta)p(\theta)\}p(\theta|y)d\theta \right] dG(y) \\ & -\frac{1}{n} \log\{p(y|\theta_0)p(\theta_0)\} + \frac{1}{n} tr\{S^{-1}(\theta_0)Q(\theta_0)\} + \frac{p}{2n} \end{aligned}$$

θ_0 is a parameter to maximize a penalized expected loglikelihood

$$\int \{\log p(y|\theta) + \log p_0(\theta)\} g(y)dy,$$

where $\log p_0(\theta) = \lim_{n\to\infty} n^{-1} \log p(\theta)$. And $Q(\theta)$ and $S(\theta)$ is defined as

$$\begin{aligned} Q(\theta) &= \int \left[\frac{\partial \log\{p(y|\theta)p_0(\theta)\}}{\partial\theta} \frac{\partial \log\{p(y|\theta)\}}{\partial\theta^T} \right] dG(y), \\ S(\theta) &= \int \left[\frac{\partial^2 \log\{p(y|\theta)p_0(\theta)\}}{\partial\theta\partial\theta^T} \right] dG(y). \end{aligned}$$

In the actual calculation, we replace the true model G to the empirical distribution function $\hat{G}$, θ_0 to $\hat{\theta}_n$, $S(\theta_0)$ and $Q(\theta_0)$ to $Q_n(\hat{\theta}_n)$ and $S_n(\hat{\theta}_n)$. Then

$$\begin{aligned} \hat{b}(\hat{G}) &= \frac{1}{n} \int p(y|\theta)p(\theta)p(\theta|y)d\theta \\ &\quad -\frac{1}{n} \log p(y|\hat{\theta}_n)p(\hat{\theta}_n) + \frac{1}{n} tr S_n^{-1}(\hat{\theta}_n) Q_n(\hat{\theta}_n) + \frac{a}{2n}, \end{aligned}$$

and

$$\begin{aligned} Q_n(\hat{\theta}_n) &= \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\partial\{\log p(y_i|\theta) + \log p(\theta)/n\}}{\partial\theta} \cdot \right. \\ &\quad \left. \frac{\partial\{\log p(y_i|\theta) + \log p(\theta)/n\}}{\partial\theta}\Big|_{\theta=\hat{\theta}_n} \right] \\ S_n(\hat{\theta}_n) &= -\frac{1}{n} \sum_{i=1}^{n} \left[\frac{\partial^2\{\log p(y_i|\theta) + \log p(\theta)/n\}}{\partial\theta\partial\theta^{\mathrm{T}}}\Big|_{\theta=\hat{\theta}_n} \right] \end{aligned}$$

Also, a is number of parameter and n is sample size.

Then, under the weak regular conditions (unimodal of the posterior distribution, consistency of the posterior mode, asymptotic normality), BPIC is defined as follows:

$$\mathrm{BPIC} = -2 \int \log\{p(y|\theta)\} p(\theta|y) d\theta + 2n\hat{b}(\hat{G}) \tag{2.1}$$

We select a lowest model of BPIC as well as other information criteria such as AIC.

Calculation In section 2.3 and 2.4, when it is assumed that data y follow a normal distribution $\mathrm{N}[\mu, \sigma^2]$ with known variance σ^2, we set $\mathrm{N}[\mu_0, \sigma_0^2]$ as prior distribution of mean parameter μ. Then, BPIC

$$\mathrm{BPIC} = -2n\eta(\hat{G}) + 2n\hat{b}(\hat{G})$$

is calculated from the following posterior loglikelihood

$$\eta(\hat{G}) = -\frac{\log(2\pi\sigma^2)}{2} - \frac{1}{2n\sigma^2} \sum_{i=1}^{n} \{(y_i - \hat{\mu}_n)^2 + \sigma_n^2\} \tag{2.2}$$

and bias

$$\hat{b}(\hat{G}) = -\left\{\frac{\sigma_n^2}{2\sigma^2} + \frac{\sigma_n^2}{2n\sigma_0^2}\right\} + \frac{S_n^{-1}(\hat{\mu}_n)Q_n(\hat{\mu}_n)}{n} + \frac{1}{2n}, \tag{2.3}$$

where

$$\begin{aligned}
\hat{\mu}_n &= \frac{\mu_0/\sigma_0^2 + \Sigma_{i=1}^n y_i/\sigma^2}{1/\sigma_0^2 + n/\sigma^2} \\
\sigma_n^2 &= \frac{1}{1/\sigma_0^2 + n/\sigma^2} \\
Q_n(\hat{\theta}_n) &= \Sigma_{i=1}^n\{(y_i - \hat{\mu}_n)/\sigma^2 + (\mu_0 - \hat{\mu}_n)/(n\sigma_0^2)\}^2/n \\
S_n(\hat{\theta}_n) &= \frac{1}{n\sigma_n^2}.
\end{aligned}$$

2.2.2 Predictive checking approach

Prior predictive checking approach

Given partition $\{E_1, E_2, \ldots, E_n\}$ of sample space Ω and any events F, if measurable event $E_1, E_2, \ldots, E_n$ are mutually exclusive and $\bigcup_{i=1}^n E_i = \Omega$, we can obtain the following equation using Bayes' theorem.

$$\Pr(E_i|F) = \frac{\Pr(E_i)\Pr(F|E_i)}{\sum_{j=1}^n \Pr(E_j)\Pr(F|E_j)}, \tag{2.4}$$

where $\Pr(E_i)$ is a generated probability of measurable event E_i, $\Pr(E_i|F)$ is a generated probability (conditional probability) of measurable event E_i under the condition F.

Though the equation expresses the calculation of the conditional probability, in an inferential problem for unknown parameter θ, by the Bayes' theorem, we can get the posterior probability

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}. \tag{2.5}$$

It shows the degree of the confidence for θ. Though we can get the posterior distribution by connecting data to a prior distribution of parameter in Bayes' theorem, it is suspicious for the model in the case where it is difficult to consider that an actual data is generated from an assumed model. When it is assumed that model including prior information is correct, a distribution of all possible sample space is a prior predictive distribution. From now, using the prior predictive distribution, we explain about a prior predictive checking approach (Box, 1980) which is an approach checking compatibility between data and prior information.

A model including prior and data information is showed by the joint probability function of parameter θ and data y

$$p(y, \theta) = p(y|\theta)p(\theta). \tag{2.6}$$

This is calculated by the product between a prior probability $p(\theta)$ of θ and a likelihood function of a sampling distribution. Then, prior predictive probability $p(y)$ is given as a distribution of all y in

$$p(y) = \int p(y, \theta)d\theta, \tag{2.7}$$

where integral region is total parameter space of θ. The probability distribution of the prior predictive probability is a prior predictive distribution.

For a known data y_d,

$$p(y_d, \theta) = p(\theta|y_d)p(y_d). \tag{2.8}$$

Here index d represents the known data or statistic obtained from the known data. The first factor in this equation is the posterior probability $p(\theta|y_d)$ of θ given y_d and we can get

$$p(\theta|y_d) = p(\theta, y_d)/p(y_d).$$

For second factor, we can get

$$p(y_d) = \int p(y_d|\theta)p(\theta)d\theta, \tag{2.9}$$

and $p(y_d)$ represents prior predictive probability for actual data y_d. Then, the model in prior predictive checking approach can be checked by comparing $p(y)$ to $p(y_d)$. So the comparison is measured by the prior predictive checking probability

$$\Pr[p(y) < p(y_d)]. \tag{2.10}$$

So in the prior predictive checking approach, a model is evaluated by comparing the prior predictive distributions of future observations to the data that have actually occurred and calculating the prior predictive checking probability (Prior-PCP). If the probability is small (i.e. <0.05), we judge that data y_d do not follow the model created by the prior distribution, and suspect the reliability for the model.

It is also possible to evaluate not only the model itself but also interesting indices or statistics by setting proper predictive checking functions. Then, we compare the prior predictive probability $p\{g(y_d)\}$ of $g(y_d)$ to the prior predictive probability $p\{g(y)\}$ of $g(y)$ and evaluate

the appropriateness of the model. The prior predictive checking probability of the predictive checking function is calculated by

$$\Pr[p\{g(y)\} < p\{g(y_d)\}]. \tag{2.11}$$

However, as the fault of the prior predictive checking approach, when it is assumed that the parameter follow improper prior distribution, it is considered that the prior predictive distribution itself is improper and the occasion that we cannot check the model even if the posterior distribution is not improper.

Posterior Predictive checking approach

Rubin (1984) proposed Post-PCA as an alternative method of the prior predictive checking approach. In the posterior predictive checking approach, a model is evaluated by comparing the posterior predictive distributions of future observations to the data that have actually occurred and calculating the posterior predictive checking probability (Post-PCP).

Setting a posterior probability as $p(\theta|y)$, we have a Bayes model

$$p(\tilde{y}, \theta|y) = p(\tilde{y}|\theta, y)p(\theta|y),$$

where $\tilde{y}$ are future observations.

Then, a posterior predictive distribution for the observations of the future, $\tilde{y}$, is obtained by

$$p(\tilde{y}|y) = \int_{\theta\in\Theta} p(\tilde{y}, \theta|y)d\theta.$$

Given the actual data $\tilde{y}_d$, Post-PCA for the model itself is calculated by comparing the density function $p(\tilde{y}|y)$ to the posterior density at $\tilde{y}_d$, $p(\tilde{y}_d|y)$, as the below:

$$\Pr[p(\tilde{y}|y) < p(\tilde{y}_d|y)|y = \tilde{y}_d] \tag{2.12}$$

As the same in the prior predictive checking approach, if the probability is small (i.e. <0.05), we judge that data $\tilde{y}_d$ do not follow the model created by the posterior distribution, and suspect the reliability for the model. For even posterior predictive checking approach as well as prior predictive checking approach, we can evaluate Post-PCA for interesting indices $g(\tilde{y})$ as the below:

$$\Pr[p\{g(\tilde{y})|y\} < p\{g(\tilde{y}_d)|y\}|y = \tilde{y}_d] \tag{2.13}$$

When we evaluate a model under both the prior predictive distribution in Pre-PCA and the posterior predictive distribution in Post-PCA, the large difference in the prior and posterior predictive checking probabilities indicates that the prior distribution is wrong.

In these prior and posterior predictive checking approach, without any specified model of alternative hypothesis, we can evaluate the fitness of the single model. Also, we think that these approaches are very useful for the selection of the model because we can also compare the predictive checking probabilities between several candidate models simultaneously.

Interruption of predictive checking probability

From a practical point of view, if large diremption between model and data exists and the predictive checking probability is near 0, the reliability of the model is suspicious because the model do not express the event which the data expresses. So generally, an improvement to a model with higher predictive checking probability is desirable. Also, because the prior predictive checking approach evaluate the model under the prior predictive distribution and the posterior predictive checking approach evaluate the model under the posterior predictive distribution, the clear difference between prior and posterior predictive checking probability implies that the prior distribution is suspicious.

However, we have to pay attention to what the predictive checking probability shows not "statistical significance" but "practical significance" (Gelman *et al.*, 2004). So a goal at predictive checking approach is not to reject the model but to judge whether data generate from the model.

Also, four major schools exist in statistical science. They are Neyman-Pearson, Fisher and likelihood school along with Bayesian school exist (Oakes, 1986). Neyman-Pearson and Fisher school criticize Bayesian school by reason of "Lack of objectivity for probability". So "Neyman-Pearson and Fisher school" and "Bayesian school" developed separately. But through the use of predictive checking approach which is complementary role of "Criticism" and "Estimation" repeated in a process of Bayes inference which formulate iterative procedure of scientific research, the connection between Neyman-Pearson/Fisher and Bayesian schools would be possible by interpreting the existing statistical method such as hypothesis test (Neyman-Pearson) and significant test (Fisher) in the framework of Bayesian approach.

Calculation As well as "Calculation" in section 2.2.2, we explain about how to calculate Prior- and Post-PCP when it is assumed that data y follow a normal distribution $\mathrm{N}[\mu, \sigma^2]$ with known variance σ^2 and we set $\mathrm{N}[\mu_0, \sigma_0^2]$ as prior distribution of mean parameter μ.

Suppose that $\bar{y}$ is $\bar{y} = \sum_{i=1}^n y_i/n$, s^2 is $s^2 = \sum_{i=1}^n (y_i - \bar{y})^2/(n-1)$.Then the likelihood is expressed as

$$p(y|\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp[\{-n(\bar{y}-\mu)^2 + \sum_{i=1}^n (y_i - \bar{y})^2\}/2\sigma^2].$$

Also, the prior predictive distribution is calculated by

$$p(y) \propto \frac{1}{\sigma^{n-1}\left(\sigma^2/n + \sigma_0^2\right)^{1/2}} \exp[-\{\sum_{i=1}^n (y_i - \bar{y})^2/\sigma^2 + (\bar{y}-\mu_0)^2/(\sigma^2/n + \sigma_0^2)\}/2].$$

Then, Prior-PCP is given by

$$\Pr[p(y) < p(y_d)] = \Pr[\chi_n^2 > g(y_d)], \tag{2.14}$$

where

$$g(y_d) = \frac{(\bar{y}_d - \mu_0)^2}{\sigma^2/n + \sigma_0^2} + \frac{(n-1)s_d^2}{\sigma^2}$$

Moreover Prior-PCP for sample mean $\bar{y}$ is given by

$$p(\bar{y}) \propto \frac{1}{(\sigma_0^2 + \sigma^2/n)^{1/2}} \exp[-(\bar{y}-\mu_0)^2/\{2(\sigma_0^2 + \sigma^2/n)\}],$$

and

$$\Pr[p(\bar{y}) < p(\bar{y}_d)] = \Pr\left[z > \left|(\bar{y}_d - \mu_0)/(\sigma_0^2 + \sigma^2/n)^{1/2}\right|\right], \tag{2.15}$$

where $z \sim \mathrm{N}[0,1]$.

Also, the posterior predictive distribution is calculate by

$$p(\tilde{y}|y) \propto \frac{1}{\sigma^{n-1}(\sigma^2/n + \sigma_n^2)^{1/2}} \exp[-\{\sum_{i=1}^n (\tilde{y}_i - \bar{\tilde{y}})^2/\sigma^2 + (\bar{\tilde{y}} - \mu_n)^2/(\sigma^2/n + \sigma_n^2)\}/2],$$

where

$$\begin{aligned} \mu_n &= (\mu_0/\sigma_0^2 + \sum_{i=1}^n y_i/\sigma^2)/(1/\sigma_0^2 + n/\sigma^2), \\ \sigma_n^2 &= 1/(1/\sigma_0^2 + n/\sigma^2). \end{aligned}$$

Then Post-PCP for model is given by

$$\Pr[p(\tilde{y}|y) < p(\tilde{y}_d|y)|y = \tilde{y}_d] = Pr[\chi_n^2 > g(\tilde{y}_d)|y = \tilde{y}_d], \tag{2.16}$$

Table 2.1: BPIC and PCP for triglyceride concentration data (Wood, 1973)

		Prior distribution $\mu \sim \mathrm{N}[\mu_0, \sigma_0^2]$			
		$\mathrm{N}[125, 20]$	$\mathrm{N}[200, 20]$	$\mathrm{N}[125, 4000]$	$\mathrm{N}[200, 4000]$
BPIC		40.35	41.36	41.98	41.98
PCP	Prior-model	0.527	0.000	0.528	0.461
	Post-model	0.527	0.003	0.528	0.527
	Prior-mean	0.887	0.334	0.978	0.255
	Post-mean	0.900	0.214	0.997	0.884

where

$$g(y_d) = \sum_{i=1}^{n} (\tilde{y}_{d,i} - \bar{\tilde{y}}_d)^2/\sigma^2 + (\bar{\tilde{y}}_d - \mu_n)^2/(\sigma^2/n + \sigma_n^2) + (n-1)s_d^2/\sigma^2.$$

Moreover, Post-PCP for sample mean $\bar{\tilde{y}}$ is given by

$$p(\bar{\tilde{y}}|y) \propto \frac{1}{(\sigma_n^2 + \sigma^2/n)^{1/2}} \exp[-(\bar{\tilde{y}} - \mu_n)^2/\{2(\sigma_n^2 + \sigma^2/n)\}]$$

and

$$\Pr[p(\bar{\tilde{y}}|y) < p(\bar{y}_d|y = \tilde{y}_d)] = \Pr[z > |(\bar{\tilde{y}}_d - \mu_n)/(\sigma_n^2 + \sigma^2/n)^{1/2}|y = \tilde{y}_d].$$

2.3 Examination on some literature example

We applied these Bayesian predictive diagnoses which were introduced in section 2.2 to data of triglyceride concentration in the plasma (Wood, 1973), and evaluated the appropriate of several models. These data (sample mean 126.8, sample variance 3973) were measured to examine whether improvement in lifestyles impact on the measurements by a team in Stanford University, and we used the pre-treatment data here. The sample size was 30. We assumed that the data followed $\mathrm{N}[\mu, \sigma^2]$ where the variance σ^2 was known and the mean μ followed the normal prior distributions $\mathrm{N}[\mu_0, \sigma_0^2]$. We set the prior mean μ_0 as 125 (close to sample mean 126.8) or 200 (not close to sample mean) and the prior variance σ_0^2 as 20 (strong prior information) or 4000 (weak prior information). Then, we calculated BPIC and PCP for model and sample mean and represented the results in Table 2.1. In the table, we gained the results that the model with the prior distribution $\mathrm{N}[125, 20]$ (close to sample mean and strong prior information) had the lowest BPIC, but the model with the prior distribution $\mathrm{N}[125, 4000]$ (close to sample mean and weak

prior information) indicated higher Prior- and Post- PCP for sample mean than the model with the prior distribution N[125, 20] in PCA. However, Prior- and Post- PCP for model were almost the same probabilities together. Moreover, BPIC for the model with weak prior information (N[125, 4000] and N[200, 4000]) had much the same value and there was no difference between them.

2.4 Simulation

In this section, to clarify the situation that Bayesian predictive diagnoses select the model including appropriate prior distribution, we conduct some simulations. Taking a notice on making the interpretation of the results easy, we conduct a setting of simulation. It is important to consider the amount of information of the prior distribution in advance because it is useful to interpret the simulation results. So we express the amount of information of the prior distribution (prior information) as "the number of observation which is required for obtaining the same estimate accuracy as Bayes estimator" (Mori, 2010) and describe it as "prior samples". In this simulation, we assume that independent samples follow a normal distribution with known variance, $y_i \sim \mathrm{N}[\mu, \sigma^2](i = 1, 2, \cdots, n)$, and a prior distribution of mean μ follows a normal distribution, $\mu \sim \mathrm{N}[\mu_0, \sigma_0^2(= \sigma^2/n_0)]$. When we set a square error as a loss function, Bayes estimator of mean μ (expectation of posterior distribution) is $\delta(y) = (n\bar{y} + n_0\mu_0)/(n + n_0)$ and Bayes risk is $\mathrm{E}[\mathrm{E}[(f(y) - \mu)^2|\mu]] = \sigma^2/(n + n_0)$. Also, Bayes risk of a sample mean in adding m observations, $\sum_{i=1}^{m+n} y_i/(m + n)$, is $\sigma^2/(m + n)$, so these Bayes risks are equal in $m = n_0$. Therefore, we can understand that the prior distribution $\mathrm{N}[\mu_0, \sigma^2/n_0(= \sigma_0^2)]$ has information about n_0 samples. We call n_0/n "Proportion of prior sample". Considered to the information about prior sample, we plan simulations.

Moreover, to investigate the impact of diremptions from the true prior mean on the results, we calculate μ_0/σ (Prior mean/Standard deviation) and call prior effect size (Prior ES). In this simulation, because we set that true prior mean μ_0 is 0, the large prior ES mean that the prior mean (which we use) is apart from true prior mean. About sample size n, because the difference of the results between BPIC and PCP was expressed even in 30 samples from the results of section 2.3, we set a broad range between $n = 10$ and $n = 1000$.

2.4.1 Simulation (1)

Purpose

It is important to evaluate an impact of a prior distribution in Bayes predictive diagnoses because Bayes model is composed by a prior distribution (or a posterior distribution) and a likelihood.

The purpose of simulation (1) is to assess the models with several different prior distributions in terms of prediction by calculating BPIC and PCP of these models from independent samples which follow true distribution and clarify these characteristics.

Method

We assume that true distribution is normal distribution with known variance ($\sigma^2 = 100$) and mean prior parameter. We define several prior distributions taking into the prior information and whether prior mean is true value or not. In detail, suppose that true value of mean parameter μ is set at 0 and prior mean μ_0 and prior variance σ_0^2 take the value of $\mu_0 = 0, 1.5$ and $\sigma_0^2 = 0.005, 0.01, 0.025, 1$. Then, prior samples are $n_0 = 100, 50, 20, 0.5$.

