รายงานวิจัยฉบับสมบูรณ์

โครงการวิธีสำหรับตัวแบบสถิติที่มีข้อจำกัดการจัดลำดับ

โดย

รองศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์

มิถุนายน 2560

สัญญาเลขที่ RSA5780005

รายงานวิจัยฉบับสมบูรณ์

โครงการวิธีสำหรับตัวแบบสถิติที่มีข้อจำกัดการจัดลำดับ

โดย

รองศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์

จุฬาลงกรณ์มหาวิทยาลัย

สนับสนุนโดยสำนักงานกองทุนสนับสนุนการวิจัย
และจุฬาลงกรณ์มหาวิทยาลัย

กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณสำนักงานกองทุนสนับสนุนการวิจัย (สกว.) จุฬาลงกรณ์มหาวิทยาลัย และคณะพาณิชยศาสตร์และการบัญชีแห่งจุฬาลงกรณ์มหาวิทยาลัย ที่ได้ให้การสนับสนุนทุนวิจัยตลอดโครงการวิจัย

รองศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์
จุฬาลงกรณ์มหาวิทยาลัย
2560

บทคัดย่อ

งานวิจัยนี้เป็นการนำเสนอวิธีทางสถิติสำหรับตัวแบบสถิติที่มีข้อจำกัดการจัดลำดับ ซึ่งประกอบด้วยผลงานสามชิ้น ผลงานชิ้นแรกนำเสนอการคำนวณการแจกแจงความน่าจะเป็นและโมเมนต์สำหรับตัวแบบสถิติที่มีข้อจำกัดการจัดลำดับในสองกรณีคือ กรณีที่ตัวแปรสุ่มเป็นอิสระกัน และกรณีที่ตัวแปรสุ่มไม่เป็นอิสระกันแต่สามารถแทนด้วยตัวแบบปัจจัยเดียว วิธีที่เสนอในงานวิจัยสามารถนำไปประยุกต์ใช้ในปัญหาการตัดสินใจเกี่ยวกับตัวแปรที่ไม่ทราบค่า แต่ทราบข้อมูลลำดับของตัวแปรนั้น คุณสมบัติทางสถิติของตัวแปร เช่น การแจกแจงก่อนและโมเมนต์ของตัวแปรจะถูกปรับค่าด้วยข้อจำกัดการจัดลำดับ เพื่อคำนวณเป็นการแจกแจงภายหลังที่ถูกต้องมากขึ้น ซึ่งโดยทั่วไปแล้วการคำนวณสำหรับปัญหาลักษณะนี้จะซับซ้อนขึ้นเมื่อข้อมูลมีมิติสูงขึ้น แต่งานวิจัยนี้แสดงการประยุกต์ใช้เทคนิคปริพันธ์เวียนเกิด เพื่อแปลงการคำนวณที่ซับซ้อนให้อยู่ในรูปของการคำนวณหาปริพันธ์ในหนึ่งมิติในกรณีของตัวแปรที่เป็นอิสระกัน หรือสองมิติในกรณีของตัวแปรที่ไม่เป็นอิสระกันแต่สามารถแทนด้วยตัวแบบปัจจัยเดียว ในผลงานชิ้นแรกยังได้แสดงการประยุกต์วิธีที่นำเสนอในการแก้ปัญหาการจัดพอร์ตการลงทุนกับข้อมูลจริง ผลงานชิ้นที่สอง และผลงานชิ้นที่สาม เป็นการประยุกต์ตัวสถิติโคโมโกรอฟในการอนุมานสถิติหลายตัวแปรพร้อม ๆ กัน ตัวสถิติโคโมโกรอฟจัดว่าเป็นตัวแบบสถิติที่มีข้อจำกัดการจัดลำดับ ซึ่งการแจกแจงความน่าจะเป็นสามารถคำนวณด้วยวิธีที่เสนอในผลงานชิ้นแรก ผลงานชิ้นที่สองเป็นการประยุกต์ตัวสถิติโคโมโกรอฟในการสร้างแถบความเชื่อมั่นของฟังก์ชันการแจกแจงแบบเบต้า และนำไปประยุกต์กับการจัดการความเสี่ยงด้านเครดิต ผลงานชิ้นที่สามเป็นการประยุกต์ตัวสถิติโคโมโกรอฟในการอนุมานสถิติสำหรับความน่าจะเป็นที่จะชนะเพื่อเปรียบเทียบตัวแบบไวบูลย์สองตัวแบบ ซึ่งผลของการเปรียบเทียบจะเป็นประโยชน์กับการตัดสินใจในงานด้านความเชื่อถือได้ของระบบ

Abstract

In this research, we propose methods for statistical models with ranking constraints, consisting of three papers. In the first paper, we discuss methods for probability and moment calculations of statistical models with ranking constraints in two important cases: the case where the statistical variables are independent and the case where the statistical variables are dependent but can be written in a one factor model. This can be useful for decision making when we do not observe the actual values of the variables, but we do observe the ordering of the variables. In such a case, prior information on the distributions and moments from the variables' specified distributions can be updated by the observed ranking to provide improved posterior information. While the calculations of the rank updated posterior distribution ostensibly involve high-dimensional integral calculations, it is shown how the recursive integration methodology can be applied so that the original high-dimensional integral can be evaluated as a series of one-dimensional integration for the case of independent variables or a two-dimensional integration for the case of dependent variables with one-factor model. In the first paper, we also show how to apply the proposed methods to solve a portfolio selection problem with a real data set. In the second and the third papers, we apply the Kolmogorov statistic to multiple comparison inference problems. The Kolmogorov statistic can be considered a statistical model with ranking constraints whose probability distribution can be computed by the methods proposed in the first paper. In the second paper, the Kolmogorov statistic is applied to construct an exact confidence band for a beta distribution function. Its application can be found in credit risk management. In the third paper, the Kolmogorov statistic is applied to win-probabilities for comparing two Weibull models. The results from the comparison will be useful for decision making in system reliability.

Project Code : RSA5780005
(รหัสโครงการ)

Project Title : Methods for Statistical Models with Ranking Constraints
(ชื่อโครงการ) โครงการวิธีสำหรับตัวแบบสถิติที่มีข้อจำกัดการจัดลำดับ

Investigator : Associate Professor Seksan Kiatsupaibul, Ph.D.
(ชื่อนักวิจัย) รองศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์

E-mail Address : seksan@cbs.chula.ac.th

Project Period : 16 June 2014 – 16 June 2017
(ระยะเวลาโครงการ) วันที่ 16 มิถุนายน 2557 ถึงวันที่ 16 มิถุนายน 2560

คำสำคัญ
การแจกแจงแบบมีเงื่อนไข
การแจกแจงแบบปกติ
ข้อจำกัดการจัดลำดับ
การสุ่มตัวอย่างแบบกลุ่มอันดับ
เทคนิคปริพันธ์เวียนเกิด

Keywords :
Conditional distribution
Normal distribution
Order restriction
Ranked set sampling
Recursive integration

1. บทนำ

รายงานฉบับนี้เป็นการสรุปผลงานจากโครงการวิธีสำหรับตัวแบบสถิติที่มีข้อจำกัดการจัดลำดับ ซึ่งประกอบด้วยสามผลงาน รายงานสรุปเน้นไปที่ผลงานแรกซึ่งเป็นบทความเรื่องการคำนวณการแจกแจงและโมเมนต์ที่มีข้อจำกัดการจัดลำดับโดย Kiatsupaibul et.al. (2017a) ข้อสรุปผลงานที่สองและสามในโครงการฯ จะปรากฏในส่วนสรุปผล

การคำนวณหาการแจกแจงและค่าคาดหวังแบบมีเงื่อนไขเป็นหลักการพื้นฐานทางสถิติที่นำไปประยุกต์ใช้ในการศึกษาทั้งทางด้านวิทยาศาสตร์ สังคมศาสตร์ และวิศวกรรมศาสตร์ โดยเฉพาะการคำนวณตามทฤษฎีของเบส์ แต่สำหรับในกรณีที่ข้อมูลประกอบอยู่ในรูปข้อมูลเชิงคุณภาพ การแจกแจงที่สนใจจะถูกจำกัดโดยเงื่อนไขของรูปหลายมิติ (polytope) ซึ่งการคำนวณค่าสำหรับปัญหาในลักษณะนี้ด้วยวิธีเชิงตัวเลข (numerical approach) ทำได้ยากและซับซ้อน (ตามตัวอย่างใน Khachiyan (1989)) ทำให้นักทฤษฎีและนักปฏิบัติส่วนมากหันไปใช้วิธีการจำลองข้อมูลด้วยวิธี Monte Carlo ในการคำนวณแทน (ตามตัวอย่างใน Smith (1984), Lovász (1999), Lovász and Vempala (2006) และ Kiatsupaibul et al. (2011)) อย่างไรก็ตามวิธีการเชิงตัวเลขมีข้อดีและมีประสิทธิภาพเหนือกว่าวิธีการจำลองข้อมูลในหลายด้าน งานวิจัยนี้จะนำเสนอวิธีการคำนวณเชิงตัวเลข ที่สามารถนำไปใช้ในการคำนวณการแจกแจงและโมเมนต์แบบมีเงื่อนไขของตัวแปร โดยจะพิจารณาในกรณีที่ข้อจำกัดดังกล่าวอยู่ในรูปของข้อมูลอันดับ (rankings) ซึ่งถือเป็นข้อจำกัดสำคัญที่พบได้ทั่วไปในทางปฏิบัติ

พิจารณาตัวแปรสุ่มแบบต่อเนื่อง (X_i) โดยกำหนดให้มีการแจกแจงเป็น $f_i(x_i), 1 \leq i \leq n$ ซึ่งโมเมนต์ของตัวแปรสุ่มจะถูกกำหนดตามการแจกแจงนี้ด้วย และให้ข้อมูลประกอบที่จะมาปรับการแจกแจงนี้อยู่ในรูปของข้อมูลอันดับ ดังนี้

$$X_1 \leq X_2 \leq \cdots \leq X_n \tag{1}$$

งานวิจัยนี้มีวัตถุประสงค์เพื่อแสดงวิธีการคำนวณการแจกแจงภายหลังและโมเมนต์ของตัวแปรสุ่ม X_i ที่ปรับด้วยข้อมูลอันดับ ซึ่งจะสามารถนำไปใช้สำหรับปัญหาในทางปฏิบัติ กรณีที่ไม่ทราบค่าของตัวแปร แต่สามารถจัดลำดับของตัวแปรดังกล่าวไป งานวิจัยด้าน Ranked Set Sampling แสดงตัวอย่างของสถานการณ์ในลักษณะนี้ เช่น งานศึกษาของ McIntyre (1952), Patil (2002), Chen et al. (2004) และ Wolfe (2004) ข้อแตกต่างของการใช้ข้อมูลอันดับในเรื่อง Ranked Set Sampling กับในงานวิจัยนี้คือ ในด้าน Ranked Set Sampling ตัวแปรสุ่มจะถูกวัดค่าในตำแหน่งที่กำหนดสำหรับแต่ละขั้นตอน โดยข้อมูลอันดับของตัวแปรจะถูกใช้ในการกำหนดตำแหน่งของข้อมูลที่จะถูกวัดค่า แต่ในบริบทของงานวิจัยนี้ข้อมูลอันดับจะถูกใช้เพื่อปรับการแจงแจงก่อน โดยอาจเป็นกรณีที่ค่าที่แท้จริงของตัวแปรไม่ได้ถูกวัดค่าเลยด้วย แม้ว่าบริบทของการวัดค่าที่แท้จริงของตัวแปรตามที่กล่าวมาข้างต้นสำหรับเรื่อง Ranked

Set Sampling กับในงานวิจัยนี้จะแตกต่างกัน แต่ทั้งคู่ถือเป็นการใช้ประโยชน์จากข้อมูลอันดับโดยไม่ทราบค่าแท้จริงของตัวแปรทุกค่าเช่นเดียวกัน

ยกตัวอย่างเช่น กำหนดให้การแจกแจงก่อนของตัวแปรสุ่ม X_i แทนระดับอาการของคนไข้ n คน ที่รวบรวมจากการวัดค่าทางการแพทย์ที่เกี่ยวข้อง จะเห็นว่าค่าที่ถูกต้องของระดับอาการประเมินเป็นตัวเลขแน่นอนได้ยากในทางปฏิบัติ แต่แพทย์สามารถใช้ข้อมูลของตัวแปรมีความสัมพันธ์กับระดับอาการที่ต้องการวัด เช่น ค่าต่างๆ ที่ได้จากผลการวิเคราะห์เลือด มารวบรวมและสร้างข้อมูลอันดับของ X_i เพื่อนำไปปรับการแจกแจงก่อน แล้วคำนวณคาดหวังและความแปรปรวนของการแจกแจงภายหลังเพื่อค่าดังกล่าวไปใช้ในการตัดสินใจต่อไป

นอกจากนี้ยังมีตัวอย่างในงานของ Patil (2002) ที่แสดงให้เห็นประยุกต์ใช้ในด้านการสำรวจพื้นที่อันตราย ที่การวัดค่าการปนเปื้อนของสารพิษมีความเสี่ยงและต้นทุนสูง จึงใช้ลักษณะของดิน เช่น คราบ สี หรือความหนาแน่นของพืชในพื้นที่ มาช่วยในขั้นตอนการจัดอันดับ แล้วนำข้อมูลอันดับดังกล่าวไปปรับการแจกแจงก่อนและโมเมนต์ของตัวแปรระดับการปนเปื้อน (X_i) ที่สนใจศึกษา

โดยทั่วไปข้อมูลอันดับที่นำมาคำนวณสามารถหาได้จากหลายแหล่ง เช่น ได้จากการรวบรวมข้อมูลของตัวแปรร่วมที่มีความสัมพันธ์กับตัวแปรที่สนใจศึกษา อย่างเช่นในงานของ Topkis (1998), Milgrom and Roberts (1994) และ Milgrom and Shannon (1994) ที่ใช้จัดอันดับโดยใช้ข้อมูล-ภาวะเศรษฐกิจด้านต่างๆ หรือข้อมูลอันดับอาจได้จากกระบวนการรวบรวมค่าที่ได้จากการจัดลำดับตามความนิยม เช่นงานของ Kemeny and Snell (1962), Young (1995) และ Ali and Meilă (2012) สำหรับวิธีการคำนวณที่อธิบายไว้ในงานวิจัยนี้เป็นเครื่องมือสำคัญที่สามารถใช้รวบรวมข้อมูลอันดับเข้ากับการแจกแจงก่อน เพื่อทำการอนุมานเชิงสถิติ อย่างเช่นในงานศึกษาของ Chiarawongse et al. (2012) ซึ่งผลการคำนวณที่ได้จะนำไปประยุกต์ใช้สำหรับปัญหาที่เกี่ยวข้องในงานศึกษาด้านต่างๆ ได้หลายด้าน ดังเช่นที่ได้ยกตัวอย่างมาข้างต้น

ในการคำนวณการแจกแจงและโมเมนต์แบบมีเงื่อนไขของตัวแปรสุ่ม X_i โดยทั่วไปจะถูกมองเป็นปัญหาการหาปริพันธ์ใน n มิติ แต่ในงานวิจัยนี้จะแสดงให้เห็นว่าสามารถใช้เทคนิคปริพันธ์เวียนเกิด (ที่ได้อธิบายไว้ในงานศึกษาของ Hayter (2006)) มาปรับรูปแบบการคำนวณให้อยู่ในรูปแบบของการหาปริพันธ์ใน 1 มิติได้ รวมทั้งจะแสดงให้เห็นด้วยว่า การนำวิธีการดังกล่าวไปใช้ไม่ได้จำกัดอยู่เพียงการคำนวณเงื่อนไขของข้อมูลอันดับแบบง่ายๆ แต่ยังสามารถนำไปใช้กับปัญหาทั่วไปที่ซับซ้อนขึ้น เช่น กรณีที่ตัวแปรสุ่มมีการแจกแจงแบบปกติหลายตัวแปร และมีโครงสร้างความสัมพันธ์อยู่ในรูปแบบที่กำหนด ได้อีกด้วย

งานวิจัยนี้ยังได้แสดงตัวอย่างการประยุกต์ใช้ในด้านต่างๆ เพื่อแสดงให้เห็นถึงผลของข้อจำกัดการจัดลำดับแบบต่างๆ ที่มีต่อการแจกแจง ค่าคาดหวัง และความแปรปรวนของตัวแปรสุ่ม X_i

โดยเฉพาะอย่างยิ่ง ผลของข้อมูลอันดับที่สอดคล้องกับค่าคาดหวังของการแจกแจงก่อน $f_i(x_i)$ ที่เรียกว่า reinforcing ranking เปรียบเทียบกับกรณีที่ข้อมูลอันดับไม่สอดคล้องกับค่าคาดหวังของการแจกแจงก่อนหรือที่เรียกว่า opposing ranking โดยจะศึกษาเมื่อระดับความไม่สอดคล้องนี้แตกต่างกันด้วย โดยผลการศึกษาจะชี้ให้เห็นถึงความแตกต่างของค่าคาดหวังและความแปรปรวนของการแจกแจงภายหลังที่คำนวณได้จากการปรับด้วยข้อมูลอันดับแบบต่างๆ รวมทั้งจะมีตัวอย่างของการประยุกต์ใช้เทคนิคนี้กับปัญหาการจัดพอร์ตการลงทุน เพื่อแสดงให้เห็นถึงประโยชน์ของการคำนวณเมื่อมีข้อมูลประกอบเชิงอันดับ โดยจะแสดงผลของการจัดพอร์ตการลงทุนเมื่อใช้ข้อมูลผลตอบแทนจริงของหลักทรัพย์มาคำนวณ

โครงสร้างของงานวิจัยนี้ประกอบด้วย การอธิบายทฤษฎีและหลักการใช้เทคนิคปริพันธ์เวียนเกิดในการคำนวณหาการแจกแจงและโมเมนต์เมื่อมีข้อจำกัดการจัดลำดับ โดยจะแสดงการคำนวณทั้งในกรณีที่ตัวแปรเป็นอิสระกัน และกรณีที่ตัวแปรมีการแจกแจงแบบปกติหลายตัวแปรที่มีโครงสร้างความสัมพันธ์ตามที่กำหนด จากนั้นในหัวข้อถัดไป จะอธิบายกระบวนการและขั้นตอนการคำนวณโดยละเอียด รวมทั้งแสดงให้เห็นถึงกลไกการลดความคลาดเคลื่อน โดยจะแสดงให้เห็นอัตราการผิดพลาดและเวลาที่ใช้ในการคำนวณประกอบกัน ถัดมาจะเป็นการยกตัวอย่างการประยุกต์ใช้เทคนิคนี้กับปัญหาในงานศึกษาด้านต่างๆ และหัวข้อสุดท้ายจะเป็นข้อสรุปที่ได้จากงานวิจัยนี้