For all pattern of prior distributions which are determined by a combination of prior mean and prior variance, generate independent samples of sample size $n = 5, 20, 50, 100$ from true distribution N$[0, 0.5]$ and calculate BPIC and Prior- and Post-PCP for model and sample mean. We repeat this process 10,000 times and summary the results.

Result

The results of BPIC were shown in Figure 2.1 and those of PCP shown in Figure 2.2- 2.5. The horizontal lines in these figures represented 25, 50, 75% points of the simulation results from the bottom. The numbers in x-axis represents the following prior distributions: $\mu \sim$ (1) N$[0, 0.005]$, (2) N$[0, 0.01]$, (3) N$[0, 0.025]$, (4) N$[0, 1]$, (5) N$[1.5, 0.005]$, (6) N$[1.5, 0.01]$, (7) N$[1.5, 0.025]$, (8) N$[1.5, 1]$.

First we considered the results of BPIC. From the results in Figure 2.1, we found that BPIC for the prior distribution (1) with true prior mean and the strongest prior information was almost the lowest value compared to BPIC for other prior distributions. Also, by comparing two cases ((1) and (5)) of the models with the strongest prior information, we observed that

BPIC for the prior distribution with true prior mean was much smaller than BPIC for the prior distribution with no true prior mean. On the other hand, by comparing two cases ((4) and (8)) of the models with different prior means and the weakest prior information, we observed that there was not much difference between these BPIC. So these results indicated that BPIC was suitable for model selection among models with strong prior information.

Next we considered the results of Prior- and Post-PCP. From the results in Figure 2.2 ($n = 5$), we found that Pre- and Post-PCP of model and sample mean in four models composed by the prior distribution with true prior mean were totally high, and especially Prior- and Post-PCP for the prior distribution (4) with the weakest prior information of them were highest. For the model composed of prior distribution (8) with no true prior mean and the weakest prior information, Post-PCP of model and sample mean were higher than those of other models with no true prior mean. Moreover, by comparing two models ((4) and (8)) with different prior means and the weakest prior information, we could not see the difference between Prior- and Post-PCP of model. However, Post-PCP of sample mean for the model (8) was much larger than Pre-PCP of sample mean for it.

Because PCA has the characteristics that the large difference between the Prior-PCP and Post-PCP indicates that the prior distribution is suspect as described in Section 2.2, it was possible to distinguish these models ((4) and (8)) even in this small sample size. The results from Figure 2.3 ($n = 20$) to Figure 2.5 ($n = 100$) were similar to those of Figure 2.2 ($n = 5$). It meant that Prior- and Post-PCP of model and sample mean for the models with the prior distributions ((1)-(4)) with true mean were totally high, especially Prior- and Post-PCP for the models with the prior distribution (4) which has the weakest prior information were highest. Moreover, as the sample size increased, Post-PCP of sample mean for the model composed by the prior distribution (8) with no true prior mean and the weakest prior information increased, but on the other hand, Pre-PCP of sample mean for the model (8) decreased. This implied that the reliability for the prior distribution clarifies as sample size increased.

As a result, we gained a clear understanding of their characteristics. Main productive findings which were obtained in our research are as follow. For models with weak prior information, BPIC was more sensitive about model selection than PCP, so selection rates of correct model in BPIC were higher than those in PCP. For models with strong prior information, BPIC was as sensitive as PCP. Furthermore, when we evaluated models with weak and strong prior distributions

simultaneously, we got much the same PCP for models with weak and strong prior information including true prior mean, so we could not distinguish between them. On the other hand, BPIC chose models with strong prior information including true prior mean more than models with week prior information including true prior mean.

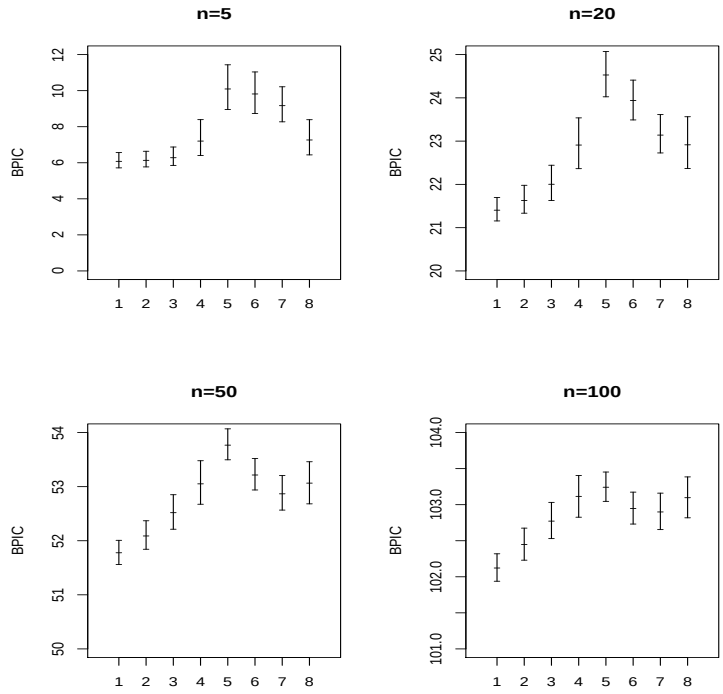


Figure 2.1: BPIC by sample size

2.4.2 Simulation (2)

Purpose

We focus on specifying the characteristics of BPIC here.

The purpose of simulation (2) is to assess the models with several different prior distributions in terms of prediction by calculating BPIC of these models from independent samples which follow true distribution and clarify the characteristics.

Method

We assume that true distribution is a normal distribution $\mathrm{N}[\mu, \sigma^2]$ with a known variance $\sigma^2 = 100$ and a mean prior parameter. We define several mean prior distributions taking into the amount of prior information and the prior mean. In detail, suppose that true value of mean

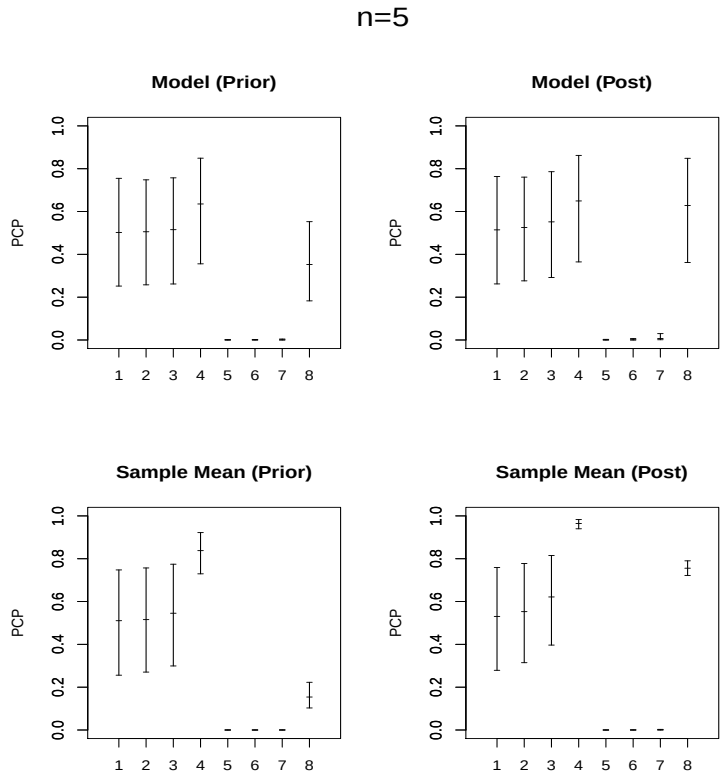


Figure 2.2: Prior- and Post- PCP for model and sample mean (n=5)

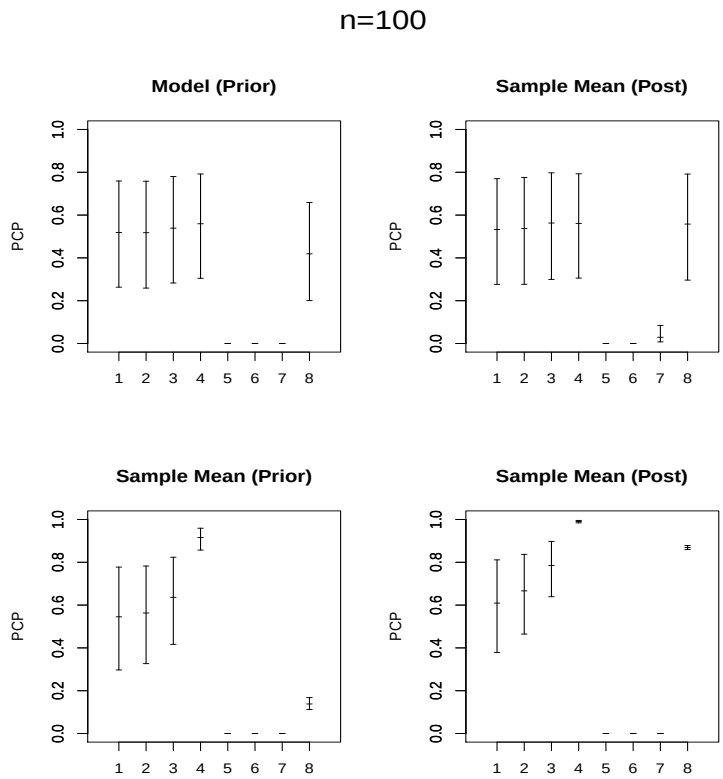


Figure 2.3: Prior- and Post- PCP for model and sample mean (n=20)

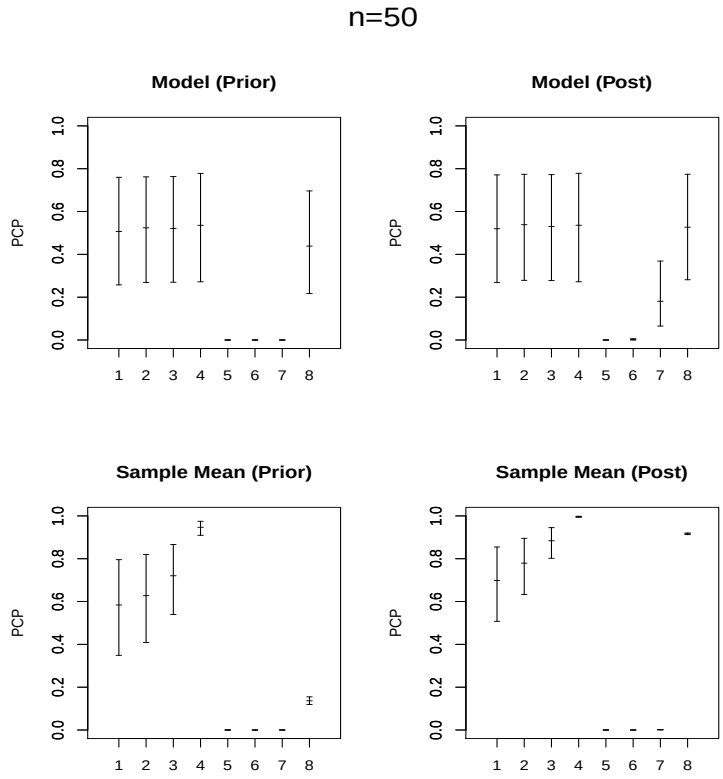


Figure 2.4: Prior- and Post- PCP for model and sample mean (n=50)

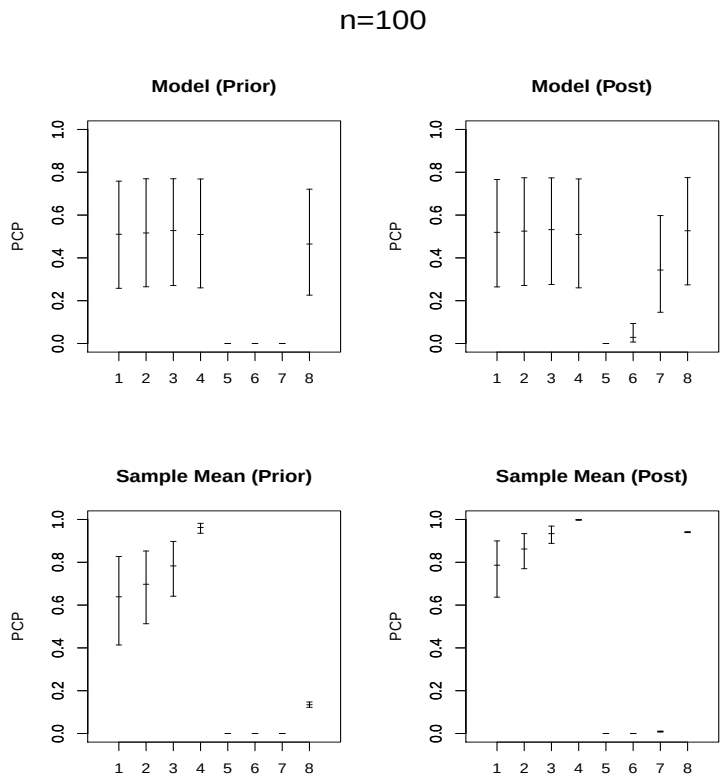


Figure 2.5: Prior- and Post- PCP for model and sample mean (n=100)

parameter μ is set at 0 and prior mean μ_0 set the following three values. When prior ES is $0, 0.5, 2$, prior mean μ_0 is $\mu_0 = 0, 5, 20$. Suppose that sample size is $n = 10, 100, 1000$ and variance σ_0^2 is calculated from the proportion of the prior sample $n_0/n = 0.001, 0.01, 0.1, 0.5, 1, 10, 50, 100, 1000$.

We generate the samples of sample size $n = 10, 100, 1000$ from true distribution N$[0, 100]$ for combination of prior distribution and sample size and calculate BPIC for the models composed by prior distribution. We repeat this process 10,000 times and summarize the results.

Result

For each sample size n, Figure 2.6 showed the results of 25%,50% and 75% points of BPIC. Actual line was the case of that prior ES was 0 ($\mu_0 = 0$), broken line was the case of that prior ES was 0.5 ($\mu_0 = 5$) and solid line was the case of that prior ES was 2 ($\mu_0 = 20$). When n_0/n was small by 0.1. Regardless of sample size n and prior ES, BPIC were almost same. However, when we took a notice on the difference between prior ES, BPIC with the case of that prior ES was 0 because larger than BPIC with the case of that prior ES was 0,5,2 as n_0/n increased. Also, Table 2.2 showed the main results of 50% of BPIC from Figure 2.6. This tables also show

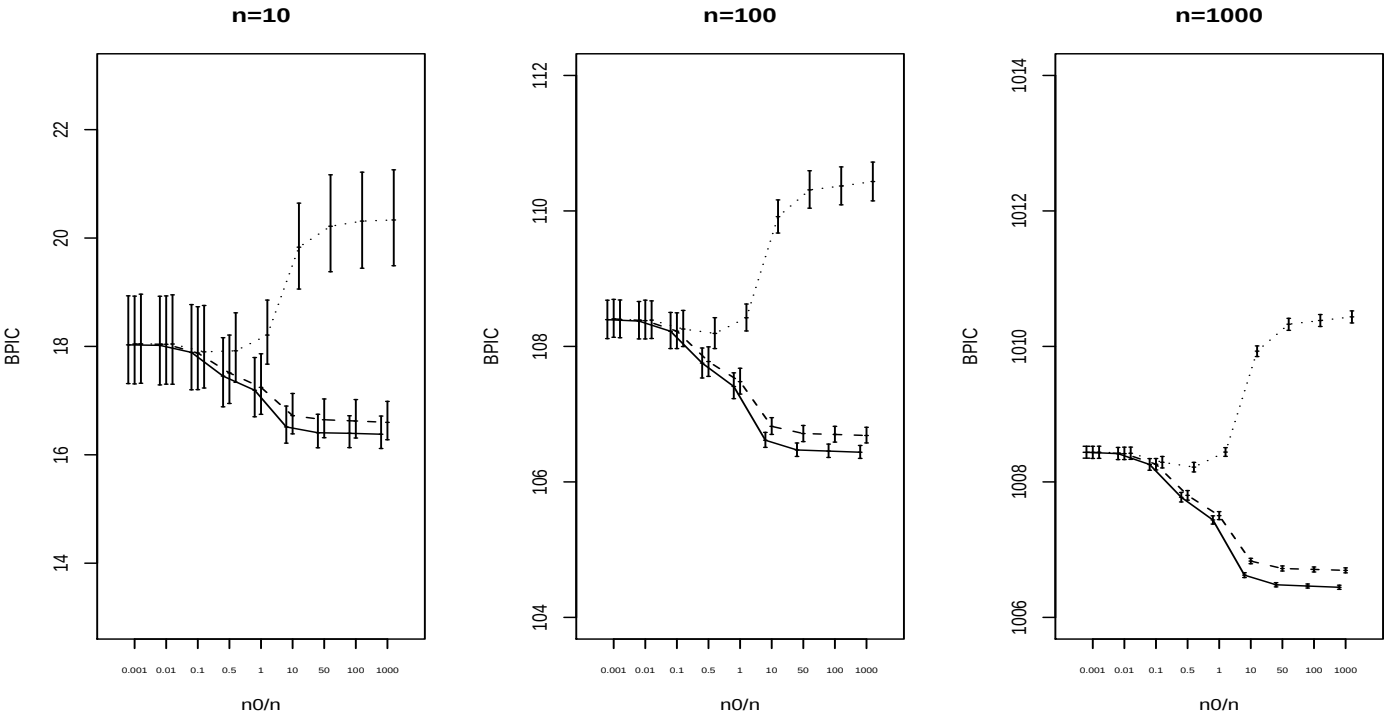


Figure 2.6: BPIC by proportion of prior sample (Prior ES:0[Actual], 0.5[Broken], 2[Solid])

Table 2.2: Summary of Figure 2.6 - 50% points of BPIC

n	n_0/n	prior ES	BPIC
10	0.001	0	18.1
		0.5	18.1
		2	18.5
	1	0	17.2
		0.5	17.3
		2	18.2
	1000	0	16.4
		0.5	16.6
		2	20.4
100	0.001	0	108.4
		0.5	108.4
		2	108.4
	1	0	107.4
		0.5	107.5
		2	108.4
	1000	0	106.4
		0.5	106.7
		2	110.4
1000	0.001	0	1008.4
		0.5	1008.4
		2	1008.4
	1	0	1007.4
		0.5	1007.5
		2	1008.4
	1000	0	1006.4
		0.5	1006.7
		2	1010.4

that BPIC became smaller as n_0/n increased when prior ES was 0.

Moreover, when prior ES was 0 and 0.5, because BPIC became small with high n_0/n regardless of sample size n, it implied that the model with strong prior information is preferable for BPIC. Therefore, as the figure indicated, we have to pay attention to selection of the prior distribution because the model with no true prior mean and strong prior information might be selected.

2.4.3 Simulation(3)

Purpose

Though we evaluated the impact of BPIC on model evaluation in Simulation(2), we focus on specifying the characteristics of the PCP here. Because PCA can express the PCP between 0 and 1 in any cases, we can evaluate the impact of sample size spontaneously.

Method

As well as Simulation(2), it is assumed that data follow true distribution $\mathrm{N}[0, 100]$, we set prior distribution $\mathrm{N}[\mu_0, \sigma_0^2]$ of mean parameter μ. And we set that prior ES is $0, 0.5, 2$, sample size is $10, 1000$ and σ_0^2 is calculated from the proportion of prior sample for sample size.

We generate the samples of sample size $n = 10, 1000$ from true distribution $\mathrm{N}[0, 100]$ for combination of prior distribution and sample size, and calculate Prior-PCP and Post-PCP for the models composed by prior distribution. We repeat this process 10,000 times and summarize the results.

Result

25%, 50% and 75% points of Prior-PCP and Post-PCP for model and sample mean were shown in Figure 2.7. In the case of that prior mean was true (prior ES was 0), Prior-PCP and Post-PCP for model and sample mean were almost same values within the same proportion of prior sample regardless of sample size n. However, strictly the difference between Prior-PCP and Post-PCP was about 0.1 at maximum when we compare 50% points of Prior-PCP with those of Post-PCP. Then sample size was 10. Also, PCP for model showed the almost same values with broad range at the same sample size n regardless of n_0/n (proportion of prior sample). However even in this case, strictly PCP increased as n_0/n decreased when we compared Prior-PCP and Post-PCP between n_0/n, and the difference was about 0.1 at maximum. Also, PCP for sample mean showed the high value with small n_0/n. These results implied that the models with weak prior information were preferable in the models with true mean at the evaluation of PCP for sample mean.

In the case of that the prior ES was 0.5 or 2 (prior mean was not true), Prior-PCP and Post-PCP for model and sample mean were low when n_0/n (proportion of prior sample) was high and n was large. When we compared different prior ES at the same n_0/n and n, Prior-PCP and Post-PCP for model and sample mean were almost same at the lowest n_0/n and $n = 10$. However, except for the case, totally PCP with the case of that prior ES was 0 were lower than PCP with the case of that prior ES was not 0. Moreover, it is possible in predictive checking approach to conduct not only the comparison between the models but also diagnosis for one model by comparing Prior-PCP with Post-PCP. In the case of that the prior ES was 0.5 and 2, the difference between Prior-PCP and Post-PCP for sample mean increased as n_0/n increased,

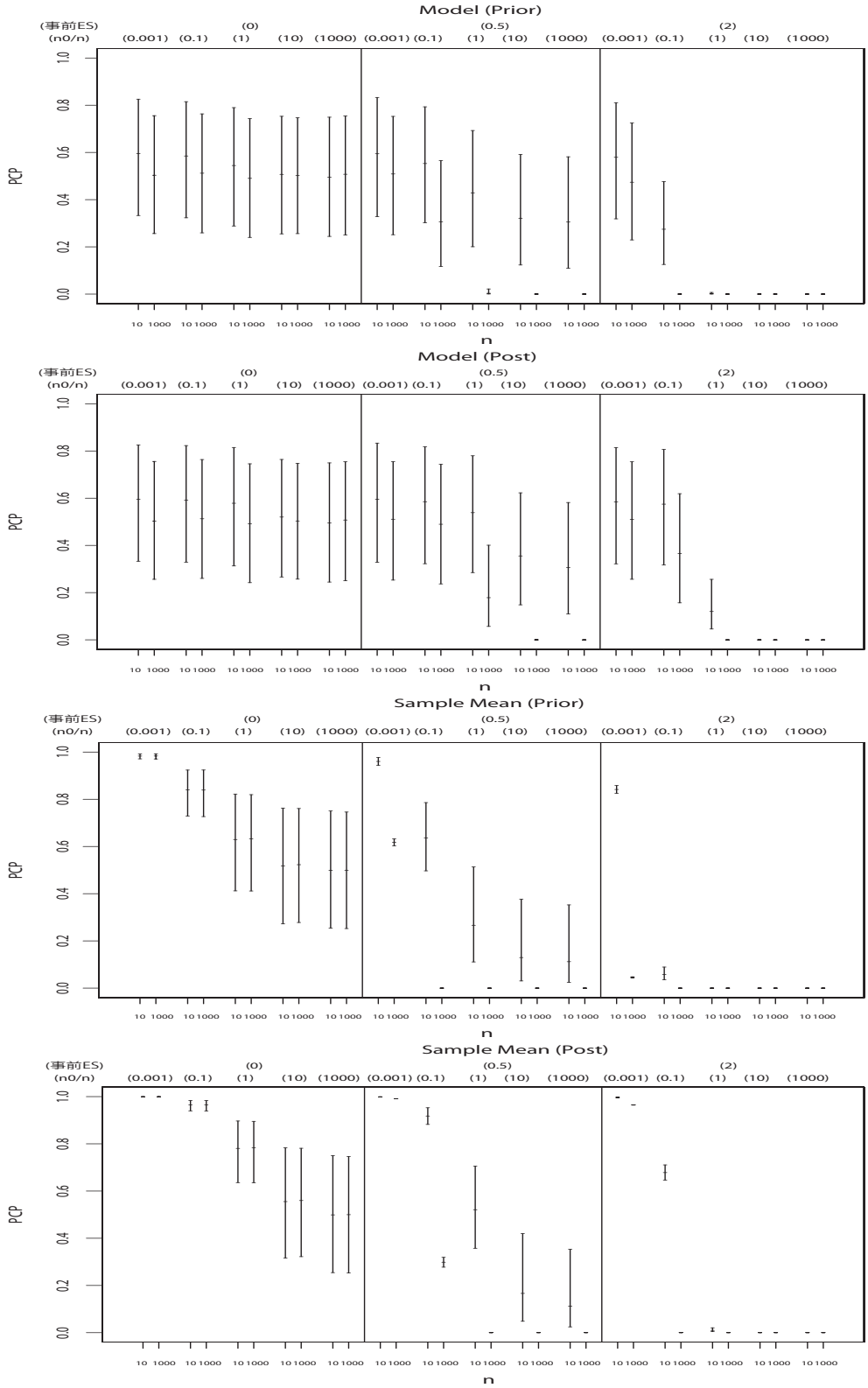


Figure 2.7: Prior-PCP and Post-PCP for model and sample mean by proportion of prior sample

Table 2.3: Summary of Figure 2.7 - 50% points of Prior-PCP and Post-PCP for model and sample mean

Prior-ES	n_0/n	n	Prior-model	Post-model	Prior-mean	Post-mean
0	0.001	10	0.593	0.593	0.983	1
		100	0.536	0.536	0.983	1
		1000	0.506	0.506	0.983	1
	1	10	0.545	0.579	0.641	0.788
		100	0.514	0.523	0.629	0.780
		1000	0.512	0.514	0.617	0.991
	1000	10	0.502	0.502	0.515	0.515
		100	0.499	0.499	0.502	0.502
		1000	0.502	0.502	0.500	0.500
0.5	0.001	10	0.588	0.588	0.960	1
		100	0.528	0.529	0.874	0.997
		1000	0.511	0.514	0.617	0.991
	1	10	0.428	0.536	0.264	0.519
		100	0.208	0.405	0.000	0.042
		1000	0.004	0.183	0	0
	1000	10	0.317	0.317	0.121	0.122
		100	0.049	0.049	0	0
		1000	0	0	0	0
2	0.001	10	0.596	0.600	0.841	0.996
		100	0.512	0.514	0.617	0.991
		1000	0.470	0.506	0.046	0.964
	1	10	0.001	0.121	0.000	0.001
		100	0	0	0	0
		1000	0	0	0	0
	1000	10	0.000	0.000	0.000	0.000
		100	0	0	0	0
		1000	0	0	0	0

($n = 100$ is newly included.)

especially more than 0.1. It implied that the models were suspicious. The difference made clear as n_0/n increased. Also, Table 2.3 showed the main results of 50% points of Prior-PCP and Post-PCP from Figure 2.7. It also included the cases of $n = 100$ which did not be shown in Figure 2.7. Again from the results of $n = 100$, it implied that the models with weak prior information were preferable in the models with true prior mean at the evaluation of PCP for sample mean.

Therefore when we conduct model diagnoses in the framework of PCA, we have to take notice on that it is possible to give high PCP for models with weak prior information. Then we pay attention to the difference between Prior-PCP and Post-PCP and judge the appropriateness of the model. These results are different from the findings of BPIC obtained from Simulation(2)

(select the models with strong prior information).

Moreover, at the evaluation of PCP for sample mean, we considered about a reason that Prior-PCP and Post-PCP were high for the models with small prior information. From (2.12) which calculate Prior-PCP, it was found that the variance of prior predictive distribution which sample mean $\bar{y}$ follow is larger in the models with weak prior information than in the models with strong prior information. So for the models with weak prior information, the values of the standardization approach to 0 and Prior-PCP becomes large. Also, from (2.14) which calculate Post-PCP, it was found that the variance of posterior predictive distribution which sample mean $\bar{y}$ follow is larger in the models with weak prior information than in the models with strong prior information, and the mean of posterior predictive distribution approaches to sample mean $\bar{\tilde{y}}$. So for the models with weak prior information, the values of the standardization approach to 0 and Post-PCP becomes large.

2.5 Conclusion

In this chapter, we focused on BPIC and PCA which evaluate models from the position of the prediction and conducted some simulations as a purpose of clarifying the features of these model diagnoses. Through our simulations, we found that regardless of whether prior mean is true or not, totally Bayesian predictive information criterion has low values in the cases of the models with strong prior information, and predictive checking probability has high value in the case of the models with weak prior information. Therefore, Bayesian predictive information criterion may select the model with strong prior information and no true prior mean than the model with weak prior information and true prior mean. Also, the predictive checking approach preferred the model with weak prior information and no true prior mean to the model with strong prior information and true prior mean in some cases of the situation defined in Simulation(2). So we have to have the findings in mind, by taking notice on the difference between prior and posterior predictive checking probability and calculating the predictive checking probability of specially interesting indices, it is very important to judge whether the model appropriately expresses the interesting occasions or not.

Though Bayesian predictive information criterion and predictive checking approach applied for relatively simple occasions here, the results were different clealy. Actually, we have to diagnose some models under various situations, however it is possible for even the cases of that

we treat another distributions and more than two parameters to capture the characteristics of these Bayesian predictive model diagnoses through the similar simulation with section 2.4. It is important to specify the characteristics of Bayesian predictive model diagnoses previously and it leads the improvement of the model selections.

Reference

1. Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, **94**, 443-458.

2. Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modeling and robustness. *J. Roy. Statist. Soc.*, A**153**, 383-430.

3. Daimon, T. and Goto, M. (2007). Predictive checking approach to Bayesian interim monitoring. *Japanese J.Appl. Statis.*, **36**(2 & 3), 119-137.

4. Gelman, A., Meng, X. L. and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, **6**, 733-807.

5. Guttman, I. (1967). The use of a concept of a future observation in goodness-of-fit problems. *J. Roy. Statist. Soc.*, B**29**, 83-100.

6. Konishi, S., Ando, T. and Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika*, **91**, 27-43.

7. Mori, H. (2010). Evaluation of prior information in Bayesian Inference. *Journal of the Japan Statistical Society*, **40**(1), 1-22 (in Japanese).

8. Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, **12**, 1151-1172.

9. Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

10. Tiao, G.C. and Xu, D. (1993). Robustness of maximum likelihood estimates for multi-step predictions; the exponential smoothing case. *Biometrika*, **80**, 623-641.

11. Wood, P. D. (1973). Personal communication.

3. A preliminary evaluation about health guidance

3.1 Introduction

A prevalence and preliminary of "lifestyle-related disease (adult disease)" increase as the lifestyle habit changes and elderly people increases. "lifestyle-related disease" is all-inclusive term of diseases due to lifestyle such as a smoking, diet, drinking, exercise and sleep (Display.3.1), and as the risk factor of lifestyle-related disease such as hypercholesteremia piles up, the incidence of cerebral stroke and ischemic cardiac disease increases.

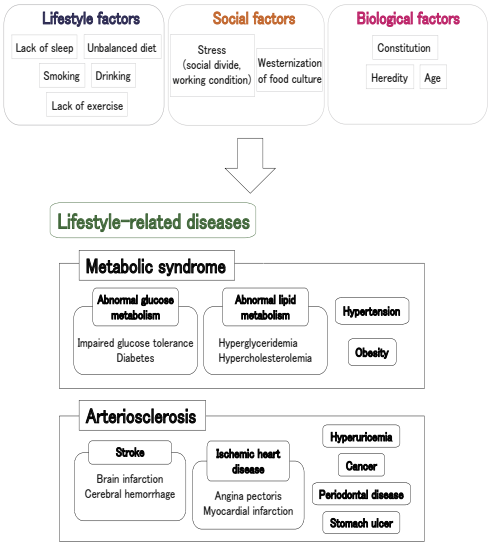
Since April 2008, Ministry of Health, Labor and Welfare of Japan has carried out "Health Checkups and Healthcare Advice" which make it obligatory for person aged 40 through 74 to reduce medical expenses and prevent lifestyle-related diseases (Health Service Bureau of Health, Labour and Welfare, 2007). Though the aim of "Health Checkups and Healthcare Advice" is for "Reduction of medical cost" and the prevention of the lifestyle-related disease, it is concerned with the lack of "Foundation for enforcement" and "evidence for Prevention" (Ohgushi, 2006: 2007). Also as Kondo (2004) indicates a lack of foundation for health checkup, we wonder about a meaning of making health checkup compulsory, too.

In this chapter, we treat data of the health checkup aims to prevent disease was carried out in April 2004 and a doctor classified subjects into uncontrolled, directed (teaching of better living) and clinical group (includes medicine), based on their results. And the teaching of better living or treatment was conducted for the directed and clinical groups and an improvement of the clinical test results was examined after that. Here the definition for the clinical group were based on the constant criterion value, the definition for the directed group were based on the judgment of the doctor. Based on this data, we explored foundation about the doctor's judgment, especially classification of the directed group, attempting to figure the doctor's character, and further evaluated directed effect for the directed group. In section 3.2, we summarize the purpose of the analysis conducted for this data and make a clear our motivation. In section 3.3, we examine

the effect of the doctor's teaching of better living. In section 3.4, we conduct the statistical diagnoses for the analysis results and consider the stablity of the results. Finally, section 3.5 contains our concluding remarks.

3.2 Analysis for the data of the health checkup

The laboratory test items used in this paper were Weight, BMI, Systolic blood pressure, Diastolic blood pressure, Total cholesterol (TC), Triglyceride (TG), High-density lipoprotein (HDL). A doctor classified subjects into uncontrolled, directed (teaching of better living) and clinical group (includes medicine), based on their results. After that, the teaching of better living or treatment was conducted for the directed and clinical groups. The subjects were 1,141 (Male 543, Female 598). As mentioned before, the definitions for the clinical group were based on the constant criterion value (Systolic blood pressure: >=160 and Diastolic blood pressure: >=100, TC: <90 or >=260, TG: >=250, HDL: <=25), the definition for the directed group were based on the judgment of the doctor. However, because 11 subjects of the total subjects were missing in more than one of TC, TG and HDL, these subjects were excluded from the analysis set in this paper. Because one subject in the uncontrolled group and three subjects in the directed group had the measurements which exceed the criterion values, these subjects were also excluded. Moreover, any clinical test results for 7 subjects in clinical groups did not



Display 3.1. Diagram related to lifestyle diseases

Display 3.2. Subject profile

Pr	Sex	Year	Uncontrolled	Directed	Clinical
1	Male	10〜30	76	3(3)	2
2	Male	30〜40	119	26(14)	14
3	Male	40〜50	81	45(34)	19
4	Male	50〜	82	42(38)	28
5	Female	10〜30	149	3(2)	1
6	Female	30〜40	100	3(3)	2
7	Female	40〜50	96	14(12)	4
8	Female	50〜	145	32(28)	34
		Male	358	116(89)	63
		Female	489	52(45)	41

(∗) The number of the subjects who had the health checkup for follow-up after 4 months.

meet the constant criterion value, so these subjects were also excluded. Subject profile (Sex, Year) in the uncontrolled, directed and clinical group was indicated in Display 3.2. The number in parentheses of the directed group expressed the number of the subjects who had the Health Checkup for follow-up after 4 months. Also, of subjects in the clinical group, those with the blood pressure more than the constant criterion values were 2, those with TC and TG more than the constant criterion values were 60 and 44.

The aims of the analysis in this paper are the following.

1 Explore foundation about the doctor's judgment for especially classification of directed group, attempting to figure the doctor's character.

2 Evaluate directed effect for subjects in the directed group.

3 Conduct the statistical diagnosis for the models used in this paper in the purpose of examining the appropriateness of the analysis results.

To achieve these purposes, following to Maruo, Shirahata, Goto and Komazawa (2008), we take note of preserving a logic consistency in the overall flow of the analysis.

3.3 A process in statistical data analysis

We assume that the clinical laboratory test result measured in the health checkup follow a power-normal distribution because the clinical laboratory test is generally positive. Then, a diagnosis of an outlier (Sample diagnosis) is conducted in the following method: 1. A method that observations beyond sample mean $\pm$ 3SD are excluded. 2. Data-adaptive probability plot (Shimokawa and Goto, 2002) 3. A method based on Dixon ratio ([Absolute deviation between largest (smallest) and second largest (smallest) observation]/[Range of total observations including extreme value]) (Dixon, 1953). Method 1 is a traditional evaluation method for an outlier. Method 2 is visually an evaluable method whether the data merely exists in the tailed parts of the distribution or is an outlier. Method 3 is an evaluation method which is suggested in a guideline of National Committee for Laboratory Standard: NCCLS, current CLSI (Clinical and Laboratory Standards Institute) (Sasse, Doumas, Miller, D'Orazio, Eckfeldt, Evans, Graham, Myers, Parsons and Stanton, 2000) and eliminate the extreme observation when Dixon ratio is over 1/3. After the diagnosis of an outlier, by Classification and Regression Tree (CART) (Breiman, Friedman, Olshen and Stone, 1984: Sugimoto, Shimokawa and Goto, 2005) which optimally find explanatory variables which have an effect on data to capture interaction and nonlinear effect of explanatory variable and data-adaptive discriminant analysis (Hatanaka, Inoue and Goto, 1981: Seo, Shimokawa, Daimon and Goto, 2002: Shimokawa and Goto, 2004) which are known as indicating high correct discriminant ratio when data follow multi-variable power-normal distribution, we explore the clinical test items which contributes to the clarification between the uncontrolled group and the directed group. Moreover, when it is assumed that clinical test results before and after the doctor's teaching of better living for subject characteristics follow bivariate power-normal distribution (Goto and Hamasaki, 2002), we identify the shape of the distribution and the extend of variation before and after the doctor's teaching, and evaluate directed effect for the directed group.

3.3.1 Power-normal distribution

As the distribution which the clinical test items follow, we set a power-normal distribution. A power-transformation of positive variable X is defined as

$$X^{(\lambda)} = \begin{cases} (X^{\lambda} - 1)/\lambda, & \lambda \neq 0 \\ \log X, & \lambda = 0 \end{cases} \tag{3.1}$$

(Box and Cox, 1964). Aiming to the normality of the transformed variables $\{X^{(\lambda)}\}$, the power-normal distribution was proposed as the distribution of the observation X on the original scale when assuming the linearity(the additivity) of model on $\{X^{(\lambda)}\}$ and the uniformity of variance (Goto, Uesaka and Inoue, 1979 : Goto, Matsubara and Tsuchiya, 1983). The probability density function is

$$f_{\mathrm{PN}}(x; \lambda, \mu, \sigma) = \begin{cases} x^{\lambda-1} \phi\left\{(x^{(\lambda)} - \mu)/\sigma\right\} / A(\lambda, \mu, \sigma), & x > 0 \\ 0, & x \leq 0, \end{cases} \tag{3.2}$$

where $\phi(\cdot)$ is a probability density function of standard normal distribution and $A(\lambda, \mu, \sigma)$ is a probability proportionality constant term

$$A(\lambda, \mu, \sigma) = \begin{cases} \Phi\left\{-(\lambda\mu + 1)/\lambda\sigma\right\}, & \lambda < 0 \\ 1, & \lambda = 0 \\ \Phi\left\{(\lambda\mu + 1)/\lambda\sigma\right\} & \lambda > 0, \end{cases} \tag{3.3}$$

where $\Phi(\cdot)$ is a cumulative distribution function of standard normal distribution. λ, μ and σ are respectively the parameter of shape, location and scale. By changing λ according to the observation X on the original scale, the power-normal distribution include several distributions. The power-normal distribution with $\lambda = 1$ expresses normal distribution, The power-normal distribution with $\lambda = 0$ expresses log-normal distribution. The main advantages of using the power-normal distribution are able to comprehend the discrepancies between ideal (model, hypothesis) and reality (data) appropriately and conduct data-adaptive analysis, and be also available for a lot of traditional methods based on normal distribution.

When the observation $X_1, X_2, \ldots, X_n$ follow the power-normal distribution f_{PN} independently, the log-likelihood is expressed as

$$l_{\mathrm{PN}}(\lambda, \mu, \sigma) = -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i^{(\lambda)} - \mu)^2 + (\lambda - 1)\sum_{i=1}^{n}\log x_i - n\log A(\lambda, \mu, \sigma). \tag{3.4}$$

Because it is generally difficult to estimate the parameter considered to $A(\lambda, \mu, \sigma)$, referring to the estimation method of Box and Cox (1964), we set $A(\lambda, \mu, \sigma) = 1$ and calculate the maximum likelihood estimator of μ and σ^2 as $A(\lambda, \mu, \sigma) = 1$ from

$$\hat{\mu}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} x_i^{(\lambda)}, \quad \hat{\sigma}^2(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(x_i^{(\lambda)} - \hat{\mu}(\lambda))^2. \tag{3.5}$$

In replacing (3.5) to (3.4), the log-likelihood can be expressed as the function of λ. So we can get the maximum likelihood estimator $\hat{\lambda}$ of λ based on the Newton-Raphson method. Moreover, in replacing $\hat{\lambda}$ to (3.5), the maximum likelihood estimators $\hat{\mu}(\hat{\lambda}),\ \hat{\sigma}(\hat{\lambda})$ of $\mu,\ \sigma^2$ given $\lambda = \hat{\lambda}$ can be calculated.