2. ทฤษฎีที่เกี่ยวข้อง

หัวข้อนี้จะอธิบายทฤษฎีที่เกี่ยวกับการใช้เทคนิคปริพันธ์เวียนเกิดในการคำนวณค่าทางสถิติ เช่น ค่าโมเมนต์เมื่อมีข้อจำกัดการจัดลำดับ โดยจะเริ่มจากการอธิบายในกรณีที่ตัวแปรสุ่มเป็นอิสระกันก่อน จากนั้นจึงจะแสดงการคำนวณในกรณีที่ตัวแปรสุ่มที่มีการแจกแจงแบบปกติหลายตัวแปรที่มีโครงสร้างความสัมพันธ์ตามที่กำหนด สุดท้ายจะกล่าวถึงการประยุกต์ใช้ในปัญหาที่ซับซ้อนขึ้น

2.1 ตัวแปรสุ่มที่เป็นอิสระกัน

กำหนดให้เซต $S \subseteq \Re^n$ ของค่าของตัวแปร $\boldsymbol{X} = (X_1, \dots, X_n)$ เป็น

$$S = S_{1,2} \cap S_{2,3} \cap \dots \cap S_{n-1,n}$$

เมื่อเซต $S_{i,i+1}$ เป็นข้อจำกัดสำหรับค่า X_i และ X_{i+1} เท่านั้น

ดังนั้น เซต S จะสอดคล้องกับอันดับตามสมการที่ (1) ยกตัวอย่างเช่น

$$S_{i,i+1} = \{ \boldsymbol{X} : X_i \leq X_{i+1} \}$$

เมื่อ $1 \leq i \leq n-1$

สำหรับทุกช่วง $(l_i, u_i), 1 \leq i \leq n+1$ จะได้ว่า

$$P(l_i \le X_i \le u_i; 1 \le i \le n \,|\boldsymbol{X} \epsilon S) = \frac{A_1}{B} \quad (2)$$

เมื่อ

$$A_1 = \int_{X \in S^*}^{\cdots} \int \prod_{i=1}^{n} f_i(x_i) dx_1 \dots dx_n$$

และ

$$B = P(\boldsymbol{X} \in S) = \int_{X \in S}^{\cdots} \int \prod_{i=1}^{n} f_i(x_i) dx_1 \dots dx_n$$

โดย

$$S^* = S^*_{1,2} \cap S^*_{2,3}, \cap \dots S^*_{n-1,n}$$

สำหรับ

$$S^*_{1,2} = S^*_{1,2} \cap \{ \boldsymbol{X} : l_1 \le X_i \le u_1, l_2 \le X_i \le u_2 \}$$

และ

$$S^*_{i,i+1} = S^*_{i,i+1} \cap \{ \boldsymbol{X} : l_{i+1} \le X_{i+1} \le u_{i+1} \}$$

เมื่อ $2 \le i \le n-1$

เช่นเดียวกับ $\mathrm{g}_i(x_i), 1 \le i \le n$ ใดๆ จะได้ว่า

$$E[\mathrm{g}_1(x_1)\mathrm{g}_2(x_2) \dots \mathrm{g}_n(x_n) | \boldsymbol{X} \in S] = \frac{A_2}{B} \quad (3)$$

เมื่อ

$$A_2 = \int_{X \in S^*}^{\cdots} \int \prod_{i=1}^{n} (\mathrm{g}_i(x_i) f_i(x_i)) dx_1 \dots dx_n$$

แม้ว่า A_1, A_2 และ B จะดูเหมือนเป็นการหาปริพันธ์ใน n มิติ แต่จากงานศึกษาของ Hayter (2006) ส่วนที่ 1 ระบุไว้ว่าการคำนวณดังกล่าวสามารถจัดให้อยู่ในรูปแบบที่ $d = 1$ ได้ ดังนั้นจึงสามารถใช้หลักการของเทคนิคปริพันธ์เวียนเกิดสำหรับข้อมูล 1 มิติ มาคำนวณได้ไม่ว่าข้อมูลจะมีขนาด n เป็นเท่าใดก็ตาม ดังนั้น ค่าความน่าจะเป็นในสมการที่ (2) และค่าคาดหวังในสมการที่ (3) ก็สามารถคำนวณได้โดยใช้เทคนิคปริพันธ์เวียนเกิดสำหรับข้อมูล 1 มิติเช่นเดียวกัน

สังเกตว่าฟังก์ชันความน่าจะเป็นสะสมร่วมแบบมีเงื่อนไขของตัวแปรสุ่ม X_i สามารถคำนวณได้จากสมการที่ (2) เมื่อ $l_i = -\infty, 1 \le i \le n$ ส่วนการแจงแจงขอบแบบมีเงื่อนไขของตัวแปรสุ่มใดๆ จะคำนวณโดยกำหนดให้ $l_i = -\infty$ และ $u_i = \infty$ สำหรับตัวแปรอื่นๆ ค่าโมเมนต์แบบมีเงื่อนไขของตัวแปรสุ่ม X_i คำนวณโดยกำหนดให้ $\mathrm{g}_i(x_i) = x_i^k$ และให้ฟังก์ชัน $\mathrm{g}_i(x_i)$ อื่นๆ มีค่าเป็น 1 นอกจากนี้ยังสามารถหาค่าความแปรปรวนร่วมแบบมีเงื่อนไขระหว่างตัวแปรสุ่ม X_{i_1} และ X_{i_2} ได้โดยกำหนดให้

$g_{i_1}(x_{i_1}) = x_{i_1}$ และ $g_{i_2}(x_{i_2}) = x_{i_2}$ แล้วให้ฟังก์ชัน $g_i(x_i)$ อื่นๆ มีค่าเป็น 1 เช่นเดียวกับการคำนวณโมเมนต์

2.2 *ตัวแปรสุ่มที่มีการแจกแจงแบบปกติหลายตัวแปรที่มีโครงสร้างความสัมพันธ์ตามที่กำหนด*

ในกรณีที่ตัวแปรสุ่ม X_i มีการแจกแจงแบบปกติหลายตัวแปร โดยมีค่าเฉลี่ย μ_i ความแปรปรวน σ_i^2 และค่าความแปรปรวนร่วมเป็น $\rho_i\rho_j$ แล้ว จะสามารถเขียนความสัมพันธ์ได้ตามสมการ ดังนี้

$$X_i = \mu_i + \rho_i M + \sqrt{\sigma_i^2 - \rho_i^2} Z_i \, , 1 \le i \le n \qquad (4)$$

เมื่อ M และ Z_i เป็นตัวแปรสุ่มมีการแจกแจงแบบปกติมาตรฐานและเป็นอิสระกัน ภายใต้เงื่อนไขของค่า M ตัวแปรสุ่ม X_i จะมีการแจกแจงปกติที่อิสระกันด้วย

การประมาณค่าตามสมการที่ (2) และ (3) ภายใต้เงื่อนไขของค่า M จำเป็นต้องคำนวณหาปริพันธ์ 1 มิติ เทียบกับแต่ละค่า m ของ M โดยการคำนวณค่าปริพัทธ์ (integrand) เทียบกับแต่ละค่า m เป็นการคำนวณใน 1 มิติ โดยรวมแล้วจึงเปรียบเสมือนการคำนวณหาค่าปริพันธ์ทั้งหมดเป็นการคำนวณใน 2 มิติ ไม่ว่าตัวแปรที่พิจารณาจะมีขนาด n เป็นกี่มิติก็ตาม

และหากกำหนดให้ค่าความแปรปรวนร่วม ρ_i มีค่าเป็นบวกเท่ากันทุกค่า เขียนแทนด้วยสัญลักษณ์ ρ แล้ว สำหรับข้อมูลลำดับของตัวแปรตามสมการที่ (1) จะได้ว่าเซต S จะขึ้นอยู่กับค่า Z_i โดยไม่ขึ้นกับค่า M ในกรณีนี้การคำนวณโมเมนต์โดยรวมแล้วจะเป็นการคำนวณหาปริพันธ์ 1 มิติ ที่ขึ้นกับฟังก์ชัน $g_i(x_i)$ การลดความซับซ้อนในการคำนวณลักษณะนี้สามารถใช้ได้ เช่น เมื่อต้องการประมาณค่าคาดหวังแบบมีเงื่อนไขของตัวแปรสุ่ม X_i เมื่อค่าคาดหวังแบบมีเงื่อนไขของตัวแปรสุ่ม X_i มีค่าเท่ากับค่าคาดหวังแบบมีเงื่อนไขของ $\mu_i + \sqrt{\sigma_i^2 - \rho^2} Z_i$

2.3 รายละเอียดเพิ่มเติมสำหรับการคำนวณภายใต้เงื่อนไขลักษณะอื่นๆ

นอกเหนือจากการใช้ลำดับอย่างง่ายตามสมการที่ (1) แล้ว เซต S ยังสามารถใช้แสดงเงื่อนไขของข้อมูลประเภทอื่นได้อีกด้วย เช่น กรณีของ Umbrella ordering

$$X_1 + c_1 \le X_2 + c_2 \le \cdots \le X_u + c_u \ge X_{n-1} + c_{n-1} \ge X_n + c_n \qquad (5)$$

ตัวอย่างเช่น สำหรับค่าคงที่ c_i ใดๆ ซึ่งลำดับลักษณะนี้ได้รับความสนใจเป็นอย่างมากในงานวิจัยเชิงสถิติ (ศึกษาเพิ่มเติมได้ในงานของ Hans and Dunson(2005), Singh and Liu (2006), Nakas and Alonzo (2007) และ Gaur et al. (2012)) นอกจากนี้ยังสามารถประยุกต์ใช้เทคนิคการคำนวณที่นำเสนอในงานวิจัยนี้กับข้อมูลลำดับแบบ tree structure ตามรายละเอียดที่ระบุไว้ในงานส่วนที่ 4 ของ Hayter (2006)

การลดความซับซ้อนในคำนวณค่าตามสมการที่ (2) และ (3) โดยการใช้วิธีปริพันธ์เวียนเกิดในการคำนวณด้วยวิธีการหาปริพันธ์ใน 1 มิตินั้น ทำได้เนื่องจากเงื่อนไข 2 ประการ ประการแรก คือ เซต S เป็นการสร้างข้อจำกัดเฉพาะตัวแปร X_i ที่อยู่ในลำดับติดกัน (แม้ว่าการกำหนดชื่อใดๆ ให้ตัวแปร n ตัวจะสามารถทำได้) เงื่อนไขประการที่ 2 คือ ปริพัทธ์สามารถแยกออกเป็นเทอมต่างหากสำหรับแต่ละตัวแปร เงื่อนไขสองประการนี้ จะเป็นจริงสำหรับข้อมูลอันดับอย่างง่ายตามสมการที่ (1) และอันดับแบบ Umbrella ordering ตามสมการที่ (5) และสำหรับตัวแปรสุ่ม X_i ที่เป็นอิสระกัน และค่าคาดหวังได้จากฟังก์ชัน $\mathrm{g}_i(x_i)$ ในความเป็นจริงแล้วภายใต้เงื่อนไขตามสมการที่ (1) เทอม B จะเป็นเพียงค่าความน่าจะเป็นของลำดับอย่างง่ายนี้ โดย Hayter and Liu (1996) ถือเป็นงานศึกษาแรกที่นำเสนอวิธีการประมาณค่าโดยใช้เทคนิคปริพันธ์เวียนเกิดมาคำนวณหาปริพันธ์ใน 1 มิติ สำหรับตัวแปรสุ่มที่เป็นอิสระกัน

ถ้าตัวแปรสุ่ม X_i ไม่เป็นอิสระกัน หรือข้อมูลที่นำมาคำนวณเป็นการกำหนดเงื่อนไขของตัวแปรสุ่มที่ไม่อยู่ติดกัน (non-adjacent) แล้ว A_1, A_2 และ B ไม่จำเป็นต้องประมาณค่าโดยใช้การหาปริพันธ์ใน 1 มิติ อย่างไรก็ตาม เทคนิคปริพันธ์เวียนเกิดสำหรับมิติที่สูงขึ้นจะทำการประมาณค่าได้โดยการคำนวณหาปริพันธ์ใน r มิติ (เมื่อ $r \geq 2$) โดยการคำนวณจะเป็นไปได้หรือไม่ขึ้นอยู่กับรูปแบบของเทอม A_1, A_2 และ B

สำหรับเซต $T \subseteq \Re^n$ ของค่า $\boldsymbol{X} = (X_1, \ldots, X_n)$ กำหนดให้เป็น

$$T = T_{1,2} \cap T_{2,3} \cap \ldots \cap T_{n-1,n}$$

เมื่อเซต $T_{i,i+1}$ เป็นข้อจำกัดสำหรับค่า X_i และ X_{i+1} เท่านั้น จะได้ว่าความน่าจะเป็นแบบมีเงื่อนไข $P(\boldsymbol{X} \in T | \boldsymbol{X} \in \boldsymbol{S})$ เท่ากับ A_1/B โดย $S^*_{i,i+1} = S_{i,i+1} \cap T_{i,i+1}$ ดังนั้นค่าความน่าจะเป็นแบบมีเงื่อนไขจะสามารถประมาณค่าได้โดยการหาปริพันธ์ใน 1 มิติ โดยใช้เทคนิคปริพันธ์เวียนเกิด

เทคนิคการคำนวณที่นำเสนอในงานวิจัยนี้จะมีประโยชน์มากหากตัวแปรสุ่ม X_i มีการแจกแจงต่างกัน เนื่องจากหาก X_i แต่ละตัวมีการแจกแจงเหมือนกันและเป็นอิสระกันแล้ว ในการคำนวณการแจกแจงและโมเมนต์แบบมีเงื่อนไขของข้อมูลตามในสมการที่ (1) จะเปรียบเสมือนการใช้ข้อมูล X_i ของข้อมูลลำดับที่ i^{th} ตามหลักของสถิติเชิงอันดับ (order statistics) ซึ่งไม่เกี่ยวข้องกับลำดับที่ถูกต้องของตัวแปร $i-1$ ที่มีค่าน้อยกว่า X_i และตัวแปร $n-i$ ที่มีค่ามากกว่า X_i ในกรณีนี้สามารถใช้วิธีตามงานศึกษาเกี่ยวกับสถิติเชิงอันดับ (เช่น Arnold et al. (1992), Harter and Balakrishnan (1996) และ David and Nagaraja (2003)) ในการหาข้อมูลเงื่อนไขของ X_i ได้ แต่สำหรับตัวแปรสุ่ม X_i แต่ละตัวที่มีการแจกแจงต่างกันแล้ว ข้อมูลอันดับตามสมการที่ (1) จะให้ข้อมูลมากกว่าการใช้ข้อมูล X_i ของข้อมูลลำดับที่ i^{th} ตามหลักของสถิติเชิงอันดับ และวิธีที่นำเสนอในงานวิจัยนี้ทำให้การปรับการแจกแจงก่อนด้วยข้อมูลลักษณะนี้ทำได้อย่างมีประสิทธิภาพ

อาจมีบางกรณีที่ข้อมูลอันดับที่ได้ไม่ถูกต้อง เนื่องจากความผิดพลาดในขั้นตอนการสร้างข้อมูล หรืออาจเกิดจากความไม่แน่นอน อย่างเช่นในส่วนที่ 6 ในงานศึกษาของ Chiarawongse et al. (2012) ที่ชี้ให้เห็นถึงความไม่แน่นอนที่เกิดขึ้นได้ในทางปฏิบัติเมื่อนักลงทุนให้ข้อคิดเห็นที่เป็นข้อมูลอันดับ แต่มีความไม่แน่นอนในแง่ของความถูกต้องของข้อมูลอันดับที่ได้ ในกรณีนี้หากสามารถระบุระดับความเชื่อมั่นที่มีต่อข้อมูลอันดับนั้นๆได้ ก็สามารถใช้ค่าความน่าจะเป็นที่ความคิดเห็นที่ได้จะถูกต้องมาประกอบการคำนวณ ซึ่งในงานของ Chiarawongse et al. (2012) จะเรียกว่าเป็นข้อมูลอันดับที่ไม่สมบูรณ์ (imperfect ranking) โดยนำค่านี้ไปคำนวณได้ตาม shrinkage model ต่อไปนี้

$$\kappa P(\boldsymbol{X} \in C \mid \boldsymbol{X} \in S) + (1-\kappa)P(\boldsymbol{X} \in C) \tag{6}$$

เมื่อ C เป็นเหตุการณ์ใดๆ และ S เป็นข้อมูลอันดับที่มี (observed ranking) สมการที่ (6) แสดงให้เห็นว่าค่าที่ได้เป็นการเฉลี่ยถ่วงน้ำหนักระหว่างความน่าจะเป็นของข้อมูลอันดับที่มีกับความน่าจะเป็นของการแจกแจงก่อน โดยพารามิเตอร์ κ แทนความน่าจะเป็นที่ข้อมูลอันดับที่ได้จะถูกต้อง

และจะได้ว่าค่าคาดหวังแบบมีเงื่อนไขก็จะคำนวณได้จาก convex combination ระหว่างค่าคาดหวังภายใต้เงื่อนไขของข้อมูลอันอับที่มี กับค่าคาดหวังกรณีที่ไม่ใช้ข้อมูลอันดับมาพิจารณา ดังแสดงได้ตามสมการ

$$\kappa E[\mathrm{g}(\boldsymbol{X})|\boldsymbol{X} \in S] + (1-\kappa)E[\mathrm{g}(\boldsymbol{X})] \tag{7}$$

นอกจากนี้ จะเห็นได้ว่าหากสามารถระบุค่าของความเชื่อมั่นตามสมการที่ (6) และ (7) ได้แล้ว การคำนวณหาค่าคาดหวังจะสามารถทำได้โดยง่าย ดังนั้น การคำนวณในส่วนต่อไปจะพิจารณาการคำนวณการแจกแจงและโมเมนต์แบบมีเงื่อนไขสำหรับข้อมูลอันดับที่มี (observed ranking) เท่านั้น โดยหากต้องการทราบค่ากรณีที่ข้อมูลอันดับไม่สมบูรณ์ (imperfect ranking) ก็สามารถคำนวณต่อได้ตาม shrinkage model ตามที่อธิบายไว้แล้วนี้

3. วิธีการคำนวณ

เนื้อหาในส่วนนี้จะอธิบายวิธีการนำเทคนิคปริพันธ์เวียนเกิดไปใช้ เริ่มจากการแสดงสูตรการคำนวณ จากนั้นจึงอธิบายขั้นตอนการคำนวณอย่างละเอียด ก่อนที่จะแสดงให้เห็นถึงกลไกการลดความคลาดเคลื่อน (self-correction) รวมทั้งเปรียบเทียบเวลาที่ใช้ในการคำนวณด้วย

3.1 สูตรในการคำนวณสำหรับวิธีปริพันธ์เวียนเกิด

การประมาณค่าสมการที่ (2) สำหรับตัวแปรสุ่มที่เป็นอิสระกันตามที่ระบุไว้ในหัวข้อ 2.1 เทอม B และ A_1 จะสามารถประมาณค่าได้ด้วยเทคนิคปริพันธ์เวียนเกิด ในการหาค่า B ฟังก์ชัน $b_1, \ldots, b_{n-1}$ สามารถหาค่าได้ตามลำดับ โดยแต่ละขั้นจะเป็นการหาปริพันธ์ใน 1 มิติ ให้ $b_0(z) = 1$ และสำหรับ $i = 1, \ldots, n-1$ จะประมาณค่าแต่ละ $z \in \Re$

$$b_i(z) = \int_{-\infty}^{z} b_{i-1}(x) f_i(x)\, dx \tag{8}$$

โดย f_i เป็นฟังก์ชันความหนาแน่นของ X_i. ดังนั้น

$$B = \int_{-\infty}^{\infty} b_{n-1}(z) f_n(z)\, dz \tag{9}$$

ในทำนองเดียวกัน การประมาณค่าเทอม A_1 หาจากฟังก์ชัน $a_1, \dots, a_{n-1}$ โดยการคำนวณตามลำดับ กำหนดให้ $a_0(z) = 1$ หาจากฟังก์ชัน $i = 1, \dots, n-1$, จะประมาณค่าแต่ละ $z \in \Re$

$$a_i(z) = \int_{l_i}^{\max\{\min\{z,\, u_i\}, l_i\}} a_{i-1}(x) f_i(x)\, dx \tag{10}$$

ดังนั้น

$$A_1 = \int_{l_n}^{u_n} a_{n-1}(z) f_n(z)\, dz \tag{11}$$

การประมาณค่าสมการที่ (3) สำหรับตัวแปรสุ่มที่เป็นอิสระกันตามที่ระบุไว้ในหัวข้อ 2.1 เทอม B สามารถประมาณค่าได้ด้วยวิธีที่อธิบายไว้ข้างต้น ส่วนเทอม A_2 จะกำหนดฟังก์ชัน $h_1, \dots, h_{n-1}$ โดยการคำนวณตามลำดับด้วยวิธีการหาปริพันธ์เวียนเกิด และกำหนดให้ $h_0(z) = 1$ และสำหรับ $i = 1, \dots, n-1$ จะประมาณค่าแต่ละ $z \in \Re$

$$h_i(z) = \int_{-\infty}^{z} h_{i-1}(x) \mathrm{g}_i(x) f_i(x)\, dx \tag{12}$$

จะได้

$$A_2 = \int_{-\infty}^{\infty} h_{\mathrm{n\text{-}1}}(\mathrm{z}) \mathrm{g}_{\mathrm{n}}(\mathrm{z}) f_{\mathrm{n}}(\mathrm{z})\, \mathrm{dz} \tag{13}$$

การประมาณค่าสมการที่ (2) สำหรับตัวแปรสุ่มที่มีการแจกแจงแบบปกติหลายตัวแปรตามที่ระบุไว้ในหัวข้อ 2.2 เทอม B และ A_1 จะสามารถประมาณค่าได้ด้วยเทคนิคปริพันธ์เวียนเกิด ในการหาค่า B ฟังก์ชัน $b_1, \dots, b_{n-1}$ สามารถหาค่าได้ตามลำดับ โดยแต่ละขั้นจะเป็นการหาปริพันธ์ใน 2 มิติ ให้ $b_0(z) = 1$ และสำหรับ $i = 1, \dots, n-1$ จะประมาณค่าแต่ละ $z \in \Re$

$$b_i(m, z) = \int_{-\infty}^{z} b_{i-1}(m, x) \phi_i(m, x)\, dx \tag{14}$$

เมื่อ ϕ_i เป็นฟังก์ชันการแจกแจงแบบ $N(\mu_i + \rho_i m,\ \sigma_i^2 - \rho_i^2)$ ของตัวแปรสุ่ม จะได้ว่าหาก ϕ เป็นฟังก์ชันความหนาแน่นของการแจกแจงปกติมาตรฐานแล้ว

$$B = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \emptyset(m) b_{n-1}(m, z) \phi_n(m, x)\, dz\, dm \tag{15}$$

ในทำนองเดียวกัน การประมาณค่าเทอม A_1 หาจากฟังก์ชัน $a_1, \dots, a_{n-1}$ กำหนดให้ $a_0(m, z) = 1$ และสำหรับ $i = 1, \dots, n-1$ จะประมาณค่าแต่ละ $z \in \Re$

$$a_i(m, z) = \int_{l_i}^{\max\{\min\{z, u_i\}, l_i\}} a_{i-1}(m, x) \phi_i(m, x)\, dx \tag{16}$$

จะได้

$$A_1 = \int_{m=-\infty}^{\infty} \int_{z=l_n}^{u_n} \phi(m) a_{n-1}(m,z) \phi_n(m,z)\, dz\, dm \quad (17)$$

การประมาณค่าสมการที่ (3) สูตรในการหาค่า A_2 เมื่อกำหนดให้ $h_1, \dots, h_{n-1}$ โดย $h_0(m,z) = 1$ และสำหรับ $i = 1, \dots, n-1$

$$h_i(m,z) = \int_{-\infty}^{z} h_{i-1}(m,x) \mathrm{g}_i(x) \phi_i(m,x)\, dx \quad (18)$$

เมื่อ $$A_2 = \int_{m=-\infty}^{\infty} \int_{z=-\infty}^{\infty} \phi(m) h_{n-1}(m,z) \mathrm{g}_n(z) \phi_n(m,z)\, dz\, dm \quad (19)$$

3.2 ขั้นตอนการคำนวณ

ในการประมาณค่าสมการที่ (9), (11) และ (13) ทำได้โดยการหาปริพันธ์ใน 1 มิติ ตามลำดับ โดยในการหาค่าปริพันธ์ในแต่ละครั้งหาจากผลรวมปริพันธ์หนึ่งมิติบนเส้นจำนวนจริงที่ถูกตัดปลายและแบ่งช่วง กระบวนการที่ 1 จะใช้สำหรับการคำนวณค่าตามสมการที่ (13) โดยใช้สูตรของ Newton-Cotes (the trapezoidal rule) ส่วนการคำนวณสมการที่ (9) และ (11) จะทำโดยการแทนค่าฟังก์ชัน $\mathrm{g}_i, i = 1, \dots, n$ ด้วยฟังก์ชันบ่งชี้ที่เหมาะสม หรืออาจแทนด้วยฟังก์ชันคงที่ที่เท่ากับ 1

ขั้นตอนการคำนวณแบบที่ 1: การคำนวณค่า A_2 ในสมการที่ (13)

1: Assume n variables with ranking $X_1 \le X_2 \dots \le X_n$.

2: Discretization grid size Δ with lower bound x_0, forming $N + 1$ grid points

$$\{x_0, x_1, \dots, x_N\}$$

where $x_j = x_{j-1} + \Delta$ for $j = 1, \dots, N$.

3: Let $h_0(x_j) = 1$ for $j = 0, 1, \dots, N$.

4: **for** $i = 1$ to n **do**

5: Let, for $j = 1, \dots, N$,

$$\bar{h}_j = \frac{h_{i-1}(x_{j-1}) g_i(x_{j-1}) f_i(x_{j-1}) + h_{i-1}(x_j) g_i(x_j) f_i(x_j)}{2}$$

6: Let, for $j = 1, \dots, N$,

$$h_i(x_j) = \sum_{k=1}^{j} \bar{h}_k \Delta$$

and let $h_i(x_0) = h_i(x_1)$

7: **end for**

8: **return** $A_2 = h_n(x_N)$.

การประมาณค่าสมการที่ (15), (17) และ (19) ทำได้โดยการหาปริพันธ์ 2 มิติ กระบวนการที่ 2 จะใช้สำหรับการคำนวณค่าตามสมการที่ (19) โดยใช้สูตรของ Newton-Cotes ส่วนการคำนวณสมการที่ (15) และ (17) จะทำโดยการแทนค่าฟังก์ชัน $g_i, i = 1, \dots, n$ ด้วยฟังก์ชันบ่งชี้ที่เหมาะสม หรืออาจแทนด้วยฟังก์ชันคงที่ที่เท่ากับ 1

ขั้นตอนการคำนวณแบบที่ 2: การคำนวณค่า A_2 ในสมการที่ (19)

1: Assume n variables with ranking $X_1 \leq X_2 \dots \leq X_n$.

2: Discretization grid size δ with lower bound m_0, forming $N + 1$ grid points

$$\{m_0, m_1, \dots, m_M\}$$

where $m_l = m_{l-1} + \delta$ for $l = 1, \dots, M$.

3: **for** $l = 1$ to M **do**

4: Discretization grid size Δ with lower bound x_0, forming $N + 1$ grid points

$$\{x_0, x_1, \dots, x_N\}$$

where $x_j = x_{j-1} + \Delta$ for $j = 1, \dots, N$.

5: Let $h_0(x_j) = 1$ for $j = 0, 1, \dots, N$.

6: **for** $i = 1$ to n **do**

7: Let, for $j = 1, \dots, N$,

$$\bar{h}_j = \frac{h_{i-1}(m_l, x_{j-1}) g_i(x_{j-1}) \phi_i(m_l, x_{j-1}) + h_{i-1}(m_l, x_j) g_i(x_j) \phi_i(m_l, x_j)}{2}$$

8: Let, for $j = 1, \dots, N$,

$$h_i(m_l, x_j) = \sum_{k=1}^{j} \bar{h}_k \Delta$$

and let $h_i(m_l, x_0) = h_i(m_l, x_1)$

9: **end for**

10: $\bar{h}(m_l) = h_n(m_l, x_N)$

11: **end for**

12: Let, for $l = 1, \dots, M$.

$$\bar{h}_l = \frac{\phi(m_l)\bar{h}(m_l) + \phi(m_{l-1})\bar{h}(m_{l-1})}{2}$$

13: **return** $A_2 = \sum_{l=1}^{M} \bar{h}_l \Delta$.

ตารางที่ 1 ค่าคลาดเคลื่อนและเวลาที่ใช้ในการคำนวณฟังก์ชันการแจกแจงสะสม โดยประเมิน 3 จุด สำหรับตำแหน่งที่ 70 ของข้อมูลสถิติอันดับ จากจำนวนตัวแปร $n = 101$ และตัวแปรสุ่มแต่ละตัวมีการแจกแจงแบบ $U[0,1]$ และเป็นอิสระกัน

ขนาดกริด	ค่าจริง	ค่าที่คำนวณได้	ค่าคลาดเคลื่อน	เวลาที่ใช้คำนวณ (วินาที)
0.01	0.03382186	0.09552326	6.170e-02	0.00
	0.60791267	0.65772572	4.981e-02	0.00
	0.99628437	0.99005003	6.234e-03	0.00
0.001	0.03382186	0.03422536	4.035e-04	0.00
	0.60791267	0.60800312	9.045e-05	0.02
	0.99628437	0.99620053	8.385e-05	0.00
0.0001	0.03382186	0.03382585	3.987e-06	0.06
	0.60791267	0.60791339	7.206e-07	0.07
	0.99628437	0.99628353	8.411e-07	0.08
0.00001	0.03382186	0.03382190	3.982e-08	1.36
	0.60791267	0.60791268	7.037e-09	1.20
	0.99628437	0.99628437	8.414e-09	1.42

ตารางที่ 2 ค่าคลาดเคลื่อนและเวลาที่ใช้ในการคำนวณฟังก์ชันการแจกแจงสะสม โดยประเมิน 3 จุด สำหรับตำแหน่งที่ 70 ของข้อมูลสถิติอันดับ จากจำนวนตัวแปร $n = 101$ และตัวแปรสุ่มแต่ละตัวมีการแจกแจงแบบปกติมาตรฐานและเป็นอิสระกัน

ขนาดกริด	ค่าจริง	ค่าที่คำนวณได้	ค่าคลาดเคลื่อน	เวลาที่ใช้คำนวณ (วินาที)
0.01	0.03382186	0.03766693	3.845e-03	0.05
	0.60791267	0.61814333	1.023e-02	0.05
	0.99628437	0.99663459	3.502e-04	0.03
0.001	0.03382186	0.03367669	1.452e-04	0.39
	0.60791267	0.60697199	9.407e-04	0.36
	0.99628437	0.99632079	3.642e-05	0.39
0.0001	0.03382186	0.03379485	2.701e-05	4.38
	0.60791267	0.60791453	1.860e-06	4.36
	0.99628437	0.99628272	1.657e-06	4.39
0.00001	0.03382186	0.03382366	1.796e-06	38.64
	0.60791267	0.60791215	5.155e-07	36.95
	0.99628437	0.99628432	5.078e-08	38.14

งานวิจัยนี้แสดงตัวอย่างการคำนวณสำหรับตัวแปรที่มีลักษณะการแจกแจงแตกต่างกัน 2 กรณี คือ กรณีที่ตัวแปรสุ่มมีการแจกแจงเหมือนกันและเป็นอิสระกันซึ่งจะทราบคำตอบของค่าที่คำนวณได้ ในที่นี้จะพิจารณาตัวแปรสุ่มจำนวน 101 ตัวแปร ซึ่งแต่ละตัวแปรกำหนดให้มีการแจกแจงแบบ $U[0,1]$ กับกรณีที่ตัวแปรสุ่มมีการแจกแจงแบบปกติมาตรฐานที่เป็นอิสระกัน ในทั้ง 2 กรณีจะประมาณค่าความน่าจะเป็นสะสมโดยใช้วิธีปริพันธ์เวียนเกิดสำหรับ 3 จุด ของตัวแปรในตำแหน่งที่ 70 หรือ X_{70} ของข้อมูลสถิติอันดับ ภายใต้เงื่อนไขที่ $X_1 \leq \ldots \leq X_{101}$ โดยคำนวณเมื่อแบ่งขนาดกริดต่างๆ กัน โดยค่าที่คำนวณได้ ค่าความคลาดเคลื่อน และเวลาที่ใช้ในการคำนวณแสดงไว้ในตารางที่ 1 และ 2 (ค่าจริงคำนวณจากความน่าจะเป็นสะสมจากการแจงแจงทวินาม) การคำนวณในงานวิจัยนี้ใช้โปรแกรม R และใช้เครื่อง 64-bit Windows ที่มี Intel Core i5-2500 3.30 GHz CPU

ในลำดับถัดไปจะแสดงการใช้เทคนิคปริพันธ์เวียนเกิดในการคำนวณค่าคาดหวังของตัวแปรในตำแหน่งที่ X_{17} และ X_{51} ภายใต้เงื่อนไขที่ $X_1 \leq \ldots \leq X_{101}$ โดยคำนวณเมื่อแบ่งขนาดกริดต่างๆ กัน โดยค่าที่คำนวณได้ ค่าความคลาดเคลื่อน และเวลาที่ใช้ในการคำนวณแสดงไว้ในตารางที่ 3 สำหรับกรณีที่ตัวแปรสุ่มมีการแจกแจงแบบ $U[0,1]$ และเป็นอิสระกัน ส่วนกรณีที่ตัวแปรสุ่มมีการแจกแจงแบบปกติมาตรฐานที่เป็นอิสระกันจะแสดงผลการคำนวณไว้ในตารางที่ 4 สำหรับกรณีแรก ทราบแล้วว่าค่าจริงของ $E[X_{17}]$ และ $E[X_{51}]$ เป็น 1/6 และ 0.5 ตามลำดับ ส่วนในกรณีที่ 2 ค่าจริงของ $E[X_{17}]$ จะไม่ทราบค่าที่แน่นอน แต่จะมีค่าประมาณ $\Phi^{-1}(1/6) \approx 0.97$ และค่าจริงของ $E[X_{51}]$ เป็นศูนย์

ตารางที่ 3 ค่าคลาดเคลื่อนและเวลาที่ใช้ในการคำนวณค่าคาดหวังของตัวแปร $X_{(17)}$ และ $X_{(51)}$ จากจำนวนตัวแปร $n = 101$ และตัวแปรสุ่มแต่ละตัวมีการแจกแจงแบบ $U[0,1]$ และเป็นอิสระกัน

ขนาดกริด	ค่าจริง	ค่าที่คำนวณได้	ค่าคลาดเคลื่อน	เวลาที่ใช้คำนวณ (วินาที)
0.01	1/6	0.14369985	2.297e-02	0.00
	0.5	0.48606226	1.394e-02	0.00
0.001	1/6	0.16661772	4.894e-05	0.00
	0.5	0.49997063	2.937e-05	0.00
0.0001	1/6	0.16666624	4.272e-07	0.05
	0.5	0.49999974	2.563e-07	0.06
0.00001	1/6	0.16666666	4.215e-09	0.92
	0.5	0.50000000	2.529e-09	0.89

ตารางที่ 4 ค่าคลาดเคลื่อนและเวลาที่ใช้ในการคำนวณค่าคาดหวังของตัวแปร $X_{(17)}$ และ $X_{(51)}$ จากจำนวนตัวแปร $n = 101$ และตัวแปรสุ่มแต่ละตัวมีการแจกแจงแบบปกติมาตรฐานและเป็นอิสระกัน

ขนาดกริด	ค่าจริง	ค่าที่คำนวณได้	ค่าคลาดเคลื่อน	เวลาที่ใช้คำนวณ (วินาที)
0.01	NA	-0.9672	NA	0.04
	0.0000	0.0000	1.397e-14	0.05
0.001	NA	-0.9779	NA	0.38
	0.0000	0.0000	1.734e-16	0.39
0.0001	NA	-0.978	NA	4.33
	0.0000	0.0000	9.826e-18	4.25
0.00001	NA	-0.978	NA	38.55
	0.0000	0.0000	1.281e-17	37.22

ในตารางที่ 1-4 นี้คอลัมน์ที่ 2 แสดงค่าจริงที่ทราบค่าจากการคำนวณที่อธิบายไปข้างต้น ส่วนคอลัมน์ที่ 3 แสดงค่าที่คำนวณได้จากวิธีการที่เสนอในงานวิจัยนี้ ในขณะที่คอลัมน์ที่ 4 และ 5 แสดงค่าคลาดเคลื่อนของค่าที่คำนวณได้แตกต่างจากค่าจริง และแสดงเวลาที่ใช้ในการคำนวณสำหรับแต่ละกรณีตามลำดับ

จากผลการคำนวณที่ได้ ค่าความคลาดเคลื่อนมีค่าน้อย และเวลาที่ใช้ในการคำนวณก็ไม่มากนัก ซึ่งหากเขียนโปรแกรมด้วยภาษา C++ หรือ Java คาดว่าจะลดเวลาที่ใช้ในการคำนวณลงได้อีก

4. ตัวอย่างการประยุกต์

หัวข้อนี้จะแสดงตัวอย่างการประยุกต์ใช้วิธีการคำนวณที่นำเสนอมาข้างต้นกับปัญหาที่แตกต่างกัน 3 ด้าน ตัวอย่างแรกเป็นปัญหาด้านสุขภาพ โดยตัวแปรสุ่มในปัญหานี้มีการแจกแจงแบบแกมมาที่เป็นอิสระกัน โดยให้ค่าพารามิเตอร์ที่แสดงรูปร่างมีค่าเท่ากัน แต่ให้สเกลพารามิเตอร์แตกต่างกัน ตัวอย่างที่สองเป็นการศึกษาปริมาณสารปนเปื้อนในดินของ การแจกแจงของตัวแปรสุ่มเป็นแบบปกติหลายตัวแปรที่มีค่าเฉลี่ยต่างกัน แต่มีค่าความแปรปรวนและความแปรปรวนร่วมเท่ากัน โดยในตัวอย่างนี้จะแสดงให้เห็นผลเมื่อข้อมูลอันดับที่ใช้ปรับค่าสอดคล้อง (reinforcing rankings) และตรงข้าม (opposing rankings) กับค่าเฉลี่ยของการแจกแจงก่อน โดยจะผลของข้อมูลอันดับแต่ละประเภท ต่อค่าของการแจกแจง ค่าคาดหวัง ความแปรปรวน และความแปรปรวนร่วมของตัวแปร ตัวอย่างที่สามเป็นการประยุกต์ใช้ด้านการเงิน โดยจะศึกษาปัญหาการจัดพอร์ตการลงทุน