When the power-transformed observations $\boldsymbol{x}^{(\boldsymbol{\lambda})} = (x_1^{(\lambda_1)}, x_2^{(\lambda_2)}, \ldots, x_p^{(\lambda_p)})^{\mathrm{T}})$ for the non-negative p-variate observations $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^{\mathrm{T}}$ follow the p-variate normal distribution approximately, the p-variate power-normal distribution is defined as the distribution which the observation $\boldsymbol{x}$ before power-transformation follow (Goto *et al.*, 1979 : Hatanaka *et al.*, 1981: Shimokawa and Goto, 2004). A probability density function $f_{\mathrm{MPN}}(\boldsymbol{x}|\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ of $\boldsymbol{x}$ is given in

$$f_{\mathrm{MPN}}(\boldsymbol{x}|\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\prod_{i=1}^{p} x_i^{\lambda_i - 1}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|A(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\Sigma})} \times \exp\left\{-\frac{1}{2}(\boldsymbol{x}^{(\boldsymbol{\lambda})} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^{(\boldsymbol{\lambda})} - \boldsymbol{\mu})\right\}, \tag{3.6}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_p)^{\mathrm{T}}$ is $p \times 1$ power parameter vector, and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are respectively mean vector and variance-covariance matrix when the transformed $\boldsymbol{z}$ follow p-variate normal distribution approximately. $A(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a probability proportionality constant term

$$A(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_{\mathrm{R}} \cdots \int \frac{1}{(2\pi)^{p/2}} \exp(-\frac{1}{2}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}) dv_1 \cdots dv_p. \tag{3.7}$$

Here, $\boldsymbol{v} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x}^{(\boldsymbol{\lambda})} - \boldsymbol{\mu})$ is $p \times 1$ probability vector. $\mathrm{R} = \{\boldsymbol{v} : \boldsymbol{x} > 0\}$ is the integrated range. As the same in the single variable case, it is difficult to estimate the parameters considered to $A(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, so we estimate the parameters based on the Newton-Raphson method as $A(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = 1$.

3.3.2 Data-adaptive discriminant analysis

In this section, we explain about data-adaptive discriminant analysis used for exploring the clinical test items which contribute to the classification of uncontrolled and directed group. It is supposed that the non-negative p-variate observations $\{x_{li}\}_{i=1}^{n_l}$ are generated from two p-variate power-normal population $\Pi_l (l = 1$: uncontrolled group, $l = 2$: directed group), where n_l is

the subject's number included in Π_l. When $\boldsymbol{\lambda}_l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l$ is known, a data-adaptive discriminant function is given by

$$\begin{aligned} g(\boldsymbol{x}) =& \log f_{\mathrm{MPN}}(\boldsymbol{x}|\boldsymbol{\lambda}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_1) - \log f_{\mathrm{MPN}}(\boldsymbol{x}|\boldsymbol{\lambda}_2, \boldsymbol{\Sigma}_2, \boldsymbol{\mu}_2) \\ =& \frac{1}{2}\left\{(\boldsymbol{x}^{(\boldsymbol{\lambda}_2)} - \boldsymbol{\mu}_2)\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}^{(\boldsymbol{\lambda}_2)} - \boldsymbol{\mu}_2) - (\boldsymbol{x}^{(\boldsymbol{\lambda}_1)} - \boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}^{(\boldsymbol{\lambda}_1)} - \boldsymbol{\mu}_1)\right\} \\ &+ \sum_{i=1}^{p}(\lambda_{1i} - \lambda_{2i})\frac{1}{2}\log x_i - \log\frac{|\boldsymbol{\Sigma_2}|}{|\boldsymbol{\Sigma_1}|}. \end{aligned} \tag{3.8}$$

Then, for newly obtained $\boldsymbol{x}$, the subject is clarified to Π_1 in the case of $g(\boldsymbol{x}) > 0$ and Π_2 in the case of $g(\boldsymbol{x}) < 0$. Because $\boldsymbol{\lambda}_l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l$ are unknown, they are replaced to the maximum likelihood estimator normally.

Also, the effect that p_r of p exploratory variables contribute to the discrimination is evaluated using Are Under Curve (AUC) of ROC curve. Given group variable l ($\Pi_l; l = 1, 2$) and exploratory variable vector $\boldsymbol{x}$, A sensitivity $F_{\mathrm{TP}}(g, u)$ and a specificity $F_{\mathrm{TN}}(g, u)$ are respectively given by

$$\begin{aligned} F_{\mathrm{TP}}(g, u) &= \Pr(g(\boldsymbol{x}) > \mu | l = 1), \\ F_{\mathrm{TN}}(g, u) &= \Pr(g(\boldsymbol{x}) < \mu | l = 2). \end{aligned}$$

ROC curve is obtained by plotting $(F_{\mathrm{TP}}(g, u), 1 - F_{\mathrm{TN}}(g, u))$ for any u $(-\infty < u < \infty)$. Moreover, AUC is calculate from

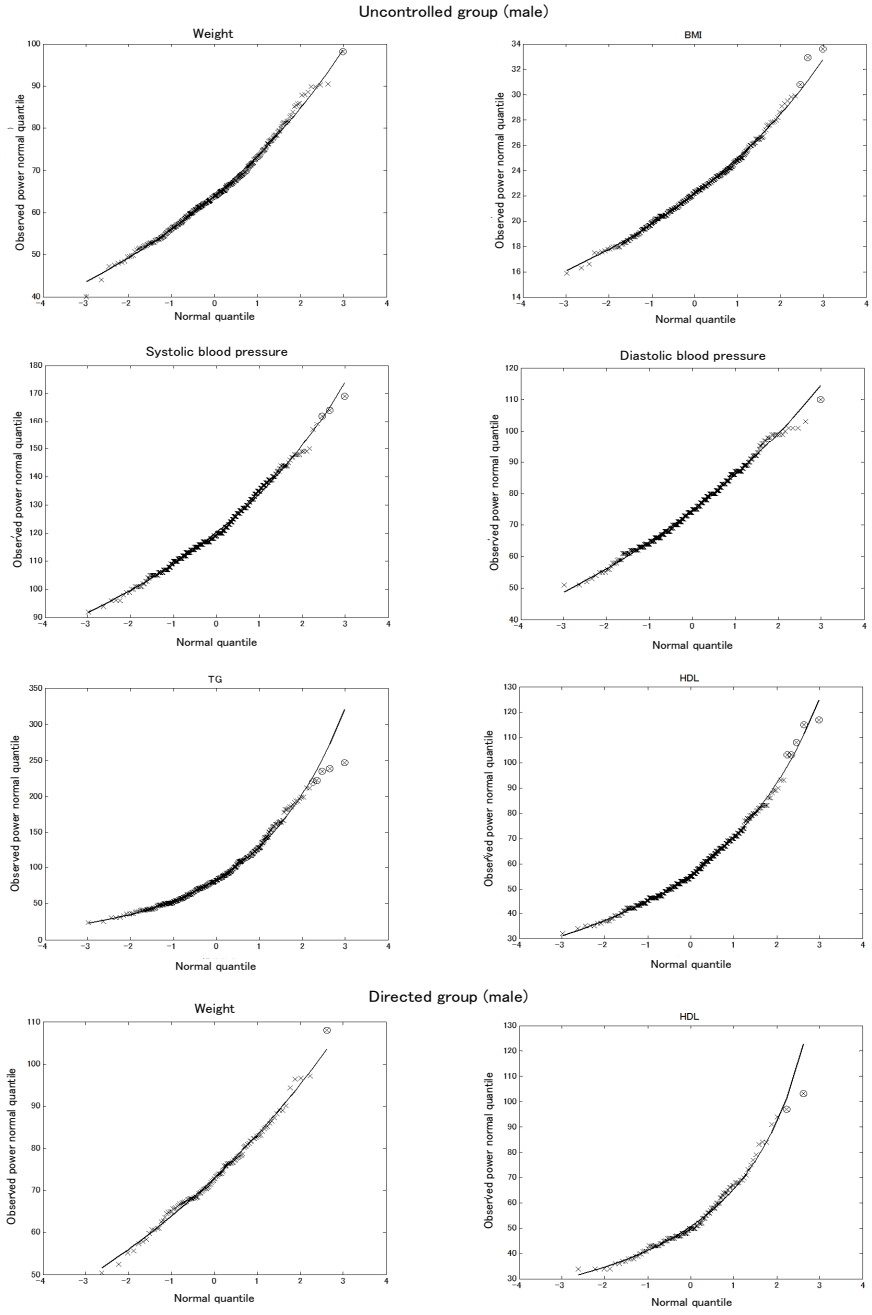
$$\mathrm{AUC} = \int_{u=-\infty}^{\infty} F_{\mathrm{TP}}(g, u) dF_{\mathrm{FP}}(g, u), \tag{3.9}$$

where $F_{\mathrm{FP}}(g, u) = 1 - F_{\mathrm{TN}}(g, u)$.

3.3.3 Exploration of clinical test items which contributes to the classification of uncontrolled and directed group

Diagnosis of outliers: Data-adaptive discriminant plots for the clinical test results in uncontrolled and directed group by sex were in Display 3.3 and Display 3.4. A circle in Display 3.3 and Display 3.4 points the observation eliminated in Method 1. However, it found that most observations eliminated in Method 1 exist in the tailed area of the distribution. The maximum value of HDL in Female (uncontrolled group) was apart from the transformation curve, but because Dixon ratio based on this observation was $0.179 < 1/3$, it was impossible to be judged as an outlier from Dixon ratio. From the above results, we use all observation in the following analyses without any removal.

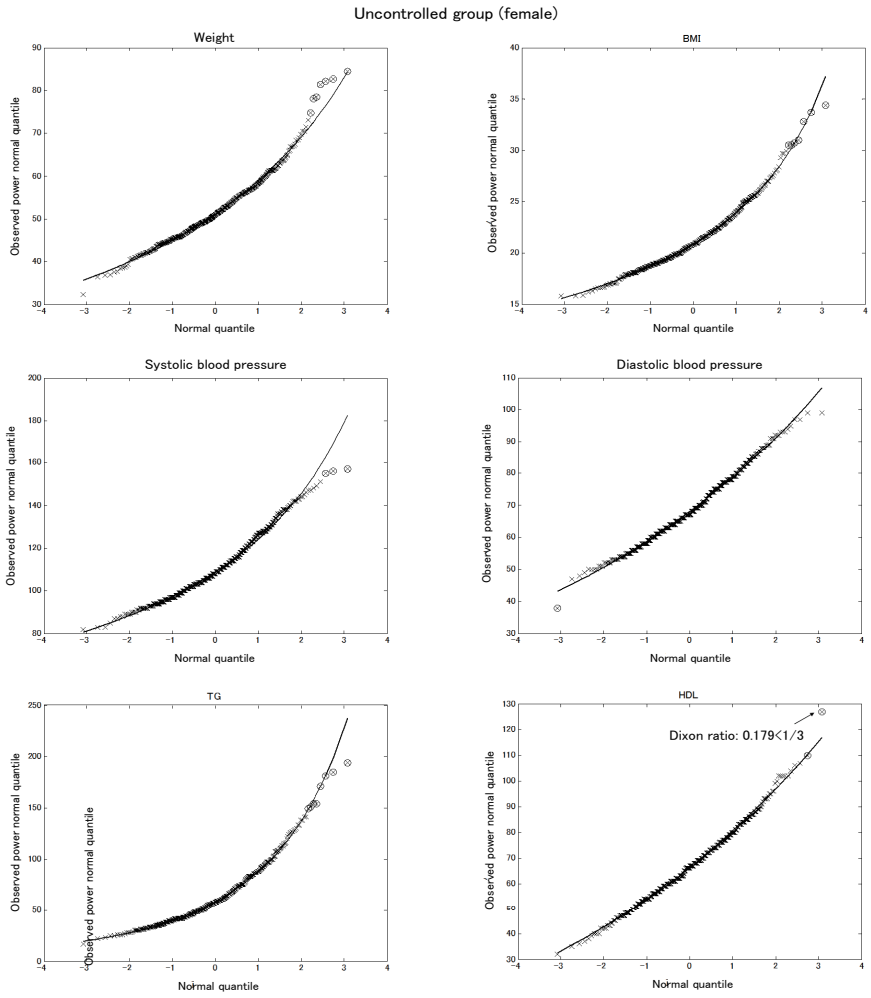
CART method: Using CART method where the explanatory variables were Weight, BMI, SBP, DBP, TC, TG, HDL, Age and Sex, we investigated a divergence pattern between uncontrolled and directed group (Display 3.5, Display 3.6). We used a cross-validation for selection of the optimum tree. As the result, the branches of the tree were in order of BMI, Weight and Age. So the classification expressed the feature of the body type clearly. However, the clinical test results in directed group might be actually affected by the criterion value because directed



Display 3.3. Diagnosis of an outlier: Male

and clinical group are divided by the criterion value. Then the misclassification rate for the uncontrolled and directed group was 13.69%. We investigated a divergence pattern between the uncontrolled and the directed/clinical group (Display 3.7). From the result, we found that not only BMI, weight and age but also TC and TG existed in the divergence pattern. Incidentally TC or TG in the clinical group exceeded the criterion value for most subjects. The classification result of TC was almost the same with the criterion value (>= 260) for clinical group, the classification result of TG was lower than the criterion value (>= 250). Therefore, the doctor might take account of TG when the doctor classified subjects into directed group.

Data-adaptive discriminant analysis: To confirm the above classification results, we conducted data-adaptive discriminant analysis. For Weight, BMI, SBP, DBP, TC, TG and HDL, we explored and evaluated the clinical test item which contributed to the classification



Display 3.4. Diagnosis of an outlier：Female

Display 3.5. Subject profile extracted in CART (Classification for uncontrolled and directed group)

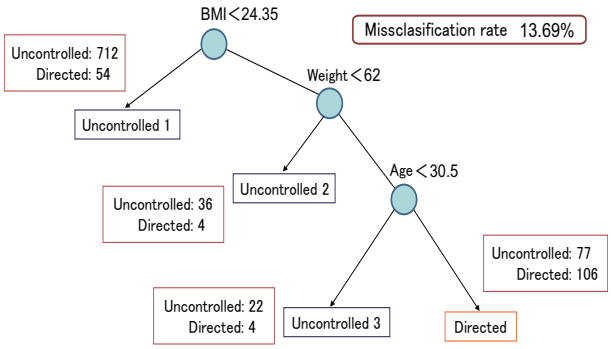
Termination node	BMI		Weight		Age		uncontrolled group	directed group
	< 24.35	$\geqq 24.35$	< 62	$\geqq 62$	< 30.5	$\geqq 30.5$		
uncontrolled 1	○						712	54
uncontrolled 2		○	○				36	4
uncontrolled 3		○		○	○		22	4
directed		○		○		○	77	106

between the uncontrolled and directed group. Though the sex did not express in the divergence pattern in the CART method, we referred to the analysis results that the factor of sex have an effect on TC, TG and HDL in Maruo *et al* (2008), and conducted data-adaptive discriminant analysis by sex. The correct classification rates in the case of using the above seven clinical test items were Male 76.0% (uncontrolled group: 74.3%, directed group: 81.0%), Female 83.6% (uncontrolled group: 83.0%, directed group: 88.5%). Also, the highest combination of variables in AUC of ROC curve were BMI, SBP, TC in Male and BMI, DBP, TC in Female (Display 3.8). We conducted data-adaptive discriminant analysis using the selected variables. Then the correct classification rates were Male 76.0% (uncontrolled group: 73.5%, directed group: 83.6%), Female 81.2% (Uncontrolled group: 80.2%, Directed group: 90.4%). Therefore, the doctor might have clearer judgment criterion for Female than Male. Again, to measure the effect on the censored value of clinical test items in the directed group, we conducted data-adaptive discriminant analysis for the uncontrolled and directed/clinical group. The correct classification rates in the case of using the above seven clinical test items were Male 81.2% (Uncontrolled group: 83.8%, Directed group: 95.7%), Female 85.7% (Uncontrolled group: 83.8%, Directed group: 95.7%), and these rates were higher than those for the uncontrolled and directed group. Again, we conducted data-adaptive discriminant analysis using BMI, SBP, TC in Male and BMI, DBP, TC in Female. As the results, the correct classification rates were Male 76.4% (Uncontrolled group: 74.3%, Directed group: 80.5%), Female 84.2% (Uncontrolled group: 83.0%, Directed group: 90.3%) and these rates were also higher than those for uncontrolled and directed group.

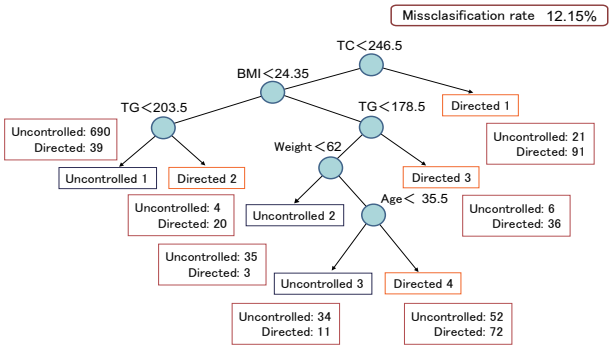
From the result in CART and data-adaptive discrimitant analysis, we can consider that the doctor mainly provided guidance about body type and blood type for the subjects in directed group.

3.3.4 The shape of the distribution and the change for the clinical test result before and after direction

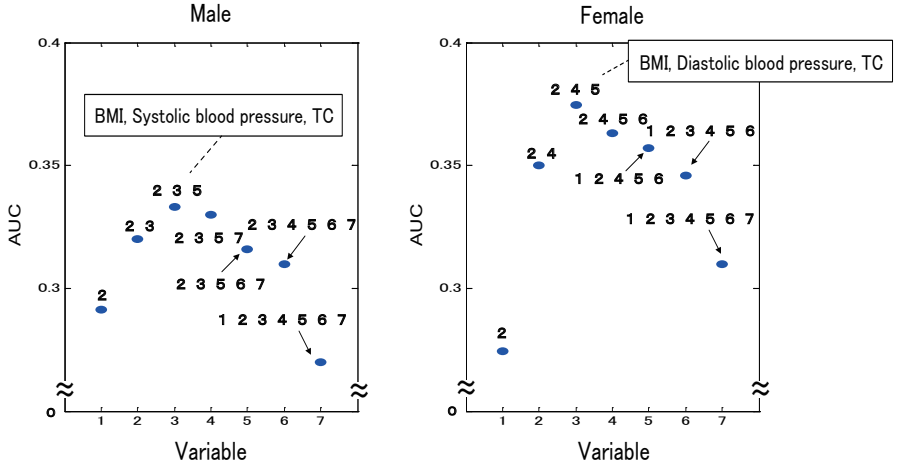
To measure the direction effect of the doctor for the subjects classified to directed group, we assume that the pre- and post-observations in the clinical test item follow bivariate power-normal distribution. In Display 3.9, the estimators of the power-parameter $\hat{\lambda}_{\mathrm{Pre}}, \hat{\lambda}_{\mathrm{Post}}$ for pre- and post-observations were showed. From the results, as generally considered, the shapes of the distribution which pre- and post-observations were almost same, but the estimated power-parameters of pre-observations for DBP and TC in Male, HDL in Female were much different from those of post-observations and it implied that the observations before and after direction



Display 3.6. Classification for uncontrolled and directed group in CART



Display 3.7. Classification uncontrolled and directed/clinical group in CART



Display 3.8. Classification for uncontrolled and directed group: AUC (1.Weight, 2.BMI, 3.SBP, 4.DBP, 5.TC, 6.TG, 7.HDL)

Display 3.9. Estimators of the power-parameter before and after the doctor's direction: directed group

	Weight		BMI		SBP		DBP		TC		TG		HDL	
	$\hat{\lambda}_{\text{Pre}}$	$\hat{\lambda}_{\text{Post}}$	$\hat{\lambda}_{\text{Pre}}$	$\hat{\lambda}_{\text{Post}}$	$\hat{\lambda}_{\text{Pre}}$	$\hat{\lambda}_{\text{Post}}$	$\hat{\lambda}_{\text{Pre}}$	$\hat{\lambda}_{\text{Post}}$	$\hat{\lambda}_{\text{Pre}}$	$\hat{\lambda}_{\text{Post}}$	$\hat{\lambda}_{\text{Pre}}$	$\hat{\lambda}_{\text{Post}}$	$\hat{\lambda}_{\text{Pre}}$	$\hat{\lambda}_{\text{Post}}$
Male	0.31	0.29	0.23	0.15	0.49	0.81	1.11	2.04	1.54	0.65	0.04	-0.11	-0.51	-0.16
Female	-0.21	-0.41	-0.19	-0.39	0.59	0.42	1.51	1.71	1.46	1.50	-0.35	-0.22	0.70	0.17

followed separate distributions.

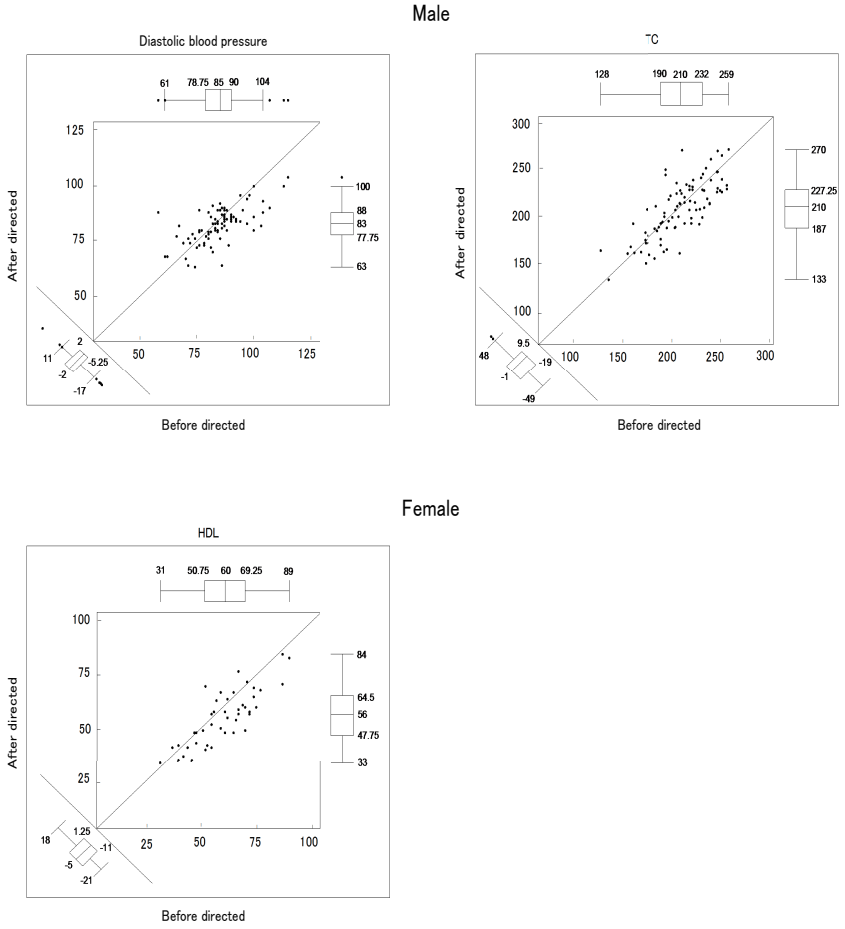
For the clinical test items that the shapes of the distribution were almost unchanged before and after the doctor's direction, we conducted the paired two-samples t-test (One-side: $\alpha = 0.05$) by noticing to the normality after power-transformation for pre- and post-observations. From the results in Display 3.9, it assumed that pre- and post-observations followed the bivariate power-normal distribution with $\lambda_{\text{Pre}} = \lambda_{\text{Post}} = 0.5$ (Square root transformation) for Weight and SBP in Male and SBP in Female, $\lambda_{\text{Pre}} = \lambda_{\text{Post}} = 0$ (Log transformation) for BMI and TG in Male, $\lambda_{\text{Pre}} = \lambda_{\text{Post}} = -0.5$ (Inverse square root transformation) for HDL in Male and Weight, BMI and TG in Female, $\lambda_{\text{Pre}} = \lambda_{\text{Post}} = 1.5$ (1.5 power transformation) for DBP and TC in Female. We set μ_{Pre} and μ_{Post} as the means of pre- and post-observations after the power transformation.

We conducted the paired two-samples t-test (One-side) with null hypothesis $\text{H}_0 : \mu_{\text{Pre}} = \mu_{\text{Post}}$, alternative hypothesis $\text{H}_1 : \mu_{\text{Pre}} > \mu_{\text{Post}}$ if $\lambda_{\text{Pre}} = \lambda_{\text{Post}} \geq 0$ and alternative hypothesis $\text{H}_1 : \mu_{\text{Pre}} < \mu_{\text{Post}}$ ($\text{H}_1 : \mu_{\text{Pre}} > \mu_{\text{Post}}$ [only HDL in Men]) if $\lambda_{\text{Pre}} = \lambda_{\text{Post}} = -0.5$. As the results, Weight, BMI and SBP in Male and Weight and BMI in Female decreased significantly. But in the results which conducted the paired two-sample t-test (One-side) for the converse alternative hypothesis H_1 (i.e. alternative hypothesis $\text{H}_1 : \mu_{\text{Pre}} < \mu_{\text{Post}}$ if $\lambda_{\text{Pre}} = \lambda_{\text{Post}} \geq 0$), TG in Male, SBP and TG in Female increased significantly, HDL in Male decreased significantly. Moreover, we showed sliding square plot for DBP and TC in Male and HDL in Female which were indicated that pre- and post-observations followed different distributions separately in Display 3.10. We found that HDL in Female was decreased after the direction especially.

3.3.5 Consideration

In section 3.1, we found that the characteristics of the body shape were mainly related to the classification between uncontrolled and directed group. Moreover, because it could be considered

that SBP, DBP and TC were also related to the classification from AUC in Display 3.8, we guess that the doctor conducted the directions about their body shapes for them in directed group. Also, because the correct classification rates in data-adaptive discriminant analysis were higher in Female than in Male, the doctor might have the clearer judgment criterion for Women. In section 3.2, we investigated the effect of the doctor's direction for the subjects in directed group. As the result, though Weight, BMI and SBP in Male and Weight and BMI in Female decreased significantly and they were improved by the direction, TG and HDL in Male and SBP, TG and HDL in Female deteriorated. Though we could not judge the efficacy of the direction only in these results, we have to pay attention that TG and HDL which lead to abnormal lipid metabolism deteriorated though Weight and BMI (Body shape) which were related much to the classification in both Male and Female decreased. From the above, it was found that embodying the subject characteristics was useful to interpret the direction effect.



Display 3.10. Sliding square plot: directed group

3.4 Statistical diagnosis and validity of results

3.4.1 Statistical diagnosis

In section 3.1, we assumed that the clinical test item followed the power-normal distributions and explored the clinical test item which contributed to the classification between uncontrolled and directed group. However, statistical diagnosis (sample diagnosis, structural diagnosis) is important to ensure the results (the findings) obtained in data analyses. Because the sample diagnosis (the diagnosis of an outlier and an influential observation) has already done, we conduct the structural diagnosis for the models (the power-normal distributions) here. We can visually judge the fitness of the power-normal distributions for clinical test results from the data-adaptive discriminant plot in Display 3.3 and Display 3.4. We found that every clinical test items fit power-normal distributions because the diremptions between the observations and the transformation curves were small. Moreover, we check the median (50% point) of the power-normal distribution in the following procedures.

Check 1. Estimate the median of the clinical test item from the power-normal distribution

Check 2. Calculate the medians from bootstrap samples 1000 times

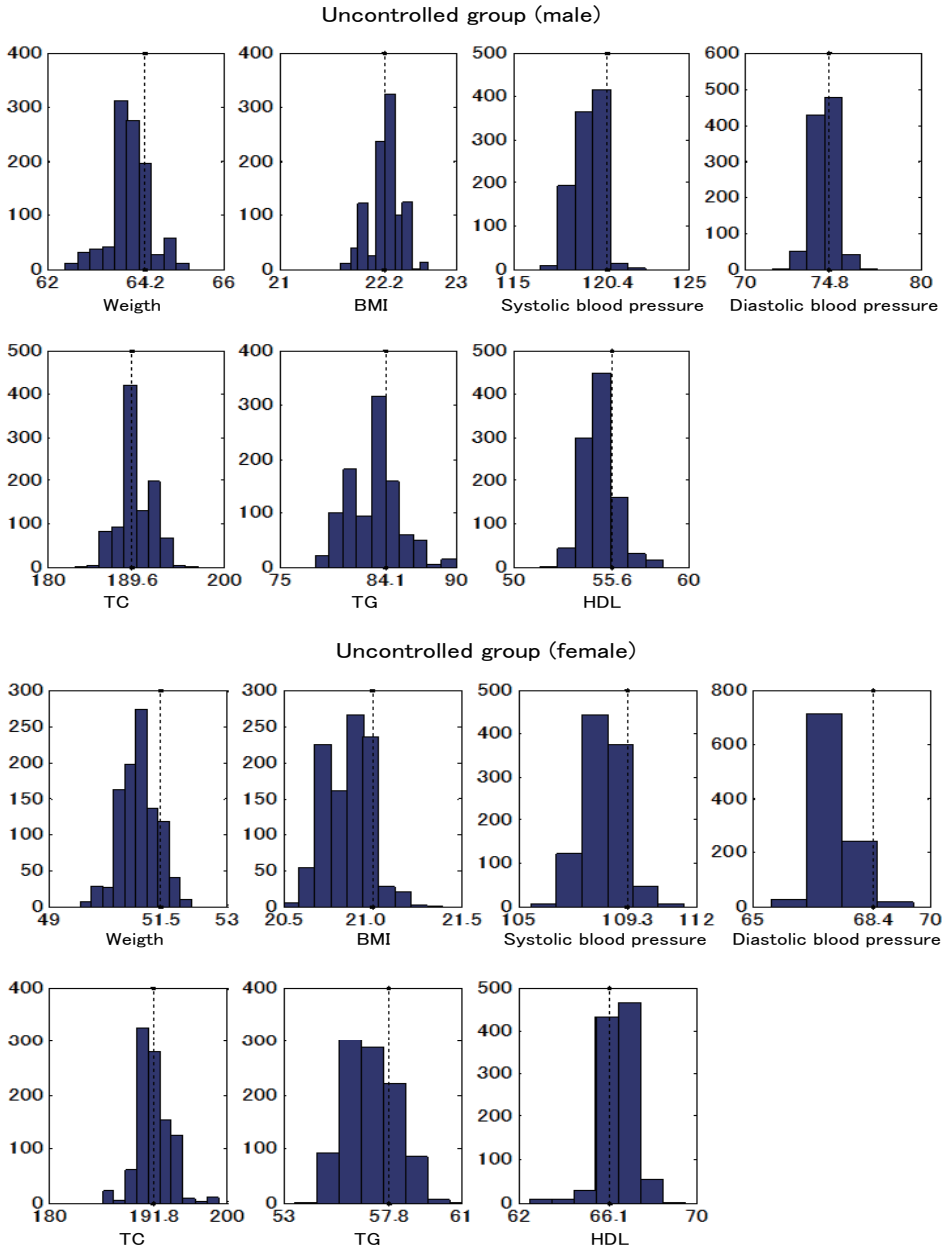
Check 3. Create the histogram of the medians calculated in Check 2 and compare it to the median based on the power-normal distribution in Check 1.

The results were showed in Display 3.11 and Display 3.12. The center line within these Displays was the median estimated from the distribution. In uncontrolled group, for clinical test items except for SBP, HDL in Male and SBP, DBP in Female, we can judge that the power-normal distributions were appropriate (for the medians) as the underlying distribution because the medians of the bootstrap samples existed near the medians estimated from the distributions. However the medians estimated from the distributions for SBP, HDL in Male and SBP, DBP in Female were located to the right-tailed parts in the histograms of the medians of the bootstrap samples, so the appropriateness of the power-normal distributions might be suspicious, but it would appear that the actual phenomenon (data) did not differed from the model (power-normal distribution) substantially because the variation of the medians of the bootstrap samples was small. In directed group, the medians of the bootstrap samples were closed to the medians

estimated from the distributions for most clinical test results, but compared to uncontrolled group, the variations of the medians of the bootstrap samples were larger due to smaller subjects.

3.5 Conclusion

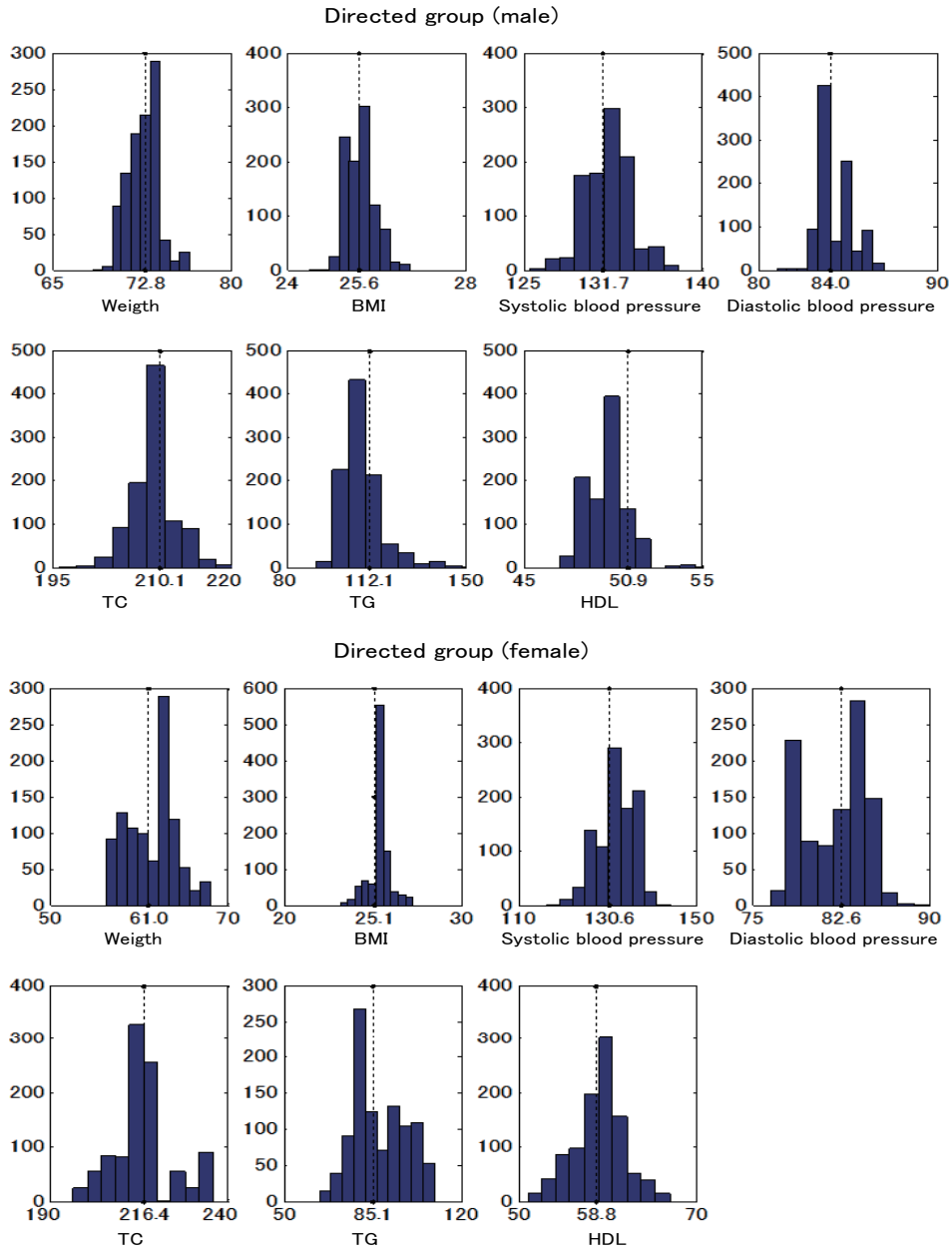
In this chapter, based on the results of the health checkup that aims to prevent disease was carried out in April 2004, we explored foundation about the doctor's judgment, especially



[Center scale shows a median estimated from power-normal distribution.]

Display 3.11. Median plot: uncontrolled group

classification of directed group, attempting to figure the doctor's character, and further evaluated directed effect for directed group. Through a process of data analysis which are conscious of logic consistency, we gained the findings of an evaluation and consideration for health care advice from "Set of cold figures" (the health checkup data) (Goto, 1986). An evaluation method showed in this paper is applicable for the occasion that it would like to express the characteristics of group where clear definition does not exist. However, when the censoring exists in upper and lower limit like in directed group, we have to pay attention to the effect. To clarify the direction effect



Display 3.12. Median plot: directed group

for better living and the prevention effect for illness, it is important to examine the effect of health care advice in particular based on the subject characteristics which was obtained through the process of data analysis which are conscious of logic consistency. However, as showed in section 3.4, we should select model for small sample carefully, so we have to approach from various viewpoints in actual application occasion.

Reference

1. Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion). *J. Roy. Statist. Soc*, B26(2), 211-246.

2. Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification And Regression Trees.* Wadsworth.

3. Dixon, W. J. (1953). Processing data for outliers. *Biometrics,* 9, 74-89.

4. Goto, M. and Hamasaki, T. (2002). The bivariate power-normal distribution. *Bulletin of Informatics and Cybernetics*, **34**(1), 29-49.

5. Goto, M., Matsubara, Y. and Tsuchiya, Y. (1983). Power-normal distribution and its applications. *Rep. Stat. Appl. Res., JUSE*, **30**, 8-28.

6. Goto, M., Uesaka, H. and Inoue, T. (1979). Some linear models for power transformed data. *Invited paper at the 10th International Biometric Conference.* August, 6-10(*Res. Rep. NO.93, Res. Instit. Fund. Infor. Sc., Kyusyu University*).

7. Hatanaka, S., Inoue, T. and Goto, M. (1981). Discrimination on power-normal distribution. *The 9th Symposium of the Japanese Society of Computational Statistics*, 113-114.

8. 近藤　誠 (2004)．成人病の真実．文芸春秋.

9. 厚生労働省健康局 (2007)．標準的な健診・保健指導プログラム（確定版）：概要．厚生労働省健康局．http://www.mhlw.go.jp/bunya/kenkou/seikatsu/pdf/01.pdf

10. Maruo, K., Shirahata, S., Goto, M. and Komazawa, T. (2008). Statistical investigation of reference intervals of clinical laboratory data. *The Japanese Journal of Behaviormetrics*, **35** (1), 73-89 (in Japanese).

11. 大櫛陽一 (2006). 検査値と病気. 間違いだらけの診断基準. 太田出版.

12. 大櫛陽一 (2007). 読売新聞 (07/10/9 付).

13. Sasse, E. A., Doumas, B. T., Miller, W. G., D'Orazio, P., Eckfeldt, J. H., Evans, S. A., Graham, G. A., Myers, G. L., Parsons, P. J. and Stanton, N. V. (2000). *How to Difine and Determine Reference Intervals in the Clinical Laboratory: Approved Guideline-Second Edition.* NCCLS.

14. Seo, H., Shimokawa, T., Daimon, T. and Goto, M. (2002). Data-adaptive discrimination based on multivariate power normal distribution, *Contributed paper at the 5th of The International Association of Statistical Computing, ARS*, 114-115.

15. Shimokawa, T. and Goto, M. (2004). Data-adaptive discriminant analysis and its diagnosis. *Bulletin of the Computational Statistics of Japan*, **17**(2), 87-108 (in Japanese).

16. Shimokawa, T. and Goto, M. (2002). Data-adaptive Probability Plot and its Applications. *The Japanese Journal of Behaviormetrics*, **29**(1), 103-119 (in Japanese).

17. Sugimoto, T., Shimokawa, T. and Goto, M. (2005). Tree-structured approaches and recent advances. *Bulletin of the Computational Statistics of Japan*, **18**(2), 123-164 (in Japanese).