4.1 ด้านสุขภาพ

ศึกษาตัวแปรสุ่ม X_i ที่แทนอาการ (Medical Condition) ของคนไข้ 5 คน ซึ่งอาการดังกล่าวไม่สามารถวัดค่าได้โดยตรง อย่างไรก็ตาม พอจะสามารถระบุตัวแบบการแจกแจงของระดับความรุนแรงของอาการด้วยการแจกแจงแบบแกมมา ได้ดังนี้

$$f(x{:}\,3,\theta) = \frac{1}{2\theta^3}x^2e^{-x/\theta}$$

โดยกำหนดให้พารามิเตอร์ที่กำหนดรูปร่างของการแจกแจงมีค่าเป็น $k = 3$ และพารามิเตอร์กำหนดสเกลแทนด้วยสัญลักษณ์ θ ค่าของพารามิเตอร์นี้จะขึ้นอยู่กับค่าอื่นๆ ที่เกี่ยวข้องของคนไข้แต่ละคน ในที่นี้จะให้คนไข้ 5 คน มีสเกลพารามิเตอร์เป็น $\theta_i = i, 1 \le i \le 5$ จากเงื่อนไขที่กำหนดนี้ จะได้ว่า การแจกแจงก่อนมีค่าคาดหวังเป็น $E(X_i) = 3$ และมีค่าความแปรปรวนเป็น $\mathrm{Var}(X_i) = 3i^2, 1 \le i \le 5$ นอกจากนี้ยังกำหนดให้ตัวแปรสุ่ม X_i แต่ละตัวเป็นอิสระกันด้วย

จากนั้นสมมติว่าเราสามารถจัดอันดับอาการของคนไข้ตามระดับอาการโดยใช้ค่าอื่นๆ ที่สังเกตได้และเกี่ยวข้องกับสภาพทางการแพทย์ที่สนใจมาเป็นข้อมูลประกอบการจัดอันดับ จนได้เป็นข้อมูลอันดับดังนี้

$$X_1 \le X_2 \le X_3 \le X_4 \le X_5$$

ข้อมูลอันดับนี้สอดคล้องกับอันดับของค่าเฉลี่ยของการแจกแจงก่อน จึงเรียกข้อมูลอันดับนี้ว่า reinforcing rankings โดยภายใต้การแจกแจงก่อนที่กำหนด ข้อมูลอันดับนี้จะมีความน่าจะเป็นเท่ากับ 0.107 (ซึ่งก็คือค่า B ในสมการที่ (2) และ (3)) จากนั้นจะใช้เทคนิคปริพันธ์เวียนเกิด ตามที่อธิบายไว้ในส่วนที่ 2 ของงานวิจัยในการคำนวณหาค่าคาดหวังแบบมีเงื่อนไข ส่วนเบี่ยงเบนมาตรฐาน และค่าสหสัมพันธ์ของตัวแปรสุ่ม X_i (เมื่อมีเงื่อนไขของข้อมูลอันดับทำให้ตัวแปรสุ่ม X_i ไม่ได้เป็นอิสระกันแล้ว) โดยแสดงผลการคำนวณไว้ในตารางที่ 5

จากผลที่ได้จะเห็นว่าภายใต้เงื่อนไขของข้อมูลอันดับที่สอดคล้องกับอันอับของค่าเฉลี่ยของการแจกแจงก่อน ค่าคาดหวังของตัวแปรสุ่ม X_i ที่ได้ ยังคงเรียงลำดับเหมือนเดิม แต่ค่าที่ได้มีการกระจายมากขึ้นกว่าค่าคาดหวังของการแจกแจงก่อน นอกจากนี้ส่วนเบี่ยงเบนมาตรฐานแบบมีเงื่อนไขจะมีค่าน้อยกว่าส่วนเบี่ยงเบนมาตรฐานของการแจกแจงก่อน และเมื่อพิจารณาค่าสหสัมพันธ์ จะเห็นว่าค่าสหสัมพันธ์ของตัวแปรสุ่มที่อยู่ติดกันจะมีค่าสูง

เมื่อคำนวณค่าความน่าจะเป็นที่คนไข้จะมีระดับอาการต่ำกว่า 5 ซึ่งจะถือว่าเป็นระดับที่ต้องได้รับการรักษาอย่างเร่งด่วน ค่าในตารางที่ 5 แสดงให้เห็นว่าเมื่อปรับการแจกแจงด้วยข้อมูลอันดับที่สอดคล้องกับการแจกแจงก่อนแล้ว จะทำให้ความน่าจะเป็นที่ตัวแปรสุ่ม X_i จะมีค่าน้อยกว่าหรือเท่ากับ 5 มีค่าเพิ่มขึ้นสำหรับคนไข้รายที่ 1 และ 2 และมีค่าลดลงสำหรับคนไข้รายที่ 3 4 และ 5

ตารางที่ 5 ตัวอย่างที่ 1 ระดับสภาพทางการแพทย์ (Medical Condition) ของคนไข้ 5 ราย

คนไข้	ค่าคาดหวัง	ส่วนเบี่ยงเบนมาตรฐาน	เมตริกซ์สัมประสิทธิ์สหสัมพันธ์					$P(X_i \leq 5)$
1	3	1.73	1	0	0	0	0	0.876
2	6	3.46		1	0	0	0	0.456
3	9	5.20			1	0	0	0.234
4	12	6.93				1	0	0.132
5	15	8.66					1	0.080
Reinforcing ranking	$X_1 \leq X_2 \leq X_3 \leq X_4 \leq X_5$							
1	2.32	1.22	1	0.36	0.17	0.08	0.03	0.968
2	4.89	2.03		1	0.49	0.24	0.10	0.578
3	8.19	3.05			1	0.52	0.23	0.132
4	12.86	4.63				1	0.45	0.010
5	21.41	8.34					1	0.000
Opposing ranking	$X_5 \leq X_4 \leq X_3 \leq X_2 \leq X_1$							
1	7.99	2.18	1	0.82	0.69	0.57	0.40	0.064
2	6.68	1.86		1	0.85	0.70	0.49	0.184
3	5.64	1.69			1	0.82	0.58	0.382
4	4.58	1.56				1	0.71	0.641
5	3.24	1.45					1	0.881

จากนั้นคำนวณค่าต่างๆ ในกรณีที่ข้อมูลอันดับตรงข้ามกับอันดับของค่าเฉลี่ยของการแจกแจงก่อน คือ

$$X_5 \leq X_4 \leq X_3 \leq X_2 \leq X_1$$

ข้อมูลอันดับที่ขัดแย้งกับอันดับของค่าเฉลี่ยของการแจกแจงก่อนนี้มีความน่าจะเป็นน้อยมากเพียง 0.00003 ซึ่งค่าคาดหวังแบบมีเงื่อนไข ส่วนเบี่ยงเบนมาตรฐาน และค่าสหสัมพันธ์ของตัวแปรสุ่ม X_i ที่คำนวณได้แสดงตามตารางที่ 5 เช่นกัน

ผลที่ได้ชี้ให้เห็นว่า เมื่อข้อมูลอันดับที่ใช้ตรงข้ามกับอันดับของการแจกแจงก่อน ลำดับของค่าคาดหวังที่คำนวณได้จะถูกปรับใหม่ให้สอดคล้องกับข้อมูลอันดับนี้ โดยค่าคาดหวังที่คำนวณได้จะมีการกระจายลดลง ส่วนเบี่ยงเบนมาตรฐานภายใต้เงื่อนไขของข้อมูลอันดับที่ตรงข้ามจะมีค่าลงลงกว่าส่วนเบี่ยงเบนมาตรฐานของการแจกแจงก่อนมาก ค่าสหสัมพันธ์ของตัวแปรที่อยู่ติดกันยังคงมีค่าสูง และโดยรวมแล้ว ค่าสหสัมพันธ์จะมีค่าสูงกว่ากรณีที่ใช้ข้อมูลอันดับที่สอดคล้องกับอันดับของค่าเฉลี่ยของการแจกแจงก่อนมาก นอกจากนี้ค่าความน่าจะเป็นที่คนไข้จะมีระดับอาการต่ำกว่า 5 ก็จะเปลี่ยนแปลงตามข้อมูลอันดับที่นำมาพิจารณาด้วย

โดยสรุปแล้ว ตัวอย่างข้างต้นแสดงให้เห็นผลของการใช้ข้อมูลอันดับมาปรับการแจกแจงก่อน โดยข้อมูลอันดับที่มีรูปแบบแตกต่างกัน ก็จะมีผลต่อการแจกแจงและโมเมนต์แบบมีเงื่อนไขแตกต่างกันไป ข้อสังเกตต่างๆ ที่ได้จากการคำนวณนี้จะเป็นประโยชน์ต่อผู้ที่จะนำวิธีการคำนวณนี้ไปใช้ในทางปฏิบัติ

4.2 ด้านมลพิษ (พื้นที่อันตราย)

พิจารณาการประเมินปริมาณสารพิษปนเปื้อนในดิน โดยสมมติว่านักวิทยาศาสตร์กำหนดตัวแบบการแจกแจงของปริมาณสารพิษปนเปื้อนให้เป็นตัวแปรสุ่มที่มีการแจกแจงแบบปกติหลายตัวแปร และสนใจศึกษาพื้นที่อันตรายทั้งหมด $n = 6$ แห่ง เวกเตอร์ค่าเฉลี่ย μ มีค่าเป็น (10, 10, 12, 15, 18, 20) ส่วนเบี่ยงเบนมาตรฐานและค่าสหสัมพันธ์มีค่าเป็น 3 และ 0.4 ตามลำดับ เท่ากันสำหรับทุกตัวแปร จากการสังเกตลักษณะภายนอกของพื้นที่ทำให้สามารถจัดลำดับความอันตรายของพื้นที่ทั้ง 6 แห่งได้ ดังนี้

$$X_1 \leq X_2 \leq X_3 \leq X_4 \leq X_5 \leq X_6$$

จากการแจกแจงก่อนที่กำหนด ข้อมูลอันดับที่สอดคล้องกับอันดับของค่าเฉลี่ยของการแจกแจงก่อน จะมีความน่าจะเป็นเท่ากับ 0.111 โดยค่าคาดหวังแบบมีเงื่อนไข ส่วนเบี่ยงเบนมาตรฐาน และค่าสหสัมพันธ์ของระดับการปนเปื้อนของสารพิษในดินที่คำนวณได้แสดงไว้ในตารางที่ 6

สังเกตว่าค่าคาดหวังแบบมีเงื่อนไขเมื่อใช้ข้อมูลอันดับที่สอดคล้องกับการแจกแจงก่อนมีค่าใกล้เคียงกับค่าคาดหวังของการแจกแจงก่อน แม้ว่าค่าคาดหวังของพื้นที่ที่ 1 จะมีค่าลดลงจาก 10 เป็น 8.21 แต่ในตำแหน่งที่ 6 ค่าคาดหวังที่ปรับด้วยข้อมูลอันดับแล้วมีค่าเพิ่มขึ้นเล็กน้อย จาก 20 เป็น 20.94 ส่วนเบี่ยงเบนมาตรฐานของทุกพื้นที่มีค่าลดลงเล็กน้อย และแต่ละพื้นที่ได้ค่าใกล้เคียงกัน ในขณะที่ค่าสหสัมพันธ์จะมีค่าสูงที่สุดสำหรับตัวแปรที่อยู่ติดกัน

ถ้ากำหนดเกณฑ์ให้พื้นที่ที่มีระดับสารพิษปนเปื้อนในดินสูงกว่า 18 ถือเป็นพื้นที่ที่จำเป็นต้องดำเนินการกำจัดสารปนเปื้อน ค่าความน่าจะเป็นที่แต่ละพื้นที่จะมีระดับสารพิษปนเปื้อนมากกว่า 18 คำนวณและแสดงผลไว้ในตารางที่ 6 เห็นได้ว่าการปรับค่าการแจกแจงด้วยข้อมูลอันดับที่สอดคล้องกับการแจกแจงก่อนทำให้ในพื้นที่ที่ 6 มีความน่าจะเป็นที่จะกลายเป็นพื้นที่ที่จำเป็นต้องดำเนินการกำจัดสารปนเปื้อนเพิ่มขึ้นจาก 0.748 เป็น 0.865 ส่วนในพื้นที่อื่นๆ อีก 5 แห่ง ค่าความน่าจะเป็นที่คำนวณได้จะมีค่าลดลง

จากนั้นเปลี่ยนข้อมูลอันดับที่ใช้มาเป็นข้อมูลอันดับที่บางส่วนไม่สอดคล้องกับอันดับของค่าเฉลี่ยของการแจกแจงก่อน

$$X_1 \leq X_2 \leq X_5 \leq X_4 \leq X_3 \leq X_6$$

กล่าวคือ ให้อันดับของพื้นที่แห่งที่ 3 4 และ 5 ตรงข้ามกับอันดับของค่าเฉลี่ยในการแจกแจงก่อน โดยข้อมูลอันดับลักษณะนี้มีค่าความน่าจะเป็น 0.002

ตารางที่ 6 ตัวอย่างที่ 2 ระดับสารพิษปนเปื้อนในดินของพื้นที่ 6 แห่ง

ตำแหน่งที่	ค่าคาดหวัง	ส่วนเบี่ยงเบนมาตรฐาน	เมตริกซ์สัมประสิทธิ์สหสัมพันธ์						$P(X_i \geq 18)$
1	10	3	1	0.4	0.4	0.4	0.4	0.4	0.004
2	10	3		1	0.4	0.4	0.4	0.4	0.004
3	12	3			1	0.4	0.4	0.4	0.023
4	15	3				1	0.4	0.4	0.159
5	18	3					1	0.4	0.500
6	20	3						1	0.748
Reinforcing ranking	$X_1 \leq X_2 \leq X_3 \leq X_4 \leq X_5 \leq X_6$								
1	8.21	2.59	1	0.79	0.67	0.59	0.56	0.52	0.000
2	10.28	2.45		1	0.78	0.65	0.60	0.55	0.001
3	12.47	2.45			1	0.74	0.62	0.56	0.012
4	15.12	2.5				1	0.71	0.58	0.125
5	17.97	2.54					1	0.69	0.494
6	20.95	2.71						1	0.863
Partially Opposing ranking	$X_1 \leq X_2 \leq X_5 \leq X_4 \leq X_3 \leq X_6$								
1	8.52	2.64	1	0.76	0.60	0.61	0.62	0.48	0.000
2	10.88	2.52		1	0.65	0.68	0.69	0.51	0.002
3	16.13	2.36			1	0.92	0.86	0.60	0.214
4	15.08	2.31				1	0.92	0.59	0.104
5	14.05	2.34					1	0.58	0.046
6	20.34	2.83						1	0.794

ค่าคาดหวังที่คำนวณได้ในตารางที่ 6 มีค่าสอดคล้องกับข้อมูลอันดับที่นำไปปรับการแจกแจงก่อน ส่วนเบี่ยงเบนมาตรฐานมีค่าลดลง แต่ในกรณีนี้ค่าสหสัมพันธ์ของพื้นที่ในตำแหน่งที่ 3 และ 4 กับ 4 และ 5 มีค่าสูงถึง 0.92 ซึ่งพื้นที่ดังกล่าวเป็นตำแหน่งที่ข้อมูลอันดับที่นำไปปรับค่ามีอันดับขัดแย้งกับอันดับของค่าเฉลี่ยของการแจกแจงก่อน ค่าความน่าจะเป็นที่แต่ละพื้นที่จะกลายเป็นพื้นที่ที่จำเป็นต้องดำเนินการกำจัดสารพิษปนเปื้อนมีลำดับของค่าความน่าจะเป็นเปลี่ยนไปตามข้อมูลอันดับที่นำมาปรับค่า โดยเฉพาะความน่าจะเป็นของพื้นที่ในตำแหน่งที่ 3 ที่ค่าเพิ่มขึ้นจาก 0.023 มาเป็น 0.214

เช่นเดียวกับตัวอย่างแรก ที่ผลการคำนวณแสดงให้เห็นผลของการใช้ข้อมูลอันดับรูปแบบต่างกัน ว่าจะมีผลต่อการแจกแจงและโมเมนต์แบบมีเงื่อนไขอย่างไรบ้าง วิธีการที่เสนอในงานวิจัยนี้สามารถนำไปประยุกต์ใช้ในทางปฏิบัติเพื่อคำนวณค่าที่เปลี่ยนแปลงไปดังกล่าวนี้ได้

4.3 ด้านการจัดพอร์ตการลงทุน

ในการจัดพอร์ตการลงทุนตามหลัก mean-variance model จำเป็นต้องทราบค่าโมเมนต์ (2 ส่วน ประกอบด้วยค่าเฉลี่ย และเมทริกซ์ความแปรปรวนและความแปรปรวนร่วม) ของผลตอบแทนสำหรับ N

สินทรัพย์ เพื่อจะหาน้ำหนักการลงทุนที่เหมาะสม โดยมีวัตถุประสงค์เพื่อให้ได้ผลตอบแทนหลังปรับด้วยความเสี่ยงแล้วสูงที่สุด ปัญหาดังกล่าวแสดงได้ด้วยสัญลักษณ์ ดังนี้

$$\max_{\boldsymbol{w}_t} \boldsymbol{\mu}_t^{\mathrm{T}} \boldsymbol{w}_t - \frac{\gamma}{2} \boldsymbol{w}_t^{\mathrm{T}} \boldsymbol{\Sigma}_t \boldsymbol{w}_t \tag{22}$$

เมื่อ $\boldsymbol{\mu}_t$ และ $\boldsymbol{\Sigma}_t$ แทน เวกเตอร์ค่าเฉลี่ยและเมทริกซ์ความแปรปรวนและความแปรปรวนร่วมของผลตอบแทน (ส่วนเกินจากผลตอบแทนที่ปราศจากความเสี่ยง) ณ เวลา t และ γ แทนพารามิเตอร์ที่แสดงการหลีกเลี่ยงความเสี่ยงของนักลงทุน ในงานวิจัยนี้กำหนดให้มีค่าเป็น 1 ส่วน $\boldsymbol{w}_t$ เป็นเวกเตอร์ของน้ำหนักในการลงทุน t ณ เวลา ซึ่งเป็นตัวแปรตัดสินใจในบริบทของปัญหานี้ ฟังก์ชันวัตถุประสงค์ตาม (22) เป็นการคำนวณค่าอรรถประโยชน์ หรือ certainty equivalent return อีกนัยหนึ่งคือผลตอบแทนที่ปรับด้วยความเสี่ยงแล้วนั่นเอง ปัญหาดังกล่าวสามารถหาคำตอบได้จากสมการต่อไปนี้

$$\boldsymbol{w}_t^* = \frac{\boldsymbol{\Sigma}_t^{-1} \boldsymbol{\mu}_t}{\mathbf{1}^T \boldsymbol{\Sigma}_t^{-1} \boldsymbol{\mu}_t} \tag{23}$$

(ดูรายละเอียดเพิ่มเติมได้จากงานของ DeMiguel et al. (2009)) ข้อแตกต่างหลักในการกำหนดกลยุทธ์การลงทุนจะมาจากวิธีการกำหนดค่าพารามิเตอร์ $\widehat{\boldsymbol{\mu}}_t$ และ $\widehat{\boldsymbol{\Sigma}}_t$

ในตัวอย่างนี้จะจัดพอร์ตการลงทุนโดยใช้ค่าคาดหวังเมื่อมีข้อมูลประกอบเชิงอันดับ $\widehat{\boldsymbol{\mu}}_t$ เช่นเดียวกับงานศึกษาของ Chiarawongse et al. (2012) แต่แทนที่จะคำนวณค่าคาดหวังด้วยวิธีลูกโซ่มาร์คอฟ (Markov Chain Monte Carlo) งานวิจัยนี้จะใช้เทคนิคปริพันธ์เวียนเกิดแทน ซึ่งค่าที่ได้จากวิธีปริพันธ์เวียนเกิดจะมีความถูกต้องมากกว่า นอกจากนี้ในงานของ Chiarawongse et al. (2012) ศึกษาโดยการจำลองข้อมูล แต่ในงานวิจัยนี้จะศึกษาข้อมูลจริง โดยใช้ข้อมูลผลตอบแทนรายเดือนของ 10 กลุ่มอุตสาหกรรมระหว่างเดือนมกราคม ค.ศ. 1999 ถึงเดือนมิถุนายน ค.ศ. 2004 ที่รวบรวมได้จากเว็ปไซต์ของ Kenneth French

วัตถุประสงค์ในการศึกษาปัญหานี้เหมือนกับในการศึกษาของ Chiarawongse et al. (2012) ที่ต้องการชี้ให้เห็นถึงประโยชน์ที่ได้รับจากการใช้ข้อมูลประกอบเชิงอันดับในการจัดพอร์ตการลงทุน โดยในงานนี้จะทดลองใช้ค่า $\boldsymbol{\mu}_t$ สองแบบ โดยแบบแรกจะใช้ค่าในเวกเตอร์ค่าเฉลี่ยของการแจกแจงก่อนแทนด้วยสัญลักษณ์ $\widetilde{\boldsymbol{\mu}}_t$ ส่วนอีกแบบจะใช้ค่าตามค่าคาดหวังแบบมีเงื่อนไขของข้อมูลอันดับ ซึ่งจะให้แทนด้วยสัญลักษณ์ $\widehat{\boldsymbol{\mu}}_t$ โดยจะเปรียบเทียบผลการจัดพอร์ตการลงทุนเมื่อใช้ข้อมูลสองแบบที่กล่าวมา

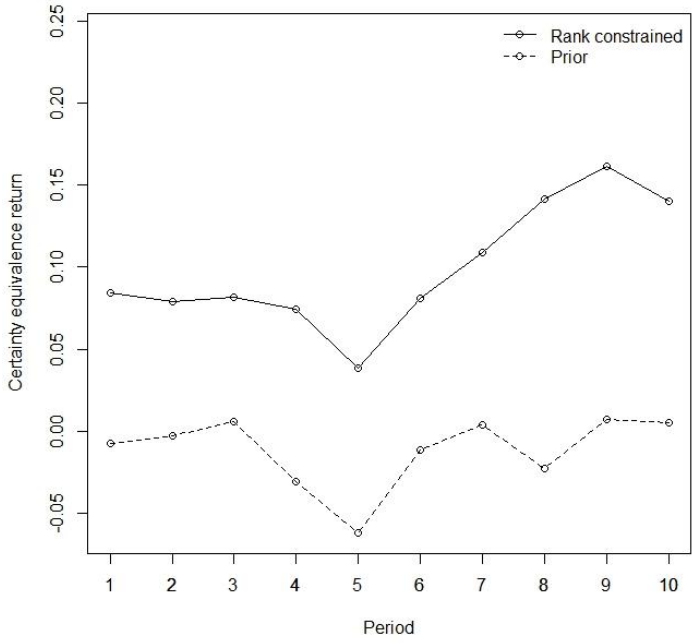
การคำนวณจะใช้วิธีเคลื่อนตัวอย่างทดลอง (rolling-sample analysis) ตามที่แสดงไว้ในงานของ DeMiguel et al. (2009) โดยในแต่ละรอบจะประมาณการแจกแจงก่อนในช่วงเวลา (window of length) 66 เดือน กล่าวคือ การคำนวณจะเริ่ม ณ เวลา $t = 66$ และจะใช้ตัวแบบ capital asset model ในการประมาณค่าพารามิเตอร์ $\overline{\boldsymbol{\mu}}_t$ และ $\overline{\boldsymbol{\Sigma}}_t$ ของการแจกแจงก่อนจากข้อมูล ณ เวลา $t - 65$ ถึง t จากนั้นใช้สมการที่ (23) ในการหาน้ำหนักที่เหมาะสมสำหรับการลงทุน ณ เวลานั้นๆ แล้วจึงทำเช่นเดียวกัน

สำหรับทุกๆ เดือน $t = 67, 68, \ldots, 186$ จนครบ หลังจากนั้น ผลตอบแทนของ 120 พอร์ตการลงทุนจะถูกนำมาแบ่งออกเป็น 10 ปี (12 เดือน) แล้วประเมินค่าอรรถประโยชน์ที่ได้รับในแต่ละปี จากสมการต่อไปนี้

$$\widehat{u_i} = \widehat{\bar{r}_i} - \frac{\widehat{s_i}^2}{2}, \quad i = 1, \ldots, 10$$

เมื่อ $\widehat{\bar{r}_i}$ และ $\widehat{s_i}^2$ เป็นค่าเฉลี่ยและความแปรปรวนของตัวอย่างในช่วงเวลาที่ i โดยค่าอรรถประโยชน์ หรือ certainty equivalent return นี้จะเป็นค่าที่ใช้ประเมินผลของการจัดพอร์ตการลงทุนในแต่ละปี โดยจะเทียบเมื่อใช้ค่าเฉลี่ยของการแจกแจงก่อนกับเมื่อใช้ค่าคาดหวังแบบมีเงื่อนไขของข้อมูลอันดับ

การทดลองทำโดยการคำนวณค่าตามสมการที่ (23) โดยใช้ค่า $\widehat{\boldsymbol{\mu}}_t$ และ $\overline{\boldsymbol{\Sigma}}_t$ มาคำนวณหาน้ำหนักการลงทุนที่เหมาะสม ณ เวลาต่างๆ ซ้ำจนครบทุกช่วงเวลาที่สนใจ $\widehat{\boldsymbol{\mu}}_t$ ใช้หลักการหาเช่นเดียวกับ Chiarawongse et al. (2012) ส่วน $\overline{\boldsymbol{\Sigma}}_t$ ใช้ค่าของการแจกแจงก่อน ในการประมาณค่า $\widehat{\boldsymbol{\mu}}_t$ พารามิเตอร์ของการแจกแจงก่อน $(\overline{\boldsymbol{\mu}}_t, \overline{\boldsymbol{\Sigma}}_t)$ จะถูกคำนวณก่อน แล้วจึงหาค่า $\widehat{\boldsymbol{\mu}}_t$ ซึ่งถือว่าเป็นตัวแปรสุ่มที่มีการแจกแจงแบบปกติหลายตัวแปร $(\overline{\boldsymbol{\mu}}_t, 0.1\overline{\boldsymbol{\Sigma}}_t)$ ที่ถูกปรับค่าด้วยข้อมูลอันดับของผลตอบแทน ณ เวลา $t + 1$ คำนวณผลตอบแทนจากพอร์ตการลงทุน $\hat{r}_i$ เช่นเดียวกันนี้ สำหรับข้อมูล ณ เวลา $t = 66, 67, \ldots, 185$



รูปที่ 1 ค่าอรรถประโยชน์ที่ได้จากการจัดพอร์ตการลงทุน 10 ปี เมื่อคำนวณโดยใช้ค่าคาดหวังแบบมีเงื่อนไขของข้อมูลอันดับ (เส้นทึบ) เปรียบเทียบกับเมื่อใช้ค่าเฉลี่ยของการแจกแจงก่อน (เส้นประ) ผลแสดงว่าการใช้ข้อมูลประกอบเชิงอันดับได้อรรถประโยชน์สูงกว่าทุกช่วงเวลา

ผลของค่าอรรถประโยชน์ในช่วง 10 ปี ที่ได้จากการใช้เวกเตอร์ผลตอบแทนแต่ละแบบแสดงไว้ตามรูปที่ 1 ซึ่งจะเห็นได้ว่าการใช้ค่าคาดหวังแบบมีเงื่อนไขของข้อมูลอันดับทำให้อรรถประโยชน์ที่ได้รับจากการลงทุนสูงกว่าการใช้ค่าคาดหวังจากการแจกแจงก่อนในทุกๆ ช่วงเวลา การวิเคราะห์ในตัวอย่างนี้ จึงเป็นการชี้ให้เห็นถึงประโยชน์ของการใช้วิธีปริพันธ์เวียนเกิดที่เสนอในงานวิจัยนี้เพื่อนำมาใช้คำนวณค่าคาดหวังแบบมีข้อจำกัดการจัดลำดับในการประยุกต์ใช้กับปัญหาด้านการลงทุนได้เป็นอย่างชัดเจน

5. สรุปผล และข้อเสนอแนะสำหรับงานวิจัยในอนาคต

งานวิจัยนี้ประกอบด้วยสามผลงาน ผลงานชิ้นแรก (Kiatsupaibul et.al. 2007) เป็นการนำเสนอวิธีการคำนวณความน่าจะเป็นและโมเมนต์ของตัวแบบสถิติที่มีข้อจำกัดการจัดลำดับสำหรับกรณีที่สำคัญสองกรณี คือ กรณีที่ตัวแปรสุ่มเป็นอิสระกัน และกรณีที่ตัวแปรสุ่มไม่เป็นอิสระกันแต่สามารถแทนด้วยตัวแบบปัจจัยเดียว ผลงานชิ้นที่สองและสามเป็นการประยุกต์ผลงานชิ้นแรกไปยังปัญหาด้านความเสี่ยง

กรณีที่ตัวแปรสุ่มที่เป็นอิสระกัน และทราบข้อมูลอันดับของตัวแปร เป็นปัญหาที่พบได้ทั่วไป เช่น ในการศึกษาด้าน Ranked Set Sampling งานวิจัยนี้ได้แสดงวิธีการคำนวณเพื่อปรับค่าการแจกแจงก่อนภายใต้เงื่อนไขของข้อมูลอันดับที่มี โดยวิธีการที่เสนอไว้ใช้หลักการปริพันธ์เวียนเกิดในการคำนวณ ซึ่งในขั้นตอนการคำนวณจะเป็นการหาปริพันธ์ใน 1 มิติ แม้ว่าตัวแปรที่สนใจศึกษาจะเป็นตัวแปรที่มีหลายมิติก็ตาม

สำหรับกรณีที่ตัวแปรไม่เป็นอิสระกันแต่สามารถแทนด้วยตัวแบบปัจจัยเดียว งานวิจัยนี้ได้แสดงการคำนวณเมื่อตัวแปรสุ่มมีการแจกแจงแบบปกติหลายตัวแปรที่มีโครงสร้างความสัมพันธ์ตามที่กำหนด และข้อมูลอันดับของตัวแปรอยู่ในรูปแบบที่ไม่ซับซ้อน อย่างไรก็ตามวิธีการที่นำเสนอสามารถนำไปประยุกต์ใช้กับข้อมูลอันดับในรูปแบบที่ซับซ้อนขึ้นได้ เช่น Umbrella Ordering หรือ Tree Ordering ซึ่งเป็นหัวข้องานวิจัยที่จะศึกษาต่อไปในอนาคต

ตัวอย่างที่แสดงในงานวิจัยนี้แสดงให้เห็นผลของข้อมูลอันดับแบบต่างๆ เช่น ข้อมูลอันดับที่สอดคล้องหรือตรงข้ามกับอันดับของค่าเฉลี่ยของการแจกแจงก่อน ที่มีต่อการแจกแจง ค่าคาดหวัง ส่วนเบี่ยงเบนมาตรฐาน และค่าสหสัมพันธ์ของตัวแปรสุ่มแตกต่างกันไป ผลที่ได้ในส่วนนี้จะเป็นประโยชน์ต่อผู้ที่จะนำเทคนิคการคำนวณด้วยวิธีปริพันธ์เวียนเกิดไปใช้จริงในทางปฏิบัติ ซึ่งแสดงในตัวอย่างสุดท้ายที่เป็นการประยุกต์เทคนิคกับปัญหาการตัดสินใจเกี่ยวกับการจัดพอร์ตการลงทุนจากข้อมูลผลตอบแทนจริง ตัวอย่างนี้แสดงให้เห็นได้ชัดเจนว่าการใช้ข้อมูลประกอบเชิงอันดับช่วยให้ผลการจัดพอร์ตการลงทุนดีขึ้นกว่าการพิจารณาเพียงค่าเฉลี่ยของการแจกแจงก่อนอย่างมีนัยสำคัญ

ผลงานชิ้นที่สอง (Kiatsupaibul et.al. 2007b) และผลงานชิ้นที่สาม (Hayter et.al. 2007) เป็นการประยุกต์ตัวสถิติโคโมโกรอฟในการอนุมานสถิติหลายตัวแปรพร้อม ๆ กัน ตัวสถิติโคโมโกรอฟจัดว่าเป็นตัวแบบสถิติที่มีข้อจำกัดการจัดลำดับ (Kiatsupaibul and Hayter 2015) ซึ่งการแจกแจงความน่าจะเป็น

สามารถคำนวณด้วยวิธีที่เสนอในผลงานชิ้นแรก ผลงานชิ้นที่สองเป็นการประยุกต์ตัวสถิติโคโมโกรอฟในการสร้างแถบความเชื่อมั่นของฟังก์ชันการแจกแจงแบบเบต้า และนำไปประยุกต์กับการจัดการความเสี่ยงด้านเครดิต ผลงานชิ้นที่สามเป็นการประยุกต์ตัวสถิติโคโมโกรอฟในการอนุมานสถิติสำหรับความน่าจะเป็นที่จะชนะเพื่อเปรียบเทียบตัวแบบไวบูลย์สองตัวแบบ ซึ่งผลของการเปรียบเทียบจะเป็นประโยชน์กับการตัดสินใจในงานด้านความเชื่อถือได้ของระบบ

รายการอ้างอิง

Ali, A., Meilă, M., 2012. Experiments with Kemeny ranking: What works when? *Math. Social Sci.* 64, 28–40.

Arnold, B.C., Balakrishnan,N., Nagaraja,H.N., 1992. *A First Course in Order Statistics*. Wiley, New Jersey.

Chen, Z., Bai, Z., Sinha, B.K., 2004. *Ranked Set Sampling*. Springer, New York.

Chiarawongse, A., Kiatsupaibul, S., Tirapat, S., Van Roy, B., 2012. Portfolio selection with qualitative input. *J. Bank. Finance* 36, 489–496.

David, H.A., Nagaraja, H.N., 2003. *Order Statistics*, third ed. Wiley, New Jersey.

DeMiguel, V., Garlappi, L., Uppal, R., 2009. Optimal versus naive diversification: how inefficient is the 1/n portfolio strategy? *Rev. Financ. Stud*. 22, 1915–1953.

Gaur, A., Mahajan, K.K., Arora, S., 2012. New nonparametric tests for testing homogeneity of scale parameters against umbrella alternative. *Statist. Probab. Lett*. 82, 1681–1689.

Hans, C., Dunson, D.B., 2005. Bayesian inferences on umbrella orderings. *Biometrics* 64 (4), 1018–1026.

Harter, H.L., Balakrishnan, N., 1996. C.R.C.Handbook of Tables for the Use of Order Statistics in Estimation. CRC Press Inc., Boca Raton.

Hayter, A.J., 2006. Recursive integration methodologies with statistical applications. *J. Statist. Plann. Inference* 136, 2284–2296.

Hayter, A.J., Liu, W., 1996. A note on the calculation of pr(x1 < ··· < xk). *Amer. Statist*. 50 (4), 365.

Hayter, A.J., P. Yang and S. Kiatsupaibul. 2017. Win-probabilities for comparing two Weibull distributions. *Quality Technology and Quantitative Management*. 14, 1-18.

Kemeny, J.L., Snell, J.G., 1962. *Mathematical Models in the Social Sciences*. Blaisdell, New York.

Khachiyan, L.G., 1989. The problem of computing the volume of polytopes is NP-hard. *Uspekhi Mat. Nauk* 44 (3), 199–200.

Kiatsupaibul, S. and A. J. Hayter. 2015 Recursive Confidence Band Construction for an Unknown Distribution Function. *Biometrical Journal*. 57, 39-51.

Kiatsupaibul, S., Smith, R.L., Zabinsky, Z.B., 2011. An analysis of a variation of hit-and-run for uniform sampling from general regions. *ACM Trans. Model. Comput. Simul*. 21 (3), Article number 16.

Kiatsupaibul, S., A. J. Hayter and W. Liu. 2017a. Rank constrained distribution and moment computations. *Computational Statistics and Data Analysis*. 105, 229-242.

Kiatsupaibul, S., A. J. Hayter and S. Somsong. 2017b. Confidence sets and confidence bands for a beta distribution with applications to credit risk management. *Insurance: Mathematics and Economics*. 75, 98-104.

Lovász, L., 1999. Hit-and-run mixes fast. *Math. Program*. 86, 443–461.

Lovász, L., Vempala, S., Fast algorithms for logconcave functions: Sampling, rounding, integration and optimization, in: Proc. Of the 47th IEEE Symposium on Foundations of Computer Science, FOCS 06, 2006, pp. 57–68.

McIntyre, G.A., 1952. A method for unbiased selective sampling, using ranked sets. *Aust. J. Agric. Res*. 3, 385–390.

Milgrom, P., Roberts, J., 1994. Comparing equilibrium. *Amer. Econ. Rev*. 84 (3), 441–459.

Milgrom, P., Shannon, C., 1994. Monotone comparative statics. *Econometrica* 62 (1), 157–180.

Nakas, C.T., Alonzo, T.A., 2007. ROC graphs for assessing the ability of a diagnostic marker to detect three disease classes with an umbrella ordering. *Biometrics* 63, 603–609.

Patil, G.P., 2002. Ranked set sampling. In: Encyclopedia of Environmetrics. Vol.3.John Wiley & Sons, Ltd., Chichester, pp.1684–1690.

Singh, P., Liu, W., 2006. A test against an umbrella or order alternative. *Comput. Statist. Data Anal.* 51, 1957–1964.

Smith, R.L., 1984. Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Oper. Res*. 32, 1296–1308.

Topkis, D.M., 1998. *Super modularity and Complementarity* .Princeton University Press.

Wolfe, D.A., 2004. Ranked set sampling: an approach to more efficient data collection. *Statist. Sci*. 19, 636–643.

Young, P., 1995.Optimal voting rules. *J. Econ. Perspect*. 9, 51–64.

Output จากโครงการวิจัยที่ได้รับทุนจาก สกว.

1. ผลงานตีพิมพ์ในวารสารวิชาการนานาชาติ (ระบุชื่อผู้แต่ง ชื่อเรื่อง ชื่อวารสาร ปี เล่มที่ เลขที่ และหน้า)

   1.1 Kiatsupaibul*, S., A. J. Hayter and W. Liu. Rank constrained distribution and moment computations. *Computational Statistics and Data Analysis*. (2017a) 105, 229-242.