4. The impact of the shape of the underlying distribution of observations on test results

4.1 Introduction

In clinical research, we consider the difference between pre- and post- treatment observations as an evaluation indicator for treatment effect. Then, pre- and post- treatment are paired, not independent. For example, when pre- and post- treatment observations are measured for I subjects, pre- treatment observation X_{B_i} and post- treatment observation X_{T_i} for i-subject can be expressed by

$$X_{Bi} = \mu + S_i + e_{Bi}, \tag{4.1}$$

$$X_{Ti} = \mu + S_i + \Delta + e_{Ti}, i = 1, \ldots, I \tag{4.2}$$

(Bonate, 2000), where μ is a population mean, S_i is i-subject effect, Δ is a clinical effect, e_{B_i} and e_{T_i} are the error terms which follow the distributions with the expectation 0 and variances σ^2. Then, the difference is presented by

$$X_{Ti} - X_{Bi} = \Delta + e_{Ti} - e_{Bi}, i = 1, \ldots, I \tag{4.3}$$

$\Delta = 0$ means no treatment effect and $\Delta \neq 0$ means treatment effect. When we examine whether the treatment effect exists or not, it is often assumed that the observations follow normal distribution, and a paired t-test in a one-sample problem and two-samples t-test in a two-sample problem are applied for them. But a lot of endpoints exist in the actual clinical research and the endpoints do not always follow the normal distribution. For example, the analysis results for 1141 subjects participating in a health checkup which was conducted at a company in 2004 (Isogawa, Ikebe, Sakamoto and Goto, 2011) and for 8815 subjects participating in a complete physical examination which was conducted at a clinic in 2003 (Maruo, Shirahata, Goto and Komazawa, 2008), the blood pressure and many items in the clinical laboratory test

did not follow the normal distribution. By now, the impact of non-normality of the potential distribution on the paired and the two-samples t-test have been discussed in several papers. Blair and Higgins (1985) compared the power in the paired t-test to it in the Wilcoxon signed rank test in small sample size when the observations follow several potential distributions. In the paper, normal distribution, uniform distribution, double exponential distribution, truncated normal distribution, exponential distribution, mixed normal distribution, log-normal distribution, chi-square distribution and Cauchy distribution were used as the potential distributions. And Yand and Tsiatis (2001) focused on the occasions that we apply the paired and two-samples t-test under the semi-parametric situations where the distributions of pre- and post- observations do not need to be specified, and inquired about the asymptotic efficacy between the sample variance and the variance estimator led by central limit theorem. However, as discussed, because we treat various endpoints in clinical research, it could be well considered that the endpoints follow the distributions without the above distributions. Also, we treat finite observations, so there are many situations that the asymptotic properties cannot be available. In this paper, when it is assumed that pre- and post- treatment observations follow various distributions, we evaluate the impact of them on tests which require the normal assumptions. Because we often conduct two-group comparison between actual group and placebo group in clinical research, we consider not only one-sample problem but also two-sample problem. We evaluate the performance of the paired t-test in one-sample problem and the two-samples t-test in two-sample problem, but also use the Wilcoxon signed rank test in one-sample and the Wilcoxon rank sum test as the comparison of the t-tests. To clarify the relationship and the structure between the distributions of pre- and post-observations and the distribution of the difference, we especially focus on the following points.

(a) Relation between non-normality of distributions of pre- and post-observations and non-normality of the distribution of the difference.

(b) Influence of non-normality of distribution of the difference on power in above tests.

(c) Availability of interpreting the test results corresponding to distributions of pre- and post-treatment samples.

As an approach to (a), we assume that pre- and post-observations follow a bivariate power-normal distribution (BPND: Goto and Hamasaki, 2002) in order to consider the relationship

between the distributions of pre- and post-observations and the distribution of the difference comprehensively and quantitatively. The bivariate power-normal distribution is the bivariate extended form of an univariate power-normal distribution (PND) which was proposed by Goto, Matsubara and Tsuchiya (1983). The univariate power-normal distribution is defined as the distribution which the observations before the power-transformation (Box and Cox, 1964) follow, and contains various distributions including well-known normal distribution and log-normal distribution, so can cover real situations to some extent and is useful to evaluate the discrepancies between ideal (model and hypothesis) and reality (data) (Goto, Uesaka and Inoue, 1979; Goto and Inoue, 1980; Goto, Matsubara and Tsuchiya, 1983). Moreover, because pre- and post-observations have the correlated relationship, the bivariate power-normal distribution including the correlation structure is suitable for assessing our problem. In fact, to analyze the health checkup data for 1141 subjects and the complete physical examination data for 8815 subjects as previously discussed, the blood pressure and the clinical laboratory test were assumed to follow the PND, and these data fitted the PND well. Because the PND express the features of the distribution which the data follow even if the distribution is not known previously, we notice on the PND in this paper. Additionally, to make clear the situation examined in this paper, we identify the distribution of pre- and post-observations by using a shape parameter (power-parameter) which expresses a skewness of the distribution and an indicator which express a variation of the distribution defined in 2.2. And we derive the distributions of the difference from numerical integral in several situations and inquire the properties about the distributions of the difference. As an approach to (b) and (c), we examine the impact of the shape of the potential distribution on the results of the t-tests.

In Section 2, we briefly describe statistical methods used in this paper. In Section 3, we examine the properties of the distributions of the difference to examine (a). In Section 4, small scale simulations are provided to examine (b), (c) and consider the results. Finally, in Section 5, we summarize some productive findings obtained by Section 3 and 4 and conclude with some further developments.

4.2 Statistical Method

It is assumed that prior- and post-observations follow the BPND.

4.2.1 Univariate power-normal distribution (PND)

The power-transformation of positive variable X is defined as

$$X^{(\lambda)} = \begin{cases} (X^{\lambda} - 1)/\lambda, & \lambda \neq 0 \\ \log X, & \lambda = 0 \end{cases} \tag{4.4}$$

(Box and Cox, 1964). The power-normal distribution on original scale X is proposed (Goto, Uesaka and Inoue, 1979: Goto, Matsubara and Tsuchiya, 1983) and the probability density function is given by

$$f_{\mathrm{PN}}(x; \lambda, \mu, \sigma) = x^{\lambda-1}\phi\{(x^{(\lambda)} - \mu)/\sigma\}/A(K), x > 0 \tag{4.5}$$

where ϕ is a probability density function of standard normal distribution and $A(K)$ is a probability proportional constant term given by

$$A(K) = \begin{cases} \Phi\{-K\}, & \lambda < 0 \\ 1, & \lambda = 0 \\ \Phi\{K\}, & \lambda > 0 \end{cases} \tag{4.6}$$

$1 - A(K)$ presents truncated probability where $K = (1 + \lambda\mu)/(\lambda\sigma)$ and $\Phi(\cdot)$ is cumulative distribution function of standard normal distribution. If the probability proportional constant term is small, it is well-known that the data cannot preserve the normality after power-transformation. Parameter λ, μ and σ are respectively called shape, local and scale parameter and the power-normal distribution for X is identified by changing λ corresponding to X. Here, X follows the normal distribution in $\lambda = 1$ and the log-normal distribution in $\lambda = 0$. The advantage of using the power-normal distribution is that it is a comprehensive model to be able to comprehend the diremption between the ideal (model and hypothesis) and the real (data), analyze real data adequately and use many traditional methods based on normal distribution.

Also, the $100p$ percent point ξ_p is presented by

$$\xi_p = \begin{cases} \{\lambda(\mu + \sigma z_{p^*}) + 1\}^{1/\lambda}, & \lambda \neq 0, \\ \exp(\mu + \sigma z_p), & \lambda = 0 \end{cases} \tag{4.7}$$

(Maruo & Goto, 2008), where z_p, z_{p^*} are the $100p, 100p^*$ percent

$$p^* = \begin{cases} 1 - A(K)(1 - p), & \lambda > 0, \\ A(K)p. & \lambda < 0 \end{cases} \tag{4.8}$$

Because the truncated term happens in anything but $\lambda = 0$, $A(K)$ is not always 1. However, in this paper, for the interpretation of the results ease, we focus on the cases of $A(K) = 1$, that is, we assume the non-truncated situation.

4.2.2 Expression of parameter transformation

Though we consider various distributions in the framework of the PND, it is difficult to interrupt μ and σ directly because these parameters change much according to λ. Therefore, in this paper, we specify the distribution in $\{\lambda, \xi_{0.5}(\text{Median}), \tau(\text{Variation of distribution})\}$ and calculate $\{K, \mu, \tau\}$ using the following relationship. Here we express variation of one distribution τ as

$$\tau = (\xi_{0.75} - \xi_{0.25})/\xi_{0.5}. \tag{4.9}$$

Given $\{\lambda, \xi_{0.5}, \tau\}$, $\{\mu, \sigma\}$ can be derived from

$$\mu = (1 + z_{0.5^*}/K)^{-1} \left[(\xi_{0.5}^{\lambda} - 1)/\{\lambda - z_{0.5^*}/(\lambda K)\} \right], \tag{4.10}$$

$$\sigma = (1 + \lambda\mu)/(\lambda K) \tag{4.11}$$

We calcucate $\{\lambda, \xi_{0.5}, \tau\}$ from $\{K, \mu, \sigma\}$ in grid search method. The relationship between the variation of the distribution (τ) and the standard deviation (SD) was shown in Figure 4.1. It was almost proportional relation regardless of λ, so we use the variation of the distribution as the alternative of the standard deviation (SD).

the p-moment of $p < |\lambda|$ does not exist in the power-normal distribution with $\lambda < 0$. So we use

$$\eta = (\xi_{0.975} - \xi_{0.5})/(\xi_{0.5} - \xi_{0.025}) \tag{4.12}$$

as the skewness indicator of the distribution. η approaches to 1 as the distribution is closed to the symmetry ($\eta > 1$ in the right skewed distribution and $\eta < 1$ in the left skewed distribution).

The relationship between the variation of the distribution and the skewness indicator was shown in Figure 4.2. It was found that the distribution was skewed to the right as λ decreased and τ increased.

4.2.3 Bivariate power-normal distribution (BPND)

An extension of PND to two-dimensional case is the bivariate power-normal distribution (BPND). Expressing power transformed variables of two positive variables (X_1, X_2) as $(X_1^{(\lambda_1)}, X_2^{(\lambda_2)})$,

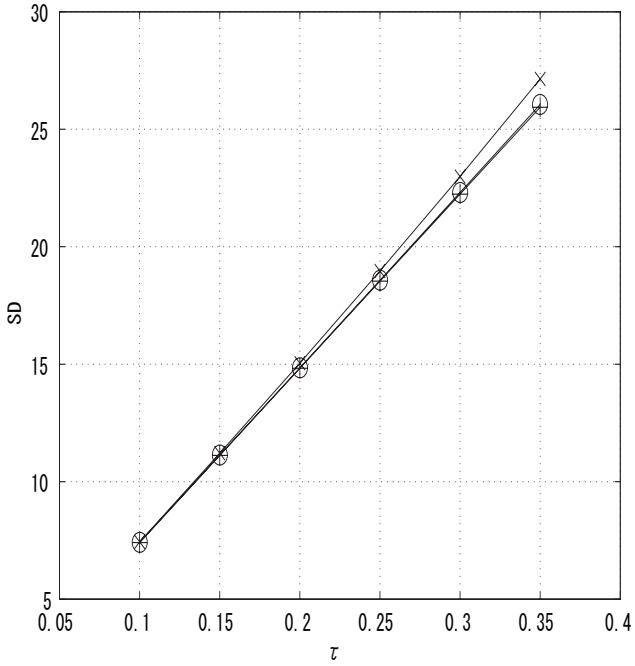


Figure 4.1: Relationship between τ and SD [$\lambda = 0$(cross),0.5(circle),1(plus)]

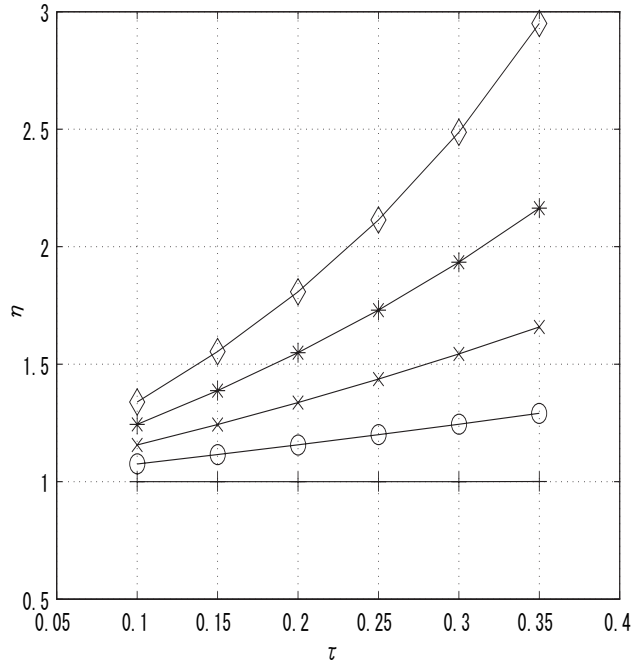


Figure 4.2: Relationship between τ and η

[$\lambda = -1$(diamond),-0.5(asterisk),0(cross),0.5(circle),1(plus)]

then we can define joint probability density function of (X_1, X_2) as

$$g(x_1, x_2) = x_1^{\lambda_1 - 1} x_2^{\lambda_2 - 1} f(x_1^{(\lambda_1)}, x_2^{(\lambda_2)}) / A(\boldsymbol{K}), x_1, x_2 > 0 \tag{4.13}$$

(Goto and Hamasaki, 2002), where

$$\begin{aligned} f(x_1^{(\lambda_1)}, x_2^{(\lambda_2)}) =& 1/(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}) \exp[-\{Q(x_1^{(\lambda_1)}, x_2^{(\lambda_2)})/2], \\ Q(x_1^{(\lambda_1)}, x_2^{(\lambda_2)}) =& 1/(1-\rho^2)[\{(x_1^{(\lambda_1)} - \mu_1)/\sigma_1\}^2 - 2\rho\{(x_1^{(\lambda_1)} - \mu_1)/\sigma_1\} \\ & \{(x_2^{(\lambda_2)} - \mu_2)/\sigma_2\} + \{(x_2^{(\lambda_2)} - \mu_2)/\sigma_2\}^2] \end{aligned} \tag{4.14}$$

and ρ is the correlation parameter. Also, $A(\boldsymbol{K})$ is a probability proportional constant term of BPND given by

$$A(\boldsymbol{K}) = \int_{a_2}^{b_2} \int_{a_1}^{b_1} \phi(x_1, x_2 : \rho) dx_1 dx_2 \tag{4.15}$$

where $\phi(x_1, x_2 : \rho)$ is joint probability density function of bivariate normal distribution which the margin distribution is a standard normal distribution. Putting $k_j = (\lambda_j \mu_j + 1)/\lambda_j (j = 1, 2)$, a_j and b_j are presented by $a_j = -k_j, b_j = \infty$ if $\lambda_j > 0$, $a_j = -\infty, b_j = \infty$ if $\lambda_j = 0$ and $a_j = -\infty, b_j = -k_j$ if $\lambda_j < 0$. The counter plots of some BPNDs were shown in Figure 4.3. We set λ as $-1, 0, 1$, τ as $0.1, 0.35$, median of pre-observation X_B as $\xi_{B0.5} = 100$ and median of post-observation X_T as $\xi_{T0.5} = 95$ and correlation coefficient parameter as $\rho = 0.75$. As the figure shown, it was found that the distribution was skewed to the right as λ decreased.

4.3 Distribution of the difference

It is assumed that pre-observation X_B and post-observation X_T follow the BPND. Because it is often considered that the distributions which X_B and X_T follow are the same and the variations of them are also the same, we set that λ (power-parameter) and τ (variation of the distribution) in X_B and X_T are equal in this paper.

When it was assumed that the potential distributions of the health checkup data in Isogawa *et al.*(2011) and the complete physical examination data in Maruo *et al.*(2008) were the PNDs, most of these data followed the right skewed-distribution, so we consider about the cases that pre- and post-observation follow the PNDs with $\lambda \leq 1$. The median of the pre-observation X_B is set as $\xi_{B0.5} = 100$, the median of the post-observation X_T as $\xi_{T0.5} = 100$ and the correlation parameter as $\rho = 0.75, 0.9$. The densities of the difference $p(D)$ were shown in Figure 4.4. When

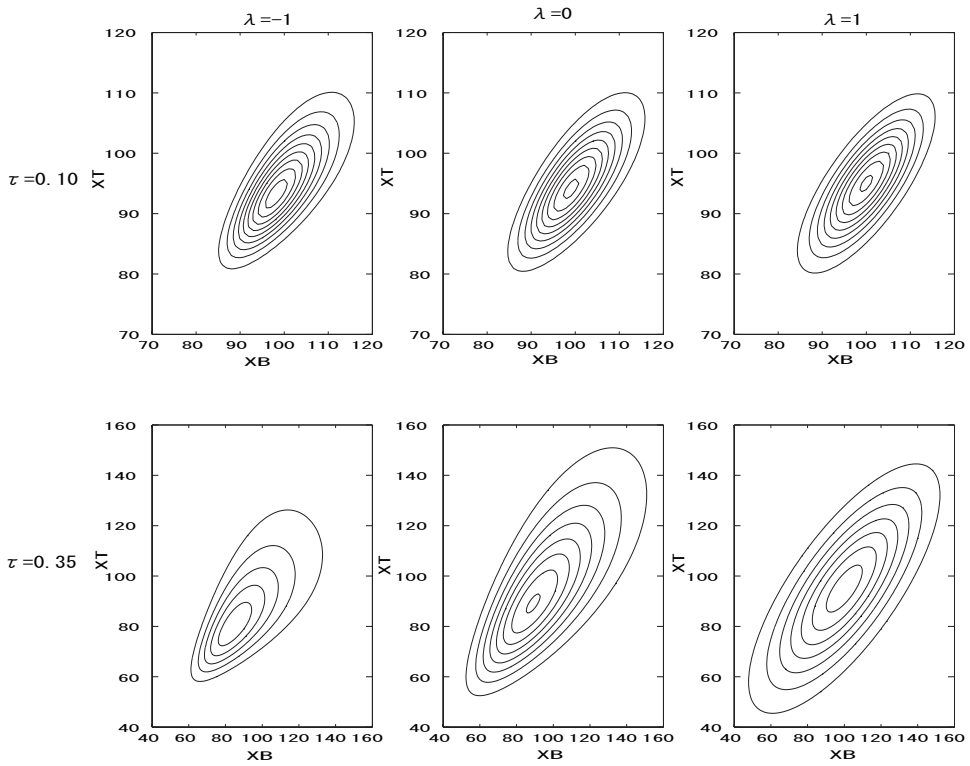


Figure 4.3: Counter plots of some bivariate power-normal distributions ($\rho = 0.75$)

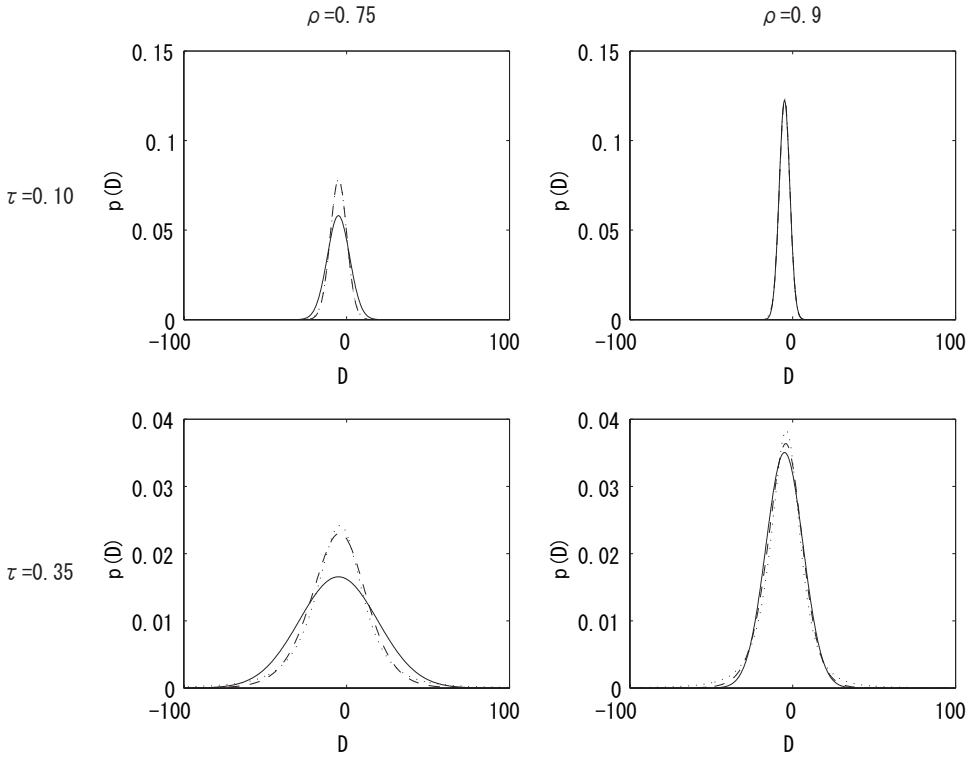


Figure 4.4: Density of the distribution of the difference D ($\lambda = -1$(dot),0(dash),1(solid))

pre-observation X_B follows the normal distribution ($\lambda = 1$), at the range between 0.1 and 0.35 in τ, the standard deviation (SD) of X_B vary between 11.1 and 25.9 and it diserves to the range between 0.111 and 0.259 in the coefficient of the variation (CV).

Though the distributions of the difference with $\lambda = -1, 0$ were more convex than those with $\lambda = 1$ in $\tau = 0.1, 0.35$ and $\rho = 0.75$, there did not exist large differences among these distributions. These distributions of the difference were almost symmetry. Moreover, by using the numerical integral, we calculated the skewness indicator of the distributions which the pre-observation X_B and the difference D follow. The results were shown in Figure 4.5. Regardless of the coefficient parameter, the skewness of the distribution of the difference approached to the symmetry compared to pre-observation even in the case but $\lambda = 1$. Especially, we notice on the fact that the skewness of the difference indicated the almost symmetrical distribution regardless of that the skewness of the pre-observation with $\lambda < 0$ was very large.