   1.2 Kiatsupaibul*, S., A. J. Hayter and S. Somsong. Confidence sets and confidence bands for a beta distribution with applications to credit risk management. *Insurance: Mathematics and Economics*. (2017b) 75, 98-104.

   1.3 Hayter, A.J., P. Yang and S. Kiatsupaibul*. Win-probabilities for comparing two Weibull distributions. *Quality Technology and Quantitative Management*. (2017) 14, 1-18.

   * ผู้ประพันธ์หลัก (corresponding author)

2. การนำผลงานวิจัยไปใช้ประโยชน์

   - เชิงพาณิชย์

     - ผลงานชิ้นที่ 1 (Kiatsupaibul et.al. 2017a) สามารถนำไปใช้ในการจัดพอร์ตลงทุน
     - ผลงานชิ้นที่ 2 (Kiatsupaibul et.al. 2017b) สามารถนำไปใช้ในการจัดการความเสี่ยงด้านเครดิต
     - ผลงานชิ้นที่ 3 (Hayter et.al. 2017) สามารถนำไปใช้ในการตัดสินใจเกี่ยวกับความน่าเชื่อถือของระบบ

   - เชิงวิชาการ

     - ผลงานชิ้นที่ 2 เป็นส่วนหนึ่งในการสร้างนักวิจัยรุ่นใหม่ โดยอดีตนิสิตในหลักสูตรวิทยาศาสตร์มหาบัณฑิตสาขาสถิติได้มีโอกาสเป็นผู้ร่วมประพันธ์บทความ

ภาคผนวก 1

บทความวิจัยปีที่ 1

Kiatsupaibul, S., A. J. Hayter and W. Liu. 2017a. Rank constrained distribution and moment computations. *Computational Statistics and Data Analysis*. 105, 229-242.

Computational Statistics and Data Analysis 105 (2017) 229–242

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Rank constrained distribution and moment computations

Seksan Kiatsupaibul [a,*], Anthony J. Hayter [b], Wei Liu [c]

[a] Department of Statistics, Chulalongkorn University, Bangkok 10330, Thailand
[b] Department of Business Information and Analytics, University of Denver, Denver, USA
[c] School of Mathematics and S3RI, University of Southampton, Southampton, UK

HIGHLIGHTS

- An integration method is proposed for random variables conditioned on their ranking.
- High dimensional integration effort is reduced to either one or two dimensional integration.
- The method possesses a self-correction mechanism supported by numerical results.
- Reinforcing ranking and opposing ranking are defined and their effects are investigated.

ARTICLE INFO

Article history:
Received 5 March 2016
Received in revised form 16 July 2016
Accepted 12 August 2016
Available online 23 August 2016

Keywords:
Conditional distribution
Moments
Normal distribution
Order restriction
Ranked set sampling
Recursive integration

ABSTRACT

Consider a set of independent random variables with specified distributions or a set of multivariate normal random variables with a product correlation structure. This paper shows how the distributions and moments of these random variables can be calculated conditional on a specified ranking of their values. This can be useful when the ordering of the variables can be determined without observing the actual values of the variables, as in ranked set sampling, for example. Thus, prior information on the distributions and moments from their individual specified distributions can be updated to provide improved posterior information using the known ranking. While these calculations ostensibly involve high dimensional integral expressions, it is shown how the previously developed general recursive integration methodology can be applied to this problem so that they can be evaluated in a straightforward manner as a series of one-dimensional or two-dimensional integral calculations. Furthermore, the proposed methodology possesses a self-correction mechanism in the computation that prevents any serious growth of the errors. Examples illustrate how different kinds of ranking information affect the distributions, expectations, variances, and covariances of the variables, and how they can be employed to solve a decision making problem.

1. Introduction

The computations of conditional probabilities and conditional moments are often the basis for a statistical analysis in the sciences, social sciences, and engineering. This can be particularly true with a Bayesian approach. When updated information is in a qualitative form, the distribution of interest will be conditioned on a polytope. In general, such a problem is intractable numerically (see, for example, Khachiyan (1989)) and both theorists and practitioners usually resort to Monte Carlo methods, e.g., Smith (1984), Lovász (1999), Lovász and Vempala (2006) and Kiatsupaibul et al. (2011). However, an efficient numerical

* Corresponding author.
E-mail address: seksan@cbs.chula.ac.th (S. Kiatsupaibul).

http://dx.doi.org/10.1016/j.csda.2016.08.009
0167-9473/

method, when it is available, can have many advantages. In this work, the envelope of numerical methods is expanded to the computation of probability distribution and moments, that are conditioned on the important class of polytopes that are formed by rankings.

Consider a set of independent continuous random variables X_i with specified probability density functions $f_i(x_i)$, $1 \leq i \leq n$. The moments of these random variables are thus determined by their specified distributions. Suppose that information becomes available which indicates the ordering

$$X_1 \leq X_2 \leq \cdots \leq X_n. \tag{1}$$

The objective of this paper is to show how the information provided by this ranking can be used to provide updated distributions and moments for the random variables X_i.

This problem has applications to many areas where the ranking of the variables can be determined without observing the actual values of the variables. The literature on ranked set sampling provides discussions of many situations where this is the case (see McIntyre (1952), Patil (2002), Chen et al. (2004) and Wolfe (2004) for example). However, in ranked set sampling the information on the ordering of a set of variables is used to determine which variables to include in the sample and to subsequently observe. In contrast, this paper considers the problem where the ordering of a set of variables is used to augment prior information on their distributions, and the variables may never actually be observed. While these applications of the ranking information are different, both cases are similar in that they utilize information on the ranking of the variables without the full realizations of the variables being available.

As an example, suppose that prior distributions for the levels X_i of a particular medical condition may be available for a set of n patients based upon covariate values of the patients. While the actual levels of this condition may be very difficult or impossible to measure, there may be an ancillary variable that can be measured for the patients and which is sufficiently correlated with the condition of interest so that it can be used to infer the ranking of the X_i. It is then useful to obtain updated expectations and variances, say, of the levels X_i of the medical condition of interest based upon the information provided by the rankings.

Alternatively, Patil (2002) discusses a problem where a hazardous waste site inspector may be able to reliably rank areas of soil with respect to concentrations of a toxic contaminant, based on features like surface staining, discoloration, or the appearance of stressed vegetation. Thus, the actually contaminant levels X_i may have specified prior distributions, but their moments can be updated based upon the ranking provided by the soil features.

In general, additional information on ranking may be derived from different sources. Besides being derived from the observation of a covariate, rankings may be derived from conditions of economic systems, as in Topkis (1998), Milgrom and Roberts (1994), and Milgrom and Shannon (1994), for example. Furthermore, rankings can also be subjectively derived from a systematic preference aggregation process, as in Kemeny and Snell (1962), Young (1995), and Ali and Meilă (2012), for example. The methodology described in this paper is an important tool to directly incorporate such ranking information into the statistical inference process, as discussed by Chiarawongse et al. (2012), which can be applied in these various areas of study.

Calculations of the conditional distributions and moments of the X_i ostensibly involve the evaluation of an n-dimensional integral expression. However, it will be seen that by employing the technique of recursive integration (discussed in Hayter (2006)) the calculations can be performed easily as a series of 1-dimensional integral calculations. In fact, it will be seen that the recursive integration technique can also be employed for more general problems when the distributions and moments are conditioned on information more complex than just a simple ranking, and when the variables X_i have a multivariate normal distribution with a product correlation structure.

It is important that the recursive integration methodology does not suffer from a growth of errors that are compiled in high dimensions. In this application of recursive integration to the particular problem of the computations of conditional probabilities and conditional expectations, it is demonstrated that there exists a self-correction mechanism in the computation that prevents any serious growth of the errors. This condition has never been discussed before in the literature of recursive integration, and it confirms that the recursive integration technique is useful for high dimensional computations such as these.

Some examples are provided to show how the information provided by the ranking can affect the distributions, expectations, and variances of the random variables X_i. In particular, a reinforcing ranking can be considered to be one which is consistent with the rankings of the expectations of the distributions $f_i(x_i)$ (the prior expectations of the X_i), while various degrees of opposing rankings have some discrepancies with the rankings of these prior expectations. As illustrated in the examples, these different kinds of rankings will have different kinds of effects on the expectations and variances of the variables. It is also illustrated how ranking information can be useful for an important problem in portfolio selection, and applications of the proposed methodology to a real data set of asset returns are provided.

The layout of this paper is as follows. The theoretical discussion of how recursive integration can be used to calculate the conditional distributions and moments is provided in Section 2 for independent random variables. An extension to random variables X_i with a multivariate normal distribution with a product correlation structure is also provided in Section 2. Section 3 contains algorithms and details of the implementation of the procedure. A self-correction mechanism is discussed together with error rates and computational times. Some illustrative examples are provided in Section 4, and finally a conclusion is provided in Section 5.

2. General theory

The general theory concerning how to use recursive integration to calculate quantities such as moments conditional on information such as rankings is presented in this section, first for independent random variables and then for random variables with a multivariate normal distribution with a product correlation structure. Finally, some extensions are also discussed.

2.1. Independent random variables

Consider the set $S \subseteq \Re^n$ of values $\boldsymbol{X} = (X_1, \ldots, X_n)$ defined by

$$S = S_{1,2} \cap S_{2,3} \cap \cdots \cap S_{n-1,n}$$

where the set $S_{i,i+1}$ places restrictions on only X_i and X_{i+1}. Thus, the set S corresponds to the simple ordering in Eq. (1), for example, with

$$S_{i,i+1} = \{\boldsymbol{X} : X_i \leq X_{i+1}\}$$

for $1 \leq i \leq n-1$.

For any intervals (l_i, u_i), $1 \leq i \leq n$, it follows that

$$P(l_i \leq X_i \leq u_i;\ 1 \leq i \leq n \mid \boldsymbol{X} \in S) = \frac{A_1}{B} \tag{2}$$

where

$$A_1 = \int_{\boldsymbol{X} \in S^*} \cdots \int \prod_{i=1}^{n} f_i(x_i)\, dx_1 \ldots dx_n$$

and

$$B = P(\boldsymbol{X} \in S) = \int_{\boldsymbol{X} \in S} \cdots \int \prod_{i=1}^{n} f_i(x_i)\, dx_1 \ldots dx_n$$

with

$$S^* = S^*_{1,2} \cap S^*_{2,3} \cap \cdots \cap S^*_{n-1,n}$$

for

$$S^*_{1,2} = S_{1,2} \cap \{\boldsymbol{X} : l_1 \leq X_1 \leq u_1, l_2 \leq X_2 \leq u_2\}$$

and

$$S^*_{i,i+1} = S_{i,i+1} \cap \{\boldsymbol{X} : l_{i+1} \leq X_{i+1} \leq u_{i+1}\}$$

for $2 \leq i \leq n-1$.

Similarly, for any functions $g_i(x_i)$, $1 \leq i \leq n$, it follows that

$$E[g_1(X_1)g_2(X_2)\ldots g_n(X_n) \mid \boldsymbol{X} \in S] = \frac{A_2}{B} \tag{3}$$

where

$$A_2 = \int_{\boldsymbol{X} \in S} \cdots \int \prod_{i=1}^{n} (g_i(x_i) f_i(x_i))\, dx_1 \ldots dx_n.$$

While A_1, A_2, and B are each ostensibly n-dimensional integrals, they are each of the form of the integral in Section 1 of Hayter (2006) with $d = 1$, and so they can each be evaluated in a straightforward manner with a series of 1-dimensional integral computations using recursive integration, regardless of the value of n. Thus, the probability in Eq. (2) and the expectation in Eq. (3) can both be evaluated in a straightforward manner with a series of 1-dimensional integral computations.

Notice that the conditional joint cumulative distribution function of the X_i can be obtained from Eq. (2) with $l_i = -\infty$, $1 \leq i \leq n$, and the conditional marginal distribution of a particular variable can be obtained by taking $l_i = -\infty$ and $u_i = \infty$ for all of the other variables. Also, the conditional moments of X_i can be calculated with $g_i(x_i) = x_i^k$ and with all the other functions $g_j(x_j)$ equal to one, while the conditional covariance of X_{i_1} and X_{i_2}, say, can be calculated with $g_{i_1}(x_{i_1}) = x_{i_1}$ and $g_{i_2}(x_{i_2}) = x_{i_2}$ and again with all the other functions $g_j(x_j)$ equal to one.

2.2. Multivariate normal distribution with a product correlation structure

If the random variables X_i have a multivariate normal distribution with means μ_i, variances σ_i^2, and covariances $\rho_i\rho_j$, then it is possible to write

$$X_i = \mu_i + \rho_i M + \sqrt{\sigma_i^2 - \rho_i^2} Z_i, \quad 1 \le i \le n, \tag{4}$$

where M and the Z_i are independent standard normal random variables. Conditional on the value of M, the random variables X_i are thus independent normal random variables.

For the evaluation of Eqs. (2) and (3), conditioning on the value of M requires a 1-dimensional integral computation over the values m of M, with the integrand being the equation evaluated at each given value m. Since the integrand can be evaluated each time as a series of 1-dimensional integral computations, the overall computational intensity will consequently be equivalent to a series of 2-dimensional integral computations, regardless of the value of n.

It can also be noted that if the covariances are all equal and positive, so that the ρ_i are all equal to ρ, say, then for the simple ordering given in Eq. (1) the set S depends only on the Z_i and not on M. In this case, for moment calculations, the overall computational intensity may only be that of a 1-dimensional integral computation, depending on the functions $g_i(x_i)$. This reduction in computational intensity is possible when evaluating the conditional expectations of the X_i, for example, since the conditional expectations of the X_i will be equal to the conditional expectations of $\mu_i + \sqrt{\sigma_i^2 - \rho^2} Z_i$.

2.3. Extensions

In addition to the simple ordering in Eq. (1), the set S upon which the expressions are conditioned can encompass other types of information, such as the "umbrella" ordering

$$X_1 + c_1 \le X_2 + c_2 \le \cdots \le X_u + c_u \ge \cdots \ge X_{n-1} + c_{n-1} \ge X_n + c_n \tag{5}$$

for example, for any constants c_i, which has received considerable attention in the statistical literature (see Hans and Dunson (2005), Singh and Liu (2006), Nakas and Alonzo (2007), and Gaur et al. (2012), for example). Extensions can also be made to orderings of the random variables which form a tree structure, as discussed in Section 4 of Hayter (2006).

The simplicity of the evaluations of Eqs. (2) and (3) as a series of 1-dimensional integral computations using recursive integration is because of two conditions. These are firstly that the set S only places restrictions on "adjacent" variables X_i (although it should be remembered that any labeling of the n variables is permissible), and secondly that the integrand factors into the product of separate terms for each of the variables. These conditions are seen to be met for the simple ordering in Eq. (1) and the umbrella ordering in Eq. (5), and for independent variables X_i where the expectation is required of the product of the functions $g_i(x_i)$. In fact, conditioning on Eq. (1), the term B is just the probability of this simple ordering, and for independent random variables its evaluation by a series of 1-dimensional integral computations using recursive integration was first shown in Hayter and Liu (1996).

If the random variables X_i are not independent, or if the conditioning information imposes restrictions on non-adjacent X_i, then A_1, A_2, and B cannot necessarily be evaluated as a series of 1-dimensional integral computations. However, recursive integration of a higher order, in which the evaluation can be performed as a series of r-dimensional integral computations (with $r \ge 2$), say, may be possible depending upon the form of the expressions for A_1, A_2, and B.

It is also worth noting that for any set $T \subseteq \Re^n$ of values $\boldsymbol{X} = (X_1, \ldots, X_n)$ defined by

$$T = T_{1,2} \cap T_{2,3} \cap \cdots \cap T_{n-1,n}$$

where the set $T_{i,i+1}$ places restrictions on only X_i and X_{i+1}, then the conditional probability $P(\boldsymbol{X} \in T \mid \boldsymbol{X} \in S)$ is also equal to A_1/B with $S^*_{i,i+1} = S_{i,i+1} \cap T_{i,i+1}$. Thus, this conditional probability can also be evaluated as a series of 1-dimensional integral computations using recursive integration.

The utility of the methodology that is presented here is paramount when the random variables X_i have different distributions. This is because if the X_i are independent and identically distributed, then for the purpose of obtaining the conditional distribution and moments of a specific X_i, say, the information provided by the simple ordering in Eq. (1) is just equivalent to the information that X_i is the ith order statistic (the actual ordering of the $i-1$ variables less than X_i and the $n-i$ variables larger than X_i is irrelevant). In this case, the standard literature on order statistics (such as Arnold et al. (1992), Harter and Balakrishnan (1996), and David and Nagaraja (2003), for example) can be used to obtain the conditional information on X_i. However, when the random variables X_i are not identically distributed, then the simple ordering in Eq. (1) provides much more information than that X_i is simply the ith order statistic, and the methodology presented here allows all of that information to be utilized.

It may be the case that the ranking provided to the experimenter is incorrect, due to errors in its construction or simply perceived uncertainties. In fact, in Section 6 of Chiarawongse et al. (2012) it is pointed out with respect to financial applications that "When an analyst offers a qualitative view but is uncertain about its validity, it is useful for the decision maker to be provided with a measure of confidence. This could take the form of a probability that the view is valid".

In this case of an "imperfect ranking", Chiarawongse et al. (2012) proposed the shrinkage model

$$\kappa P(\boldsymbol{X} \in C \mid \boldsymbol{X} \in S) + (1-\kappa)P(\boldsymbol{X} \in C) \tag{6}$$

where C is any event and S is the observed ranking. Eq. (6) simply indicates that the experimenter uses a convex combination of the probability with the observed ranking and the prior probability. The additional parameter κ represents the estimate of the probability that the observed ranking is valid.

Notice that in this case the expectation of a variable can be expressed as a convex combination of the expectation with the observed ranking and that without ranking information

$$\kappa E[g(\boldsymbol{X}) \mid \boldsymbol{X} \in S] + (1-\kappa)E[g(\boldsymbol{X})]. \tag{7}$$

Also, it can be seen that once the expressions under the observed ranking are obtained, Eqs. (6) and (7) follow readily without additional computational efforts. In what follows, we consider distribution and moment computations only for the observed ranking. The computations for the case of an imperfect ranking under this shrinkage model then follow naturally from these computations.

3. Implementation with recursive integration

In this section first some formulas are provided for the implementation of the methodology with recursive integration, and the algorithms are explicitly provided. Finally, a discussion is provided of a self-correction mechanism and computational times.

3.1. Formulas for the recursive integration

To evaluate Eq. (2) for the independent random variables case in Section 2.1, B and A_1 can be evaluated according to the following recursive integration methodology. To evaluate B, the intermediate functions $b_1, \ldots, b_{n-1}$ can be sequentially evaluated, each with a one-dimensional integration. Let $b_0(z) = 1$ and for $i = 1, \ldots, n-1$ evaluate for each $z \in \Re$

$$b_i(z) = \int_{-\infty}^{z} b_{i-1}(x) f_i(x)\, dx \tag{8}$$

where f_i is the density of X_i. Then

$$B = \int_{-\infty}^{\infty} b_{n-1}(z) f_n(z)\, dz. \tag{9}$$

To evaluate A_1 in a similar manner, the intermediate functions $a_1, \ldots, a_{n-1}$ are sequentially evaluated. Here $a_0(z) = 1$ and for $i = 1, \ldots, n-1$, evaluate for each $z \in \Re$

$$a_i(z) = \int_{l_i}^{\max\{\min\{z, u_i\}, l_i\}} a_{i-1}(x) f_i(x)\, dx \tag{10}$$

so that

$$A_1 = \int_{l_n}^{u_n} a_{n-1}(z) f_n(z)\, dz. \tag{11}$$

To evaluate Eq. (3) for the independent random variables case in Section 2.1, B can be evaluated as above. To evaluate A_2, the intermediate functions $h_1, \ldots, h_{n-1}$ can be sequentially evaluated, each with a one-dimensional integration. Let $h_0(z) = 1$ and for $i = 1, \ldots, n-1$ evaluate for each $z \in \Re$

$$h_i(z) = \int_{-\infty}^{z} h_{i-1}(x) g_i(x) f_i(x)\, dx. \tag{12}$$

Then

$$A_2 = \int_{-\infty}^{\infty} h_{n-1}(z) g_n(z) f_n(z)\, dz. \tag{13}$$

To evaluate Eq. (2) for the multivariate normal case in Section 2.2, B and A_1 can be evaluated according to the following recursive integration methodology. To evaluate B, the intermediate functions $b_1, \ldots, b_{n-1}$ can be sequentially evaluated, each with a two-dimensional integration complexity. Let $b_0(m, z) = 1$ and for $i = 1, \ldots, n-1$, evaluate for each $m, z \in \Re$

$$b_i(m, z) = \int_{-\infty}^{z} b_{i-1}(m, x) \phi_i(m, x)\, dx, \tag{14}$$

where ϕ_i is the density of a $N(\mu_i + \rho_i m, \sigma_i^2 - \rho_i^2)$ random variable. Then with ϕ as the standard normal density

$$B = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(m) b_{n-1}(m, z) \phi_n(m, z)\, dz\, dm. \tag{15}$$

To evaluate A_1 in a similar manner with the intermediate functions $a_1, \ldots, a_{n-1}$, let $a_0(m, z) = 1$ and for $i = 1, \ldots, n-1$, evaluate for each $m, z \in \Re$

$$a_i(m, z) = \int_{l_i}^{\max\{\min\{z, u_i\}, l_i\}} a_{i-1}(m, x) \phi_i(m, x)\, dx. \tag{16}$$

Then

$$A_1 = \int_{m=-\infty}^{\infty} \int_{z=l_n}^{u_n} \phi(m) a_{n-1}(m, z) \phi_n(m, z)\, dz\, dm. \tag{17}$$

To evaluate Eq. (3), the formulas for evaluating A_2 with the intermediate functions $h_1, \ldots, h_{n-1}$ are $h_0(m, z) = 1$, and for $i = 1, \ldots, n-1$

$$h_i(m, z) = \int_{-\infty}^{z} h_{i-1}(m, x) g_i(x) \phi_i(m, x)\, dx \tag{18}$$

with

$$A_2 = \int_{m=-\infty}^{\infty} \int_{z=-\infty}^{\infty} \phi(m) h_{n-1}(m, z) g_n(z) \phi_n(m, z)\, dz\, dm. \tag{19}$$

3.2. Algorithms

The evaluations of Eqs. (9), (11) and (13) are accomplished with a sequence of n 1-dimensional numerical integrations, with each integration being evaluated by a sum on a truncated-discretized real line. Algorithm 1 can be used to compute Eq. (13) with the integrals being performed with a first order Newton–Cotes formula (the trapezoidal rule). To compute Eqs. (9) and (11), simply replace g_i, $i = 1, \ldots, n$ in the algorithm with appropriate indicator functions or with a constant function equal to one, respectively. R code is available from the authors to implement this algorithm.

Algorithm 1 Computation of A_2 in equation (13)

1: Assume n variables with ranking $X_1 \le X_2 \cdots \le X_n$.
2: Discretization grid size Δ with lower bound x_0, forming $N + 1$ grid points

$$\{x_0, x_1, \ldots, x_N\},$$

where $x_j = x_{j-1} + \Delta$ for $j = 1, \ldots, N$.
3: Let $h_0(x_j) = 1$ for $j = 0, 1, \ldots, N$.
4: **for** $i = 1$ to n **do**
5: Let, for $j = 1, \ldots, N$,

$$\overline{h}_j = \frac{h_{i-1}(x_{j-1}) g_i(x_{j-1}) f_i(x_{j-1}) + h_{i-1}(x_j) g_i(x_j) f_i(x_j)}{2}.$$

6: Let, for $j = 1, \ldots, N$,

$$h_i(x_j) = \sum_{k=1}^{j} \overline{h}_k \Delta.$$

and let $h_i(x_0) = h_i(x_1)$.
7: **end for**
8: **return** $A_2 = h_n(x_N)$.

The evaluations of Eqs. (15), (17) and (19), require a sequence of 2-dimensional numerical integrations. Algorithm 2 can be used to compute (19) with the integrals being performed with the first order Newton–Cotes formula. To compute Eqs. (15) and (17), simply replace g_i, $i = 1, \ldots, n$ in Algorithm 2 with appropriate indicator functions or with a constant function equal to one, respectively. Again, R code is available from the authors to implement this algorithm.

Algorithm 2 Computation of A_2 in equation (19)

Assume n variables with ranking $X_1 \leq X_2 \cdots \leq X_n$.
Discretization grid size δ with lower bound m_0, forming $M+1$ grid points

$$\{m_0, m_1, \ldots, m_M\},$$

where $m_l = m_{l-1} + \delta$ for $l = 1, \ldots, M$.
for $l = 0$ to M **do**
  Discretization grid size Δ with lower bound x_0, forming $N+1$ grid points

$$\{x_0, x_1, \ldots, x_N\},$$

  where $x_j = x_{j-1} + \Delta$ for $j = 1, \ldots, N$.
  Let $h_0(m_l, x_j) = 1$ for $j = 0, 1, \ldots, N$
  for $i = 1$ to n **do**
    Let, for $j = 1, \ldots, N$,

$$\overline{h}_j = \frac{h_{i-1}(m_l, x_{j-1}) g_i(x_{j-1}) \phi_i(m_l, x_{j-1}) + h_{i-1}(m_l, x_j) g_i(x_j) \phi_i(m_l, x_j)}{2}.$$

    Let, for $j = 1, \ldots, N$,

$$h_i(m_l, x_j) = \sum_{k=1}^{j} \overline{h}_k \Delta.$$

    and let $h_i(m_l, x_0) = h_i(m_l, x_1)$.
  end for
  $\tilde{h}(m_l) = h_n(m_l, x_N)$.
end for
Let, for $l = 1, \ldots, M$,

$$\overline{h}_l = \frac{\phi(m_l)\tilde{h}(m_l) + \phi(m_{l-1})\tilde{h}(m_{l-1})}{2}.$$

return $A_2 = \sum_{l=1}^{M} \overline{h}_l \delta$.

3.3. Self-correction mechanism

It is useful to point out that the conditional probability and the conditional expectation computations based on the recursive integration methodology possess a self-correction mechanism. To see this, observe that the target value is of the form

$$f(A, B) = \frac{A}{B}$$

where A is either A_1 in Eq. (2) or A_2 in Eq. (3). Let $\hat{A}$ and $\hat{B}$ denote the computed values of A and B, respectively. The computed value of the target quantity is then $f(\hat{A}, \hat{B})$, and with a first order approximation

$$\begin{aligned} f(\hat{A}, \hat{B}) &\approx f(A, B) + f_1'(A, B)(\hat{A} - A) + f_2'(A, B)(\hat{B} - \hat{B}) \\ &= f(A, B) + \frac{1}{B}(\hat{A} - A) - \frac{A}{B^2}(\hat{B} - \hat{B}) \end{aligned} \tag{20}$$

where f_1' and f_2' are the partial derivatives of f with respect to its first and second arguments. If we let ε, ε_A and ε_B be the computational error of the target value, that of A and that of B, respectively, so that

$$\begin{aligned} f(\hat{A}, \hat{B}) &= f(A, B) + \varepsilon \\ \hat{A} &= A + \varepsilon_A \\ \hat{B} &= B + \varepsilon_B, \end{aligned}$$

then Eq. (20) implies that

$$\varepsilon \approx \frac{1}{B}\varepsilon_A - \frac{A}{B^2}\varepsilon_B. \tag{21}$$

Thus, when A and both B are positive, and when ε_A and ε_B have the same sign, the computational errors of A and B tend to cancel each other in producing the total error of the target value. Notice that for the applications in this paper, B is positive since it is a probability, and A is positive when it is a probability and may be positive when it is an expectation.

Table 1
The numerical errors and the computational times for the cumulative distribution function evaluated at three points for the 70th order statistic of $n = 101$ independent $U[0, 1]$ random variables.

Grid size	True value	Comp. value	Error	Comp. time (s)	Comp. value	Error
					With exact denominator	
0.01	0.03382186	0.09552326	6.170e−02	0.00	231.1455	$\gg 1$
	0.60791267	0.65772572	4.981e−02	0.00	1591.5334	$\gg 1$
	0.99628437	0.99005003	6.234e−03	0.00	2395.7061	$\gg 1$
0.001	0.03382186	0.03422536	4.035e−04	0.00	0.03739016	3.568e−03
	0.60791267	0.60800312	9.045e−05	0.02	0.66422472	5.631e−02
	0.99628437	0.99620053	8.385e−05	0.00	1.08831847	9.203e−02
0.0001	0.03382186	0.03382585	3.987e−06	0.06	0.03385576	3.390e−05
	0.60791267	0.60791339	7.206e−07	0.07	0.60845101	5.383e−04
	0.99628437	0.99628353	8.411e−07	0.08	0.99716462	8.802e−04
0.00001	0.03382186	0.03382190	3.982e−08	1.36	0.03382220	3.387e−07
	0.60791267	0.60791268	7.037e−09	1.20	0.60791805	5.380e−06
	0.99628437	0.99628437	8.414e−09	1.42	0.99629317	8.797e−06

Table 2
The numerical errors and the computational times for the cumulative distribution function evaluated at three points for the 70th order statistic of $n = 101$ independent standard normal random variables.

Grid size	True value	Comp. value	Error	Comp. time (s)	Comp. value	Error
					With exact denominator	
0.01	0.03382186	0.03766693	3.845e−03	0.05	0.08004251	4.622e−02
	0.60791267	0.61814333	1.023e−02	0.05	1.31355904	7.056e−01
	0.99628437	0.99663459	3.502e−04	0.03	2.11785572	1.122e+00
0.001	0.03382186	0.03367669	1.452e−04	0.39	0.03393352	1.117e−04
	0.60791267	0.60697199	9.407e−04	0.36	0.61160101	3.688e−03
	0.99628437	0.99632079	3.642e−05	0.39	1.00391914	7.635e−03
0.0001	0.03382186	0.03379485	2.701e−05	4.38	0.03379548	2.638e−05
	0.60791267	0.60791453	1.860e−06	4.36	0.60792588	1.320e−05
	0.99628437	0.99628272	1.657e−06	4.39	0.99630131	1.694e−05
0.00001	0.03382186	0.03382366	1.796e−06	38.64	0.03382172	1.365e−07
	0.60791267	0.60791215	5.155e−07	36.95	0.60787742	3.525e−05
	0.99628437	0.99628432	5.078e−08	38.14	0.99622740	5.697e−05

This error cancellation effect or self-correction mechanism is strongest when the values of A and B are comparable, and when the values of ε_A and ε_B are comparable. For the problems considered in this paper, the processes of computing $\hat{A}$ and $\hat{B}$ share some common numerical integration sequences, and hence ε_A and ε_B will tend to have the same sign. However, the level of the error cancellation depends upon the relative values of A and B, and the relative values of ε_A and ε_B. This is illustrated and further discussed in the following numerical examples.

This self-correction mechanism caused by the cancellation of the errors from the numerator and the denominator is a useful property of the implementation of the recursive integration methodology for the problems discussed in this paper. In fact, even in the case where the exact value of the denominator B might be known, according to (21) it may be better that $\hat{B}$ is computed and employed in evaluating the ratio since the self-correction mechanism applies. The estimate obtained by employing $\hat{B}$ can be interpreted as an estimate that has been formed by a legitimate discrete distribution induced by the discretization procedure. The resulting estimate approaches the true value when the discretization becomes finer. In the following examples it is shown that not taking advantage of this self-correction mechanism causes a significant increase in the error of the estimate if the discretization is not fine enough.

Some calculations are now presented to demonstrate the errors for problems with independent identically distributed random variables where the solutions are known. Specifically, consider the cases of $n = 101$ independent uniform $[0, 1]$ random variables or independent standard normal random variables. In both cases the cumulative distribution at three points of X_{70}, under the condition $X_1 \leq \cdots \leq X_{101}$, was evaluated by the recursive integration methodology for various grid sizes, and the computed values of the probability, the computational errors, and the computational times are shown in Tables 1 and 2 (note that the true values can be obtained from the cumulative distribution of a binomial distribution). The computations were implemented in R on a 64-bit Windows machine with an Intel Core i5-2500 3.30 GHz CPU.

Next, the expectations of X_{17} and X_{51}, under the condition $X_1 \leq \cdots \leq X_{101}$, were evaluated by the recursive integration methodology for various grid sizes, and the computed values of the expectations, the computational errors, and the computational times are shown in Table 3 for the case of independent uniform $[0, 1]$ variables and in Table 4 for the case of independent standard normal variables. In the first case the true values of $E[X_{17}]$ and $E[X_{51}]$ are known to be $1/6$ and 0.5,

Table 3
The numerical errors and the computational times for the expectations of $X_{(17)}$ and $X_{(51)}$ of $n = 101$ independent $U[0, 1]$ random variables.

Grid size	True value	Comp. value	Error	Comp. time (s)	Comp. value	Error
					With exact denominator	
0.01	1/6	0.14369985	2.297e−02	0.00	347.7224	≫1
	0.5	0.48606226	1.394e−02	0.00	1176.1651	≫1
0.001	1/6	0.16661772	4.894e−05	0.00	0.18202474	1.536e−02
	0.5	0.49997063	2.937e−05	0.00	0.54620256	4.620e−02
0.0001	1/6	0.16666624	4.272e−07	0.05	0.16681363	1.470e−04
	0.5	0.49999974	2.563e−07	0.06	0.50044193	4.419e−04
0.00001	1/6	0.16666666	4.215e−09	0.92	0.16666814	1.469e−06
	0.5	0.50000000	2.529e−09	0.89	0.50000442	4.416e−06

Table 4
The numerical errors and the computational times for the expectations of $X_{(17)}$ and $X_{(51)}$ of $n = 101$ independent standard normal random variables.

Grid size	True value	Comp. value	Error	Comp. time (s)	Comp. value	Error
					With exact denominator	
0.01	NA	−0.9672	NA	0.04	−2.0552	NA
	0.0000	0.0000	1.397e−14	0.05	0.0000	2.970e−14
0.001	NA	−0.9779	NA	0.38	−0.9854	NA
	0.0000	0.0000	1.734e−16	0.39	0.0000	1.747e−16
0.0001	NA	−0.9780	NA	4.33	−0.9781	NA
	0.0000	0.0000	9.826e−18	4.25	0.0000	9.826e−18
0.00001	NA	−0.9780	NA	38.55	−0.9780	NA
	0.0000	0.0000	1.281e−17	37.22	0.0000	1.281e−17

while in the second case the true value of $E[X_{17}]$ is unknown but is about $\Phi^{-1}(1/6) \approx 0.97$, and the true value of $E[X_{51}]$ is known to be zero.

In these tables the second column shows the known true values, while the third column shows the estimates from the proposed methodology. The fourth and fifth columns show the errors (from the true values) and the computational time of the proposed methodology. Also, both for the independent uniform [0, 1] variables and the independent standard normal variables the true values of the denominators B in Eqs. (2) and (3) are known to be $1/n!$. The sixth and seventh columns then show the estimates and errors when the true values of the denominators are employed and the recursive integration is used only for computing the numerators A_1 and A_2.

First of all, it can be seen from these tables that these calculations which involve 101 successive one-dimensional numerical integrations attain a small error with a very reasonable computation time. In fact, with the potential optimization of the coding on a low level computer programming language such as C++ or Java, the computation can be expected to be accelerated even more.

Furthermore, special attention should be given to the results in the sixth and seventh columns. In the sixth column the estimates computed by employing the true values for the denominators do not take advantage of the self-correction mechanism feature of the proposed methodology given in Eq. (21). It can be seen that when the discretization grid sizes are not very small, the errors can become so large that the estimates are unreasonable. Specifically, some estimates for the conditional probabilities in Table 1 and in Table 2 when the grid sizes are 0.01 and 0.001 are much greater than one. Note that such unreasonable values of the estimates do not occur when the methodology is applied appropriately to both the numerators and the denominators and the self-correction mechanism applies (as shown in the third column of each table).

In almost all cases in Table 1 and in Table 2 the errors from the proposed methodology with self-correction mechanism are much smaller than the corresponding values without the feature. The exceptions are row 4, row 7 and row 10 in Table 2. The errors with the self-correction mechanism are not smaller than those without the feature in these cases, although they are close. One explanation for this is that the distribution function is evaluated at a low quantile, so that A is much smaller than B in Eq. (21). Furthermore, at this low quantile $\hat{A}$ and $\hat{B}$ do not share a lot of common integration sequences, so that ε_A and ε_B may be quite different. Consequently, the error cancellation in Eq. (21) does not apply substantially, although in these exceptional cases the differences between the errors are very small.

In Table 3 the estimates with the self-correction mechanism are much more accurate than those without the self-correction mechanism. According to Eq. (21), this is because the value of A is comparable to that of B, causing a strong error cancellation. In Table 4 when the true value of $E[X_{51}]$ is known to be zero, A is zero. According to Eq. (21) there is therefore no error cancellation, and consequently the errors of the estimates with or without self-correction mechanism are quite similar. In the case where the true value of $E[X_{17}]$ is unknown it can be observed that the estimate converges to a certain value as the grid size becomes smaller. The estimate with the self-correction mechanism seems to converge more quickly than the estimate without this feature.

Table 5
Example 1—Medical conditions of five patients.

Reactor	Expectation	Standard deviation	Correlation matrix					$P(X_i \leq 5)$
Prior								
1	3	1.73	1	0	0	0	0	0.876
2	6	3.46		1	0	0	0	0.456
3	9	5.20			1	0	0	0.234
4	12	6.93				1	0	0.132
5	15	8.66					1	0.080
Reinforcing ranking $X_1 \leq X_2 \leq X_3 \leq X_4 \leq X_5$								
1	2.32	1.22	1	0.36	0.17	0.08	0.03	0.968
2	4.89	2.03		1	0.49	0.24	0.10	0.578
3	8.19	3.05			1	0.52	0.23	0.132
4	12.86	4.63				1	0.45	0.010
5	21.41	8.34					1	0.000
Opposing ranking $X_5 \leq X_4 \leq X_3 \leq X_2 \leq X_1$								
1	7.99	2.18	1	0.82	0.69	0.57	0.40	0.064
2	6.68	1.86		1	0.85	0.70	0.49	0.184
3	5.64	1.69			1	0.82	0.58	0.382
4	4.58	1.56				1	0.71	0.641
5	3.24	1.45					1	0.881

4. Examples

Three examples are presented in this section. The first is a healthcare example where the random variables are taken to have independent gamma distributions with equal shape parameters but different scale parameters. The second is a soil contamination example where the random variables are taken to have a multivariate normal distribution with different means but equal variances and covariances. It is shown how information on both reinforcing rankings and opposing rankings affects the distributions, expectations, variances, and covariances of the variables. The third example concerns portfolio selection in finance.