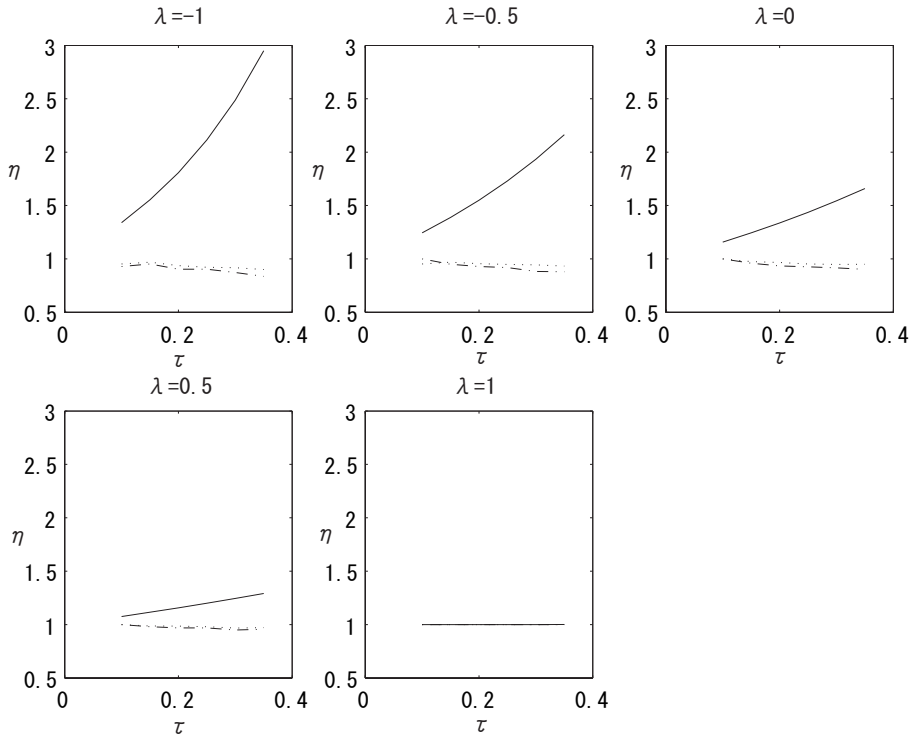


Figure 4.5: Skewness indicator η of the distribution of the pre-treatment and the difference(X_B(solid),D($\rho = 0.75$(dot),0.9(dash)))

4.4 Simulation

4.4.1 One-sample problem

Purpose We examine the impact of the distribution of the difference on the paired t-test which assumes normality.

For details, we assume that the pre- and post-observations follow several BPNDs. We conduct the paired t-test for the pre- and post-observations and those after the power-transformation. By comparing the powers of the paired t-test for the pre- and post-observations and those after the power-transformation, we can evaluate the loss of information about the distribution. Because it focuses on the occasion that the pre- and post-observations do not follow normal distribution, we also provide further insights into the Wilcoxon signed-rank test (a typical non-parametric test) which is selected as alternative of the paired t-test in the cases of that the data do not follow the normal distribution.

Method Set the pre- and post-observations as those generated from the BPNDs whose medians of the pre- and post-observations are 100 and 95.

In each case of the BPND with $\lambda = -1, -0.5, 0, 0.5, 1$, $\tau = 0.1, 0.15, 0.2, 0.25, 0.3, 0.35$ and $\rho = 0.75, 0.9$, we calculated the minimum sample size that the observations after power-transformation indicate more than 90% power in the paired t-test. However, the cases with $\tau = 0.1$ were eliminated in $\rho = 0.75, 0.9$ because the sample size were less than 10 with $\tau = 0.1$ and $\rho = 0.9$. The sample size in each simulation was indicated in Table 4.1.

We generated the pre- and post-observations from each BPNDs and calculated the type I errors and the power in the paired t-test and the Wilcoxon signed rank test for the pre- and post-observations and the paired t-test for those after the power-transformation.

Then the type I error was calculated as follows: After getting the observations of the pre- and post-observations from the above BPNDs whose medians of the pre- and post-treatment were 100, we conducted the above three tests for the observations. The procedures were repeated in 50,000 and the proportion that the null hypothesis was rejected in each test was calculated as the type I error. The power was also calculated in the same way of the type I error except for that the medians of the post-treatment was 95. These results were showed in Figure 4.6-4.9.

Table 4.1: Sample size in each simulation

		One-sample					Two-sample				
		τ					τ				
λ	ρ	0.15	0.2	0.25	0.3	0.35	0.15	0.2	0.25	0.3	0.35
1	0.75	25	45	69	100	136	51	91	142	204	277
	0.9	11	18	28	41	55	21	37	57	82	112
0.5	0.75	25	44	69	99	135	51	89	140	201	273
	0.9	10	18	28	40	54	21	36	56	81	110
0	0.75	25	44	69	99	133	50	88	137	197	266
	0.9	10	18	28	40	54	20	35	55	79	107
-0.5	0.75	25	44	68	97	131	49	86	134	191	258
	0.9	10	18	28	39	53	20	35	54	77	104
-1	0.75	25	43	66	94	126	48	83	129	183	245
	0.9	10	18	27	38	51	19	34	52	74	99

Result At first, we noticed the relationship between the variation of the distribution τ and the type I error. From the results of the type I error in Figure 4.6 and Figure 4.7, the type I errors indicated about 0.05 in not only $\lambda = 1$ but also almost all cases. So we consider that the type I error were almost preserved to 0.05 and focus on the power in Figure 4.8 and Figure 4.9.

As expected, the powers in the paired t-test for those after power-transformation were almost preserved to 0.9. In the case of $\lambda = 1, 0.5$ in $\rho = 0.75, 0.9$, the powers in three tests were almost the same 0.9. In the case of $\lambda = 0$ in both $\rho = 0.75, 0.9$, there were no large differences among the powers in the three tests, but the powers were strictly the paired t-test for the pre- and post-observations after power-transformation, the paired t-test for those and the Wilcoxon signed-rank test in ascending order. Also though the powers of the paired t-test for the pre- and post-observations and those after power-transformation when τ is low, the powers in the paired t-test for those after power-transformation approached to those in the Wilcoxon signed-rank test as τ increased. In the case of $\lambda = -0.5, -1$ in $\rho = 0.75, 0.9$, the powers in the paired t-test for those after power-transformation decreased largely as τ increased, and they were under 0.6 especially in $\lambda = -1$ and $\tau = 0.35$. Though the powers in the Wilcoxon signed-rank test

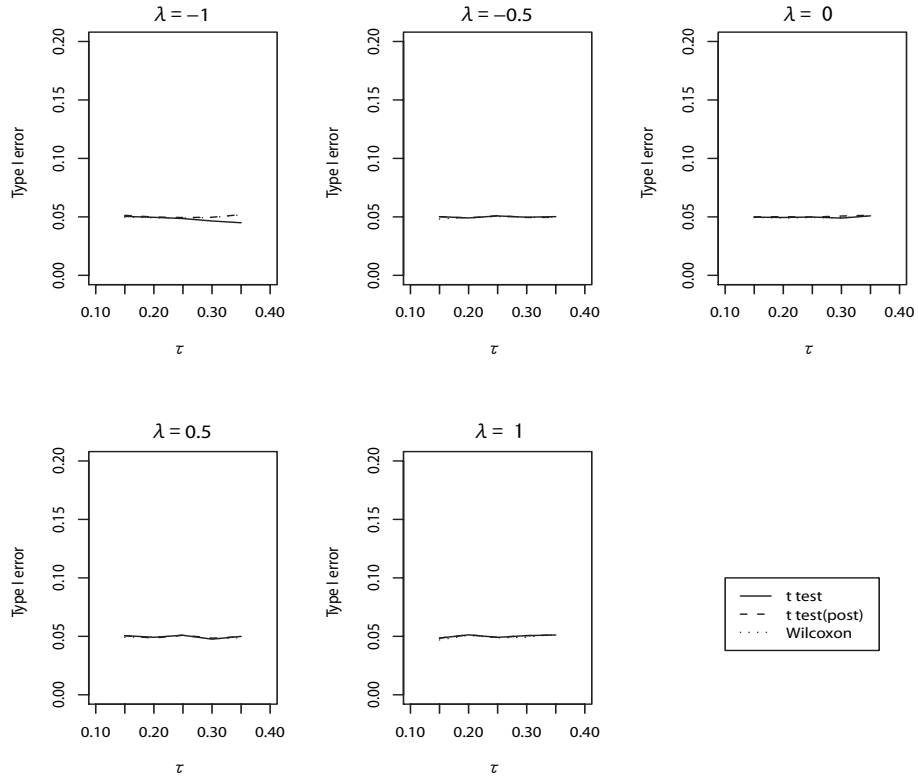


Figure 4.6: One-sample: Relationship between τ and type I error ($\rho = 0.75$)

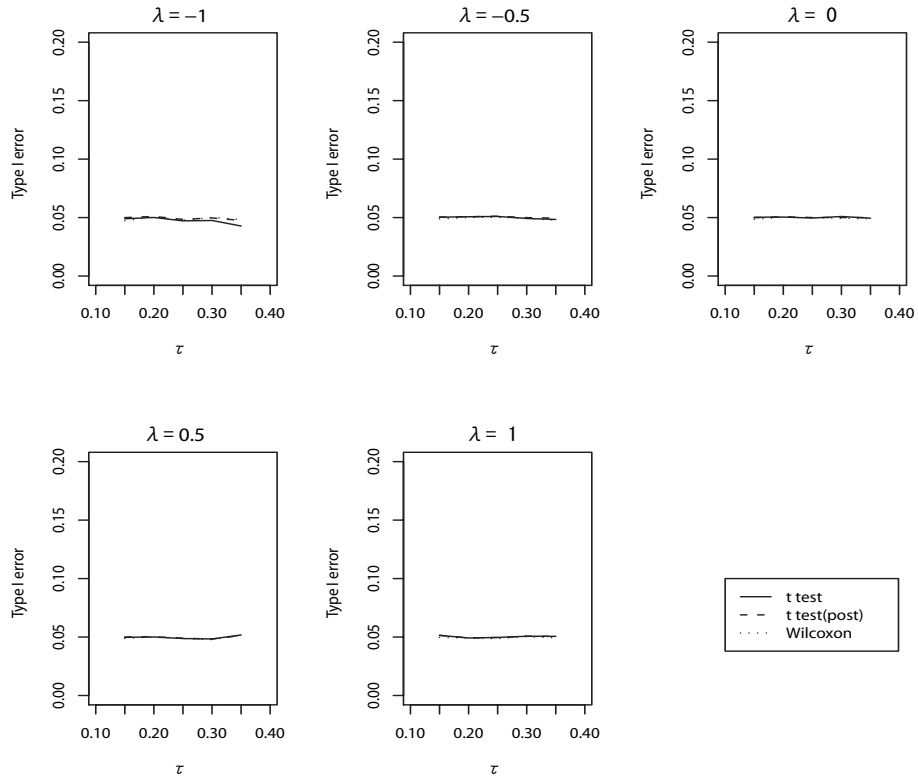


Figure 4.7: One-sample: Relationship between τ and type I error ($\rho = 0.9$)

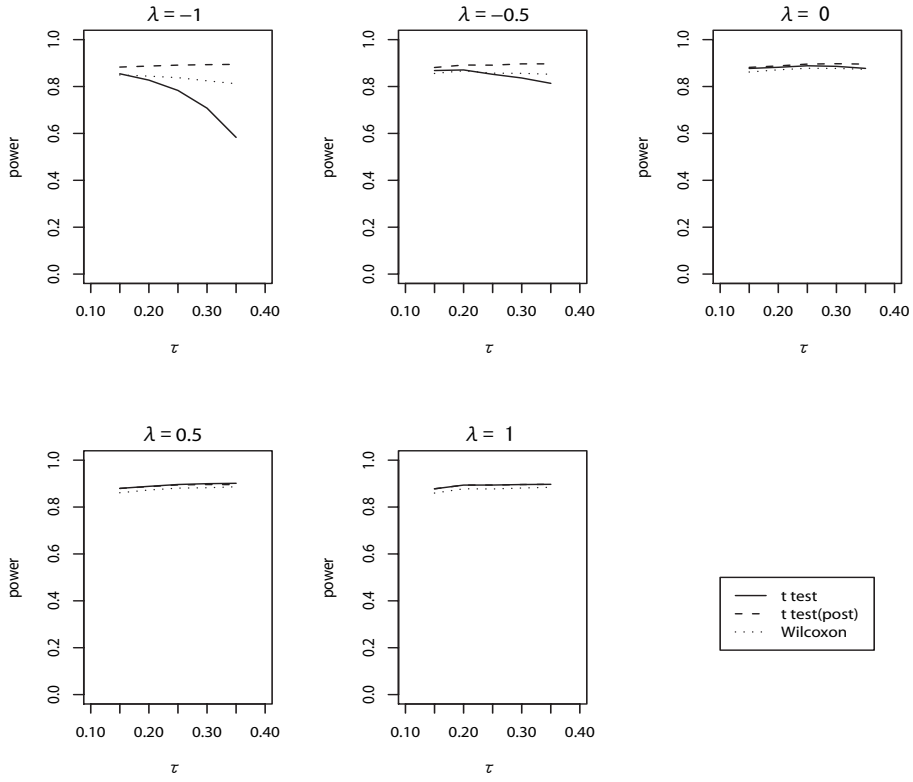


Figure 4.8: One-sample: Relationship between τ and power ($\rho = 0.75$)

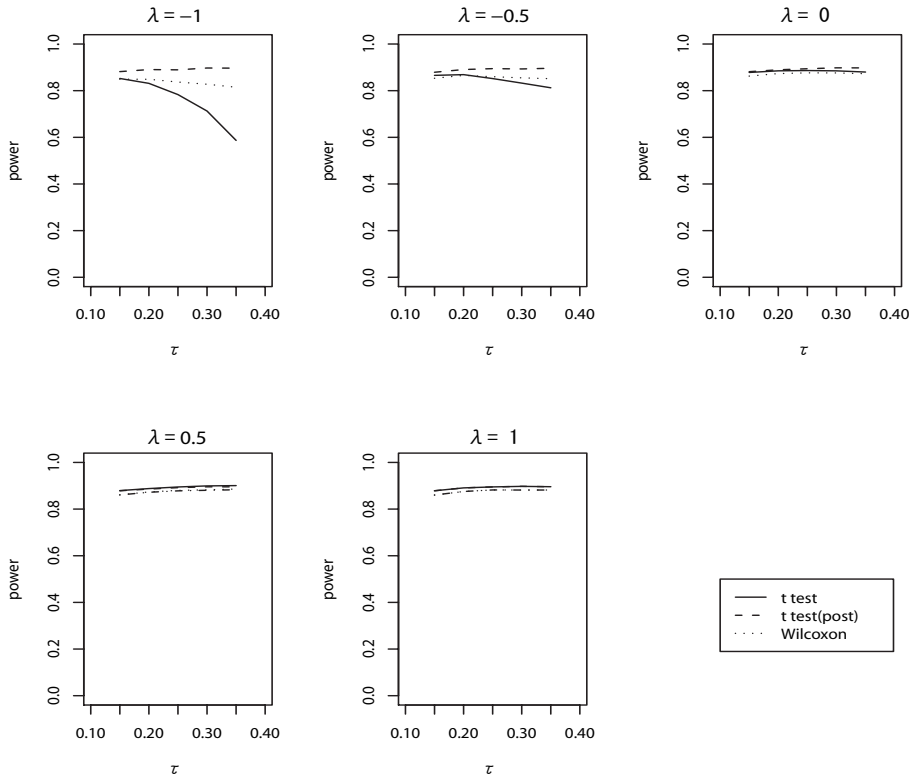


Figure 4.9: One-sample: Relationship between τ and power ($\rho = 0.9$)

Table 4.2: Percent points of some staistics obtained in the simulation for one-sample problem ($\rho = 0.75$)

			Percent point				
τ	Statistics	λ	5%	25%	50%	75%	95%
0.15	Test statistics	-1	1.47	2.45	3.19	3.98	5.26
		-0.5	1.54	2.52	3.24	4.02	5.27
		0	1.58	2.55	3.28	4.07	5.33
		0.5	1.58	2.57	3.30	4.08	5.37
		1	1.59	2.56	3.29	4.08	5.37
	Sample mean	-1	-7.78	-6.17	-5.08	-3.98	-2.41
		-0.5	-7.66	-6.10	-5.04	-3.98	-2.48
		0	-7.59	-6.07	-5.03	-3.98	-2.47
		0.5	-7.58	-6.06	-5.02	-3.97	-2.49
		1	-7.58	-6.06	-5.02	-3.97	-2.49
	SD	-1	5.94	7.08	7.94	8.87	10.3
		-0.5	5.91	6.97	7.77	8.61	9.92
		0	5.86	6.89	7.64	8.43	9.63
		0.5	5.84	6.85	7.58	8.36	9.48
		1	5.84	6.84	7.58	8.34	9.45
0.35	Test statistics	-1	0.24	1.46	2.22	2.96	3.99
		-0.5	1.18	2.19	2.88	3.57	4.58
		0	1.50	2.47	3.15	3.83	4.83
		0.5	1.62	2.59	3.27	3.96	4.99
		1	1.59	2.56	3.26	3.95	4.96
	Sample mean	-1	-10.2	-7.21	-5.45	-3.66	-0.71
		-0.5	-8.41	-6.56	-5.30	-4.03	-2.21
		0	-7.85	-6.26	-5.17	-4.06	-2.48
		0.5	-7.64	-6.11	-5.07	-4.03	-2.52
		1	-7.50	-6.04	-5.00	-3.95	-2.46
	SD	-1	20.4	23.5	26.5	31.3	57.2
		-0.5	17.9	19.6	20.9	22.4	25.0
		0	16.7	17.9	18.8	19.8	21.2
		0.5	16.1	17.2	18.0	18.7	19.9
		1	16.1	17.1	17.8	18.6	19.7

were also decreasing, the tendency was slower pace than those in the paired t-test for pre- and post-observations, and the powers were almost 0.8 even in $\lambda = -1$ and $\tau = 0.35$. To find the causes of why the powers in the paired t-test for the pre- and post-observations decreased in the cases of $\lambda = -1, -0.5$, we showed the percent points (5%, 25%, 50%, 75%, 95%) of the test

statistics, the sample means and the standard deviations of the difference between pre- and post-observations for the simulated 50,000 data in Table 4.2. In $\tau = 0.15$, the percernt points of the test statistics, the sample means and the standard deviations indicated approximately the same values regardless of λ. However, in $\tau = 0.35$, the percent points in $\lambda = -1, -0.5$ were distinctly different than those in $\lambda = 0, 0.5, 1$. In the details, the test statistics in $\lambda = -0.5, -1$ were smaller than those in $\lambda = 0, 0.5, 1$ and the range of the standard deviations and the sample means in $\lambda = -0.5, -1$ was wider than the those in $\lambda = 0, 0.5, 1$. Thus, it is considered that the test statistics decreased as the standard deviations increased, so the powers decreased. Additionally, the results in $\rho = 0.9$ were similar to those in $\rho = 0.75$. We can consider that it has influence on the powers in the paired t-test that the distributions of the difference with $\lambda = -1, -0.5$ and $\tau = 0.35$ are longer tailed in both sides, compared to the distributions of the difference with $\lambda = 1$ (Normal distribution).

Also, the correlation coefficient parameter in BPNDs had little influence on the type I errors and the powers in this simulation.

From these results, we found that the paired t-test is robust even if the assumptions of the normality are slightly violated.

4.4.2 Two-sample problem

Purpose Because we often conduct a comparison between actual group and placebo group in clinical research, we consider two-sample problem in this section. It is often assumed that the difference of the drug effect between the actual drug and the placebo follow the normal distribution, and we conduct the two-samples t-test and the analysis of covariance which assume the normality. However, the clinical endpoints do not always follow a normal distribution as expected. It may happen that the pre- and post-observations do not follow the normal distribution as a result. The purpose of the simulation in this section is to evaluate how these situations have the influence on the results of the two-samples t-test. We also consider the Wilcoxon rank-sum test as the alternatives of the two-samples t-test.

Method Set the pre- and post-observations in actual group as those generated from the BPNDs whose medians of the pre- and post-observations are 100 and 95 respectively, and those in placebo group as those from the BPNDs whose medians of the pre- and post-treatment are

both 100. In each case of the BPND with $\lambda = -1, -0.5, 0, 0.5, 1$, $\tau = 0.1, 0.15, 0.2, 0.25, 0.3, 0.35$ and $\rho = 0.75, 0.9$ in a similar way to one-sample problem, we represented the minimum sample size that the observations after power-transformation indicate more than 90% power in the two-samples t-test in Table 4.1.

We generated the observations from each BPND and calculated the type I errors and the power of the two-samples t-test and the Wilcoxon rank-sum test for the pre- and post-observations and the two-samples t-test for those after the power-transformation in a similar way to one-sample problem. These results were showed in Figure 4.10-4.13.

Result At first, we notice the relationship between the variation of the distribution τ and the type I error. From the results of the type I error in Figure 4.10 and Figure 4.11, the type I errors indicated about 0.05 in not only $\lambda = 1$ but also almost all cases. So we consider that the type I error were almost preserved to 0.05 and compare the power in Figure 4.12 and Figure 4.13.

As expected, the powers in the two-samples t-test for those after power-transformation were almost preserved to 0.9. In the case of $\lambda = 1, 0.5, 0$ in $\rho = 0.75, 0.9$, these results were similar to those in one-sample problem, and the powers in three tests were the almost same 0.9.

In the case of $\lambda = -0.5, -1$ in $\rho = 0.75, 0.9$, the powers in the two-samples t-test for pre- and post-observations decreased largely as τ increased, and the extent of the decreases was larger than it in one-sample problem. Especially, the power was under 0.5 in $\lambda = -1$ and $\tau = 0.35$. Though the powers in the Wilcoxon rank-sum test were also decreasing as τ increased, the tendency was with a slower pace than those in the two-samples t-test for pre- and post-observations.

To find the causes of why the powers in the two-samples t-test for pre- and post-observations in the cases of $\lambda = -1, -0.5$, we showed the percent points (5%, 25%, 50%, 75%, 95%) of the test statistics, the difference of the sample means and the pooled variance in two groups for the simulated 50,000 data (without the details). As in one-sample problem, the test statistics in $\lambda = -0.5, -1$ were smaller than those in $\lambda = 0, 0.5, 1$ and the range of the square roots of the pooled variance and the difference of the sample means in $\lambda = -0.5, -1$ was wider than the those in $\lambda = 0, 0.5, 1$. Thus, it is considered that the test statistics decreased as the square roots of the pooled variance increased, so the powers decreased. Again, we can consider that it has influence on the powers in the two-samples t-test that the distributions of the difference with $\lambda = -1, -0.5$ and $\tau = 0.35$ are longer tailed in both sides, compared to the distributions

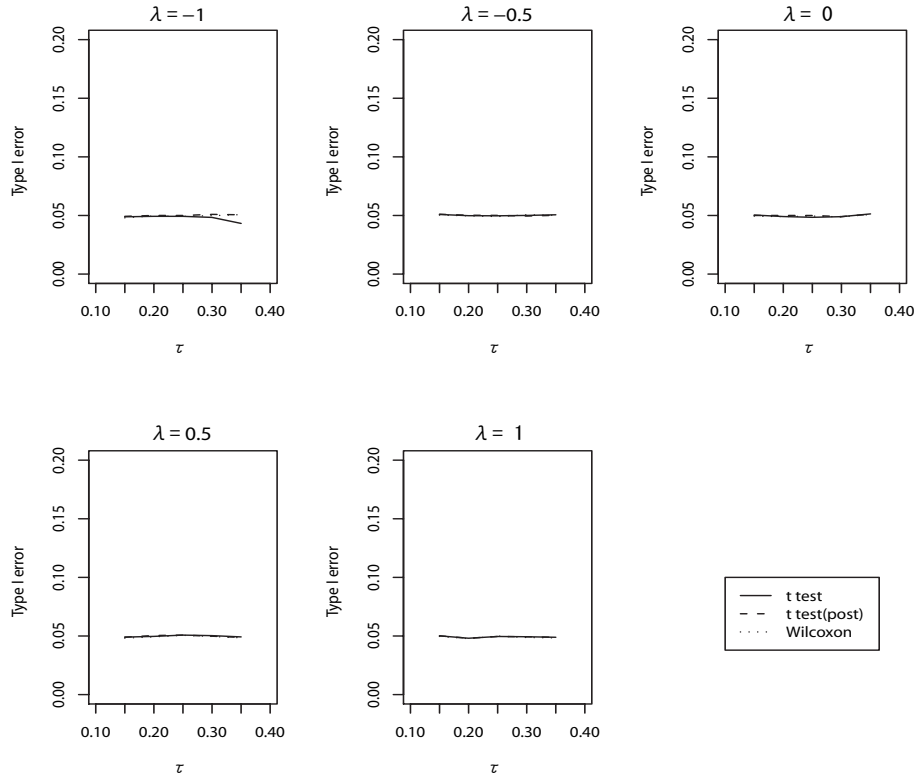


Figure 4.10: Two-sample: Relationship between τ and type I error ($\rho = 0.75$)

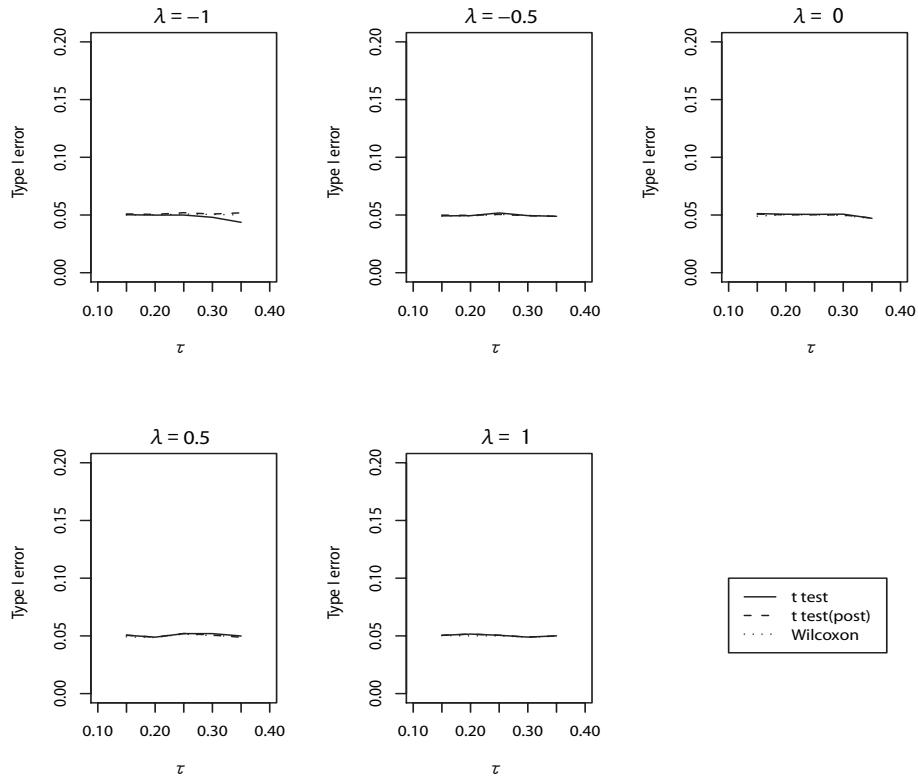


Figure 4.11: Two-sample: Relationship between τ and type I error ($\rho = 0.9$)

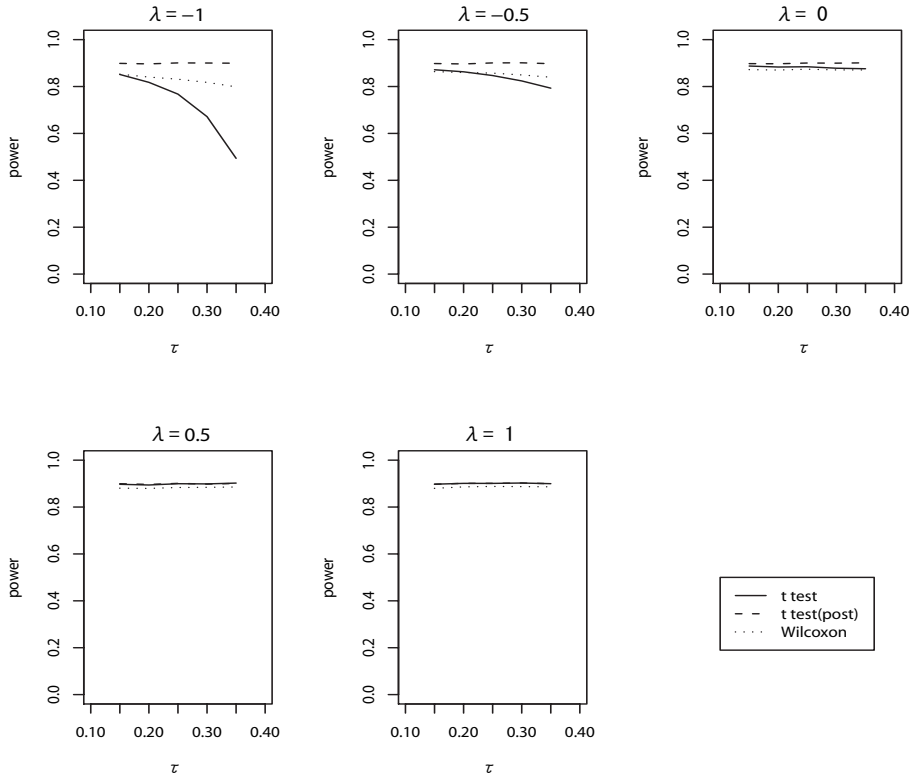


Figure 4.12: Two-sample: Relationship between τ and power ($\rho = 0.75$)

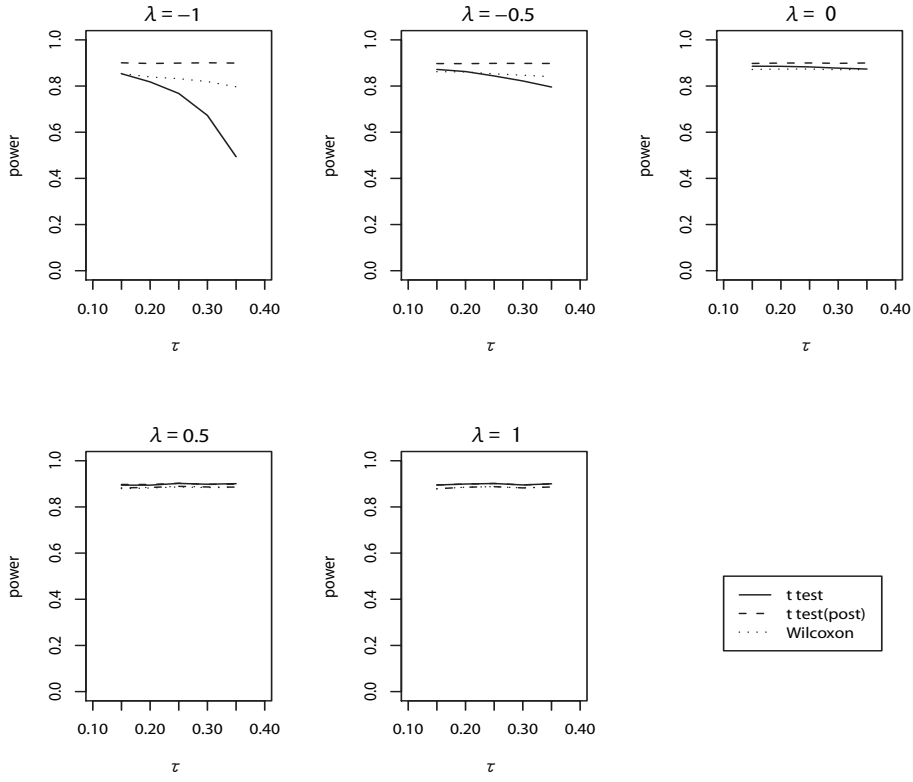


Figure 4.13: Two-sample: Relationship between τ and power ($\rho = 0.9$)

of the difference with $\lambda = 1$ (normal distribution). Also the correlation coefficient parameter in BPNDs had little influence on the type I errors and the powers in this simulation.

From these results, we found that the two-samples t-test is robust even if the assumptions of the normality are slightly violated.

4.5 Conclusion

In this paper, we comprehensively discussed how the potential distribution which the observations of pre- and post-treatments follow impacts the distribution of the difference and test results. As the result, regardless of the coefficient parameter, the distribution of the difference approached to the symmetry even if the distribution which pre- and post-observations follow was right-skewed. From the results of the simulations which were conducted to examine the impact of the distribution of the difference on t-test which assumes normality, even when the potential distribution was actually log-normal distribution though it was assumed that the potential distribution was a normal-distribution, the powers in the paired t-test and the two-samples t-test were preserved high.

So we found that the paired and two-samples t-test were robust even if the assumptions of the normality were slightly violated. But when the potential distribution is longer tailed to the right than log-normal distribution and the variation of the distribution is large, the power decreased remarkably because the distribution of the difference was longer tailed in both sides.

Thus, we found that the potential distribution has influence on the distribution of the difference and the paired and two-samples t-test. It is desirable to interpret the test results after making clear the potential distribution which pre- and post-treatments follow at first and having a sufficient understanding of the characteristics of the distribution of the difference.

Reference

1. Blair, R.C. and Higgins, J.J. (1985). A comparison of the power of the paired samples rank transform statistics to that of Wilcoxon's signed ranks statistic. *JEBS* **10**(4), 368-383.

2. Bonate, P.L. (2000). Analysis of pretest-posttest designs. CHAPMAN and HALL/CRC.

3. Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations (with discussion). *J Roy Stat Soc Ser B* **26**(2), 211-246.

4. Goto, M. and Hamasaki, T. (2002). The bivariate power-normal distribution. *Bull Informatics Cybern* **34**(1), 29-49.

5. Goto, M. and Matsubara, Y. and Tsuchiya, Y. (1983). Power-normal distribution and its applications. *Rep Stat Appl Res JUSE*, **30**, 8-28.

6. Goto, M., Uesaka, H. and Inoue, T. (1979). Some linear models for power transformed data. Invited paper at the 10th International Biometric Conference. August, 6-10.(Res Rep NO.93, Res Instit Fund Infor Sc, Kyusyu University)

7. Isogawa, N., Ikebe, T., Sakamoto, W. and Goto, M. (2011). A preliminary evaluation about health guidance. *J Behaviormetrics* **38**(1), 51-63.

8. Maruo, K. and Goto, M. (2008). On estimation of parameters in power-normal distribution. Joint Meeting of 4th World Conference of the IASC and 6th Conference of the Asian Regional Section of the IASC on Computational Statistics & Data Analysis

9. Maruo, K., Shirahata, S., Goto, M. and Komazawa, T. (2008). Statistical investigation of reference intervals of clinical laboratory data. *J Behaviometrics*, **35**(1), 73-89.

10. Yang, L. and Tsiatis, A.A. (2001). Efficiency study of estimators for a treatment effect in a pretest-posttest trial. *Amer Statist* **55**(4), 314-321.

5. Conclusions

In this paper, we focused on three different topics. They are "Predictive performance of Bayesian diagnoses", "A preliminary evaluation about health guidance" and "The impact of the shape of the underlying distribution". In chapter 2, we explained about BPIC and the predictive checking approach, and described new findings obtained from the simulation to make clear the the predictive performance. In chapter 3, we conducted a preliminary evaluation about health guidance for data of 1,141 subjects who had the health checkup that was carried out in April 2004. As the results, we found that it was very important to interpret data through a process of data analysis which are conscious of logic consistency. In chapter 4, we examined the impact of the shape of the underlying distribution of observations on test results and specifically present occasions where t-test works well. We found that the paired and two-samples t-test were robust even if the assumptions of the normality were slightly violated. But when the potential distribution is longer tailed to the right than log-normal distribution and the variation of the distribution is large, the power decreased remarkably because the distribution of the difference was longer tailed in both sides. In this chapter, we propose some subjects for future investigation.

5.1 Future problem

Predictive performance of Bayesian diagnoses: We considered about the case that sample follow normal distribution with known variance and mean parameter follow prior distribution. We think that we can clarify the characteristics of BPIC and predictive checking approach even in the case of that sample follow another distribution except for normal distribution or that the number of parameters which have prior distribution increases. To maximize the advantage of Bayesian approach which can select appropriate model in terms of prediction, it is very important to make a clear the profiles of these predictive diagnoses in the application situation before conducting predictive diagnoses.

Also prior and posterior predictive checking approaches have been under the development yet

and so it is expected for the application to various occasions. For example, in actual situation, though model diagnosis which is conducted by residual display and another effective plotting is unprogrammed, we can re-consider them using prior and posterior predictive checking approach through the logical framework. These attempts may lead to propose direct model evaluation (Okuda, 1999). As our goals, we would like to connect Neyman-Pearson to Bayesian, which have been developed separately by now, through predictive checking approach.

A preliminary evaluation about health guidance: As described in section 3.1, though Ministry of Health, Labor and Welfare of Japan has carried out "Health Checkups and Healthcare Advice" which make it obligatory for person aged 40 through 74 to reduce medical expenses and prevent lifestyle-related diseases (Health Service Bureau of Health, Labour and Welfare, 2007) since April 2008, it is concerned with the lack of "Foundation for enforcement" and "Evidence for prevention" (Ohgushi, 2006: 2007).

We also think that the effect of "Health Checkups and Healthcare Advice" should be clarified. In chapter 2, based on the results of the Health Checkup that aims to prevent disease are carried out in April 2004, we explored foundation about the doctor's judgment, especially classification of the directed group, attempting to figure the doctor's character, and further evaluated directed effect for the directed group. Through a process of data analysis which is conscious of logic consistency, we gained the findings of an evaluation and consideration for health care advice from "Set of cold figures" (health checkup data) (Goto, 1986). If possible, we would like to analyze the actual data of "Health Checkups and Healthcare Advice" and examine the effects in fact.

The impact of the shape of the underlying distribution: We comprehensively discussed how the potential distribution which the observations of pre- and post-treatments follow impacts the distribution of the difference and test results. As the result, regardless of the coefficient parameter, the distribution of the difference approached to the symmetry even if the distribution which pre- and post-observations follow was right-skewed. Also we found that the paired and two-samples t-test were robust even if the assumptions of the normality were slightly violated.

As the future problem, we would like to clarify the relationship between the difference and pre-/post-observation numerically and examine how the difference distribution expresses the information of pre- and post-observations.

List of publications

- Isogawa, N. (2008). A preliminary evaluation about health guidance. *Proceedings of the 92th Symposium of Behaviormetrics*, Okayama, Japan (in Japanese).

- Isogawa, N. (2009). Predictive checking function and the performance evaluation: New development of Bayes approach. *Unpublished Master thesis*, Osaka University (in Japanese).

- Isogawa, N. (2010). Data replication based on Bayesian approach. *The 24th Symposium of the Japanese Society of Computational Statistics*, 101-104, Osaka, Japan (in Japanese).

- Isogawa, N. (2011). Predictive performance of Bayesian diagnoses. *The 58th World Statistics Congress of the Internetional Statistical Institute*, Dublin, Ireland.

- Isogawa, N. and Goto, M. (2011). Predictive performance of Bayesian diagnoses. *Journal of the Japanese Society of Computational Statistics*, 24(2) (in Japanese) (in press).

- Isogawa, N., Ikebe, T., Sakamoto, W. and Goto, M. (2009). A preliminary evaluation about health guidance. *The 37th Annual Meeting of Behaviormetrics*, 104-105, Oita, Japan (in Japanese).

- Isogawa, N., Ikebe, T., Sakamoto, W. and Goto, M. (2011). A preliminary evaluation about health guidance. *The Japanese Journal of Behaviormetrics*, 38(1), 51-63 (in Japanese).

- Isogawa, N., Sakamoto, W. and Goto, M. (2009). Model diagnosis using predictive checking function. *The 23th Symposium of Behaviormetrics*, 17-20, Fukuoka, Japan (in Japanese).

- Isogawa, N., Sakamoto, W., Shirahata, S. and Goto, M. (2008). Evaluation of prior and posterior predictive checking function. *The 22th Symposium of Behaviormetrics*, 203-206, Hyogo, Japan (in Japanese).

- Isogawa, N., Shirahata, S. and Goto, M. (2008). Effect of asymmetry in underlying distributions on performance of paired tests. *Joint Meeting of 4th World Confernce of the*

IASC and 6th Conference of the Asian Regional Section of the IASC on Computational Statistics & Data Analysis(2008.12.5-8), Program & Abstracts, 131, Yokohama, Japan.

- Sakamoto, W., Isogawa, N. and Goto, M. (2008). Statistical issues on Japanese criteria of metabolic syndrome. *The Japanese Journal of Behaviormetrics*, 35(2), 177-192 (in Japanese).