4.1. Healthcare

Suppose that the levels X_i of a medical condition of $n = 5$ patients are of interest, but that they cannot be directly measured. However, the levels can be modeled with a gamma distribution

$$f(x; 3, \theta) = \frac{1}{2\theta^3} x^2 e^{-x/\theta}$$

with a shape parameter $k = 3$ and with a scale parameter θ that depends upon some covariate values of the patients. Specifically, suppose that the five patients have scale parameters $\theta_i = i$, $1 \leq i \leq 5$, so that based upon these distributions the prior expectations and variances are $E(X_i) = 3i$ and $\mathrm{Var}(X_i) = 3i^2$, $1 \leq i \leq 5$. It should also be noted that the variables X_i are modeled to be independent.

Now suppose that an ancillary measurement becomes available for the five patients that provides the information that

$$X_1 \leq X_2 \leq X_3 \leq X_4 \leq X_5$$

(or equivalently, this same ranking with strict inequalities). This is a reinforcing ranking since it matches the ranking of the prior expectations of the X_i. It is interesting to note that under the prior distributions this reinforcing ranking has a probability of 0.107 (this is B in Eqs. (2) and (3)). Using the recursive integration techniques discussed in Section 2, the conditional expectations, standard deviations, and correlations of the X_i (conditionally the X_i are no longer independent) are shown in Table 5.

It can first be noted that with this reinforcing ranking the conditional expectations of the X_i have maintained their ordering but are now more spread out than the prior expectations. Furthermore, the conditional standard deviations are each smaller than the prior standard deviations. Also, it can be seen that the correlations are largest for adjacent variables.

In addition, suppose that it has been decided that urgent corrective action needs to be taken whenever the level of this deterioration condition is less than 5. Table 5 also shows how these probabilities change under the knowledge provided by the reinforcing ranking. It can be seen that the probabilities that urgent corrective action needs to be taken become larger for patients 1 and 2, and become smaller for patients 3, 4 and 5.

Now consider the opposing ranking

$$X_5 \leq X_4 \leq X_3 \leq X_2 \leq X_1$$

Table 6
Example 2—Toxic contamination levels at six locations.

Location	Expectation	Standard deviation	Correlation matrix						$P(X_i \geq 18)$
Prior									
1	10	3	1	0.40	0.40	0.40	0.40	0.40	0.004
2	10	3		1	0.40	0.40	0.40	0.40	0.004
3	12	3			1	0.40	0.40	0.40	0.023
4	15	3				1	0.40	0.40	0.159
5	18	3					1	0.40	0.500
6	20	3						1	0.748
Reinforcing ranking $X_1 \leq X_2 \leq X_3 \leq X_4 \leq X_5 \leq X_6$									
1	8.21	2.59	1	0.79	0.67	0.59	0.56	0.52	0.000
2	10.28	2.45		1	0.78	0.65	0.60	0.55	0.001
3	12.47	2.45			1	0.74	0.62	0.56	0.012
4	15.12	2.50				1	0.71	0.58	0.125
5	17.97	2.54					1	0.69	0.494
6	20.95	2.71						1	0.863
Partially opposing ranking $X_1 \leq X_2 \leq X_5 \leq X_4 \leq X_3 \leq X_6$									
1	8.52	2.64	1	0.76	0.60	0.61	0.62	0.48	0.000
2	10.88	2.52		1	0.65	0.68	0.69	0.51	0.002
3	16.13	2.36			1	0.92	0.86	0.60	0.214
4	15.08	2.31				1	0.92	0.59	0.104
5	14.05	2.34					1	0.58	0.046
6	20.34	2.83						1	0.794

which is completely opposite to the ranking of the prior expectations of the X_i. In fact, under the prior distributions this opposing ranking has a very small probability of 0.00003. Under this opposing ranking the conditional expectations, standard deviations, and correlations of the X_i are also shown in Table 5.

It can be seen that with this opposing ranking the order of the conditional expectations has switched to match this ranking, and that the conditional expectations are less spread out than the prior expectations. The conditional standard deviations are also much smaller than the prior standard deviations, and their order also matches the opposing ranking. Again, the correlations are largest for the adjacent variables, and they are all much larger than the correlations for the reinforcing ranking. Also, the probabilities that urgent corrective action needs to be taken are now ordered to match the opposing ranking.

In summary, this example illustrates how knowledge of the ranking can result in important changes in the distributions and moments of the variables, which will be important information for practitioners.

4.2. Hazardous waste sites

Suppose that based upon knowledge of polluting activities, scientists originally model the unknown toxic contamination levels X_i at $n = 6$ locations with a multivariate normal distribution with means $\boldsymbol{\mu} = (10, 10, 12, 15, 18, 20)$, standard deviations all equal to 3, and correlations all equal to 0.4. Then suppose that subsequently surface features indicate the reinforcing ranking

$$X_1 \leq X_2 \leq X_3 \leq X_4 \leq X_5 \leq X_6.$$

Under the prior distribution this reinforcing ranking has a probability of 0.111. In this case the conditional expectations, standard deviations, and correlations of the toxic contamination levels are shown in Table 6.

It can be seen that with this reinforcing ranking the conditional expectations are quite similar to the prior expectations, although for location 1 the expectation has decreased from 10 to 8.21, while for location 6 the expectation has increased from 20 to 20.94. The standard deviations have all decreased and are all fairly similar, while the correlations have increased and are largest for the adjacent variables.

Also, suppose that it has been decided that decontamination needs to be taken whenever the toxic contamination level is larger than 18. It can be seen from Table 6 that the reinforcing ranking has increased the probability that decontamination needs to be taken at location 6 from 0.748 to 0.865, while these probabilities have fallen at the other 5 locations.

Now suppose that subsequently surface features indicate the partially opposing ranking

$$X_1 \leq X_2 \leq X_5 \leq X_4 \leq X_3 \leq X_6$$

where the ordering of the toxic contamination levels at locations 3, 4, and 5 is opposite to their prior expectations. This partially opposing ranking has a probability of 0.002 under the prior distribution.

Table 6 shows that the conditional expectations are now ordered in the same way as this ranking. In addition, the standard deviations have decreased, but now there are very high correlations of 0.92 between locations 3 and 4, and between locations

4 and 5, which are the locations where the ranking contradicts the prior expectations. The probabilities that decontamination needs to be taken are now ordered in the same way as the partially opposing ranking, and specifically the probability at location 3 has risen from 0.023 to 0.214.

As with Example 1, this example illustrates how the different rankings can result in important changes in the distributions and moments of the variables, and the methodology presented in this paper allows practitioners to calculate those changes.

4.3. Portfolio selection

The celebrated mean–variance portfolio selection model requires two sets of moments (the means and the variance–covariance matrix) of the returns on N assets in order to recommend the proportions of capital, or the portfolio weights, to be invested in these N assets. The objective is to optimize the risk-return trade-off of the entire portfolio.

The problem can be expressed as

$$\max_{\boldsymbol{w}_t} \quad \boldsymbol{\mu}_t^\top \boldsymbol{w}_t - \frac{\gamma}{2}\boldsymbol{w}_t^\top \boldsymbol{\Sigma}_t \boldsymbol{w}_t \tag{22}$$

where $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$ are the mean vector and the variance–covariance matrix of the return (in excess of risk free) at time t, γ is a parameter representing risk aversion of the investor which we set equal to 1 in this analysis, and $\boldsymbol{w}_t$ is the vector of decision variables that represent the portfolio weights at time t. The objective function in (22) is the certainty-equivalent return which is a utility function interpreted as the return that is penalized by risk. The solution to the optimization problem is

$$\boldsymbol{w}_t^* = \frac{\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t}{\mathbf{1}^\top \boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t} \tag{23}$$

(see, for example, DeMiguel et al. (2009) for more details). The main difference between one portfolio selection strategy and another is how the parameter estimates $\hat{\boldsymbol{\mu}}_t$ and $\hat{\boldsymbol{\Sigma}}_t$ are obtained.

In this example the estimate $\hat{\boldsymbol{\mu}}_t$ is based on the rank constrained statistical estimates proposed by Chiarawongse et al. (2012), where the estimates are obtained by a Markov chain Monte Carlo. Here the calculations are replaced by the recursive integration methodology presented in this paper in order to obtain more accurate estimates. Furthermore, in Chiarawongse et al. (2012) the authors perform their experiments based on simulated data sets, whereas in this example a real data set is considered instead of ten industry monthly asset returns during the period of 01/1999–06/2014 obtained from Kenneth French's web site.

Our objective is the same as that in Chiarawongse et al. (2012), which is to demonstrate the potential benefit of adopting a rank constrained statistical estimate. Two experiments are performed, one using *prior* parameter estimates $\tilde{\boldsymbol{\mu}}$ and the other using rank constrained statistical estimates $\hat{\boldsymbol{\mu}}$. The performances of the two models are then compared.

In the analysis that follows the rolling-sample approach appearing in DeMiguel et al. (2009) is employed. An estimation window of length 66 months is fixed. Then, starting from $t = 66$ and applying the capital asset pricing model, the data from month $t - 65$ to month t are used to estimate the *prior* parameters $(\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t)$. These are then applied to (23) to obtain the weight vector $\tilde{\boldsymbol{w}}_t$ for $t = 66, 67, \ldots, 185$. The weight at time t is applied to the out-of-sample returns at $t + 1$ and summed across assets to provide 120 out-of-sample portfolio returns $\tilde{r}_t$, $t = 67, 68, \ldots, 186$. Subsequently, the 120 portfolio returns are then divided into 10 twelve-month periods. In each period a certainty equivalent return is estimated as

$$\tilde{u}_i = \bar{\tilde{r}}_i - \tilde{s}_i^2/2, \quad i = 1, \ldots, 10,$$

where $\bar{\tilde{r}}_i$ and $\tilde{s}_i^2$ are the averages and the sample variances of the $\tilde{r}_t$ in period i. These certainty equivalent returns are the performance measurements of the prior model which were compared with those obtained from the following rank constrained statistical counterpart.

The experiment was repeated with $(\hat{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t)$ in (23) to obtain different portfolio weights, where $\hat{\boldsymbol{\mu}}_t$ are the rank constrained estimates based on Chiarawongse et al. (2012) and $\tilde{\boldsymbol{\Sigma}}_t$ is the same as in the prior model. To estimate $\hat{\boldsymbol{\mu}}_t$, $(\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t)$ are first estimated as in the prior model, and then for each $t = 66, 67, \ldots, 185$, $\hat{\boldsymbol{\mu}}_t$ is estimated as the expectation of a multivariate normal distribution parameter $(\tilde{\boldsymbol{\mu}}_t, 0.1\tilde{\boldsymbol{\Sigma}}_t)$ conditioned on the ranking obtained from that of the ten industry returns at time $t + 1$. This can be interpreted as a process to improve the quality of prior mean estimates by a one-step ahead ranking. Note that in practice the ranking would usually be obtained from another database of investor views. However, this approach affords an understanding of the potential benefit of the one-step ahead ranking methodology. As with the prior model, the portfolio returns $\hat{r}_t$, $t = 67, 68, \ldots, 186$ are computed for the rank constrained model, and finally the certainty equivalent returns $\hat{u}_i$ are computed for the 10 twelve-month periods.

The certainty equivalence returns for the ten periods obtained from the prior model and the rank constrained model are plotted against each other in Fig. 1. The certainty equivalence returns from the rank constrained model outperform those from the prior model consistently in every period. This analysis indicates the potential benefit of the rank constrained model and demonstrates the advantages in this area of study available from employing the recursive integration methodology discussed in this paper.

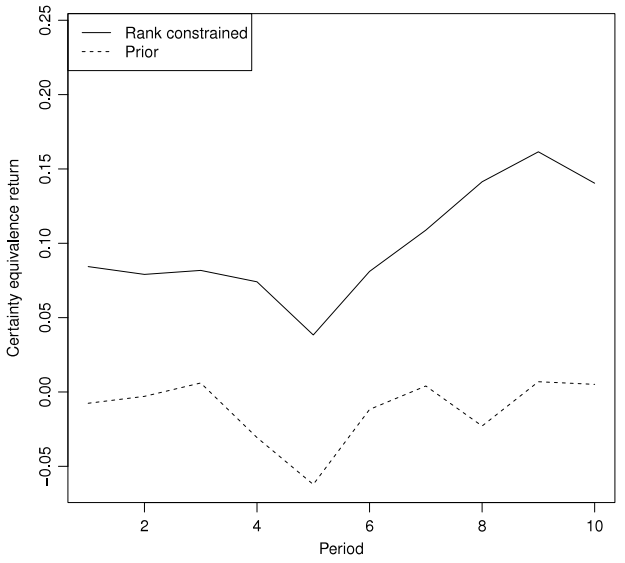


Fig. 1. Certainty equivalence returns of the 10 twelve-month period from the rank constrained model (solid line) and the prior model (dashed line). The returns from the rank constrained model consistently outperform those from the prior model.

5. Conclusion

This paper has addressed the situation where a set of variables has independent specified prior distributions, and where some information becomes available on the ordering of the variables. This is a common phenomenon which has been discussed and utilized in other statistical methodologies such as ranked set sampling. In this paper it is shown how updated distributions and moments of the variables can be calculated conditional on the knowledge provided by the ranking. It has been shown how the technique of recursive integration can be used to perform these calculations in a straightforward manner as a series of one-dimensional integral computations regardless of the number of variables.

For these particular problems of conditional probability and conditional expectation computations, it has been demonstrated that the errors in the numerators and the denominators can partially cancel each other, providing a self-correction mechanism that improves the accuracy of the recursive integration methodology.

The methodology presented in this paper has been implemented for a simple ordering of the variables. The methodology has also been generalized to variables with a multivariate normal distribution with a product correlation structure. In principle, the methodology can be extended to more complicated orderings such as an umbrella ordering or tree orderings, which are topics planned for future research.

Examples have been presented which illustrate how different kinds of rankings, such as reinforcing rankings and opposing rankings, can have different and substantial effects on the distributions, expectations, standard deviations, and correlations of the variables. This can be valuable information for practitioners, and the methodology presented in this paper allows this information to be obtained. Finally, the methodology has been applied to a decision problem in portfolio selection with ranking information, where it has been shown to provide a potential benefit. R code is available from the authors to replicate the tables and examples in this paper, and to implement the algorithms discussed

Acknowledgments

We would like to thank the reviewers whose comments have substantially improved this work. This work was supported by the Thailand Research Fund and Chulalongkorn University [RSA5780005].

References

Ali, A., Meilă, M., 2012. Experiments with Kemeny ranking: What works when? Math. Social Sci. 64, 28–40.
Arnold, B.C., Balakrishnan, N., Nagaraja, H.N., 1992. A First Course in Order Statistics. Wiley, New Jersey.
Chen, Z., Bai, Z., Sinha, B.K., 2004. Ranked Set Sampling. Springer, New York.
Chiarawongse, A., Kiatsupaibul, S., Tirapat, S., Van Roy, B., 2012. Portfolio selection with qualitative input. J. Bank. Finance 36, 489–496.
David, H.A., Nagaraja, H.N., 2003. Order Statistics, third ed. Wiley, New Jersey.
DeMiguel, V., Garlappi, L., Uppal, R., 2009. Optimal versus naive diversification: how inefficient is the $1/n$ portfolio strategy? Rev. Financ. Stud. 22, 1915–1953.
Gaur, A., Mahajan, K.K., Arora, S., 2012. New nonparametric tests for testing homogeneity of scale parameters against umbrella alternative. Statist. Probab. Lett. 82, 1681–1689.
Hans, C., Dunson, D.B., 2005. Bayesian inferences on umbrella orderings. Biometrics 64 (4), 1018–1026.
Harter, H.L., Balakrishnan, N., 1996. C.R.C. Handbook of Tables for the Use of Order Statistics in Estimation. CRC Press Inc., Boca Raton.

Hayter, A.J., 2006. Recursive integration methodologies with statistical applications. J. Statist. Plann. Inference 136, 2284–2296.
Hayter, A.J., Liu, W., 1996. A note on the calculation of $\mathrm{pr}(x_1 < \cdots < x_k)$. Amer. Statist. 50 (4), 365.
Kemeny, J.L., Snell, J.G., 1962. Mathematical Models in the Social Sciences. Blaisdell, New York.
Khachiyan, L.G., 1989. The problem of computing the volume of polytopes is NP-hard. Uspekhi Mat. Nauk 44 (3), 199–200.
Kiatsupaibul, S., Smith, R.L., Zabinsky, Z.B., 2011. An analysis of a variation of hit-and-run for uniform sampling from general regions. ACM Trans. Model. Comput. Simul. 21 (3), Article number 16.
Lovász, L., 1999. Hit-and-run mixes fast. Math. Program. 86, 443–461.
Lovász, L., Vempala, S., Fast algorithms for logconcave functions: Sampling, rounding, integration and optimization, in: Proc. of the 47th IEEE Symposium on Foundations of Computer Science, FOCS 06, 2006, pp. 57–68.
McIntyre, G.A., 1952. A method for unbiased selective sampling, using ranked sets. Aust. J. Agric. Res. 3, 385–390.
Milgrom, P., Roberts, J., 1994. Comparing equilibrium. Amer. Econ. Rev. 84 (3), 441–459.
Milgrom, P., Shannon, C., 1994. Monotone comparative statics. Econometrica 62 (1), 157–180.
Nakas, C.T., Alonzo, T.A., 2007. ROC graphs for assessing the ability of a diagnostic marker to detect three disease classes with an umbrella ordering. Biometrics 63, 603–609.
Patil, G.P., 2002. Ranked set sampling. In: Encyclopedia of Environmetrics. Vol. 3. John Wiley & Sons, Ltd., Chichester, pp. 1684–1690.
Singh, P., Liu, W., 2006. A test against an umbrella ororder alternative. Comput. Statist. Data Anal. 51, 1957–1964.
Smith, R.L., 1984. Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. Oper. Res. 32, 1296–1308.
Topkis, D.M., 1998. Supermodularity and Complementarity. Princeton University Press.
Wolfe, D.A., 2004. Ranked set sampling: an approach to more efficient data collection. Statist. Sci. 19, 636–643.
Young, P., 1995. Optimal voting rules. J. Econ. Perspect. 9, 51–64.

ภาคผนวก 2

บทความวิจัยปีที่ 2

Kiatsupaibul, S., A. J. Hayter and S. Somsong. 2017b. Confidence sets and confidence bands for a beta distribution with applications to credit risk management. *Insurance: Mathematics and Economics*. 75, 98-104.

Insurance: Mathematics and Economics 75 (2017) 98–104

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime

Confidence sets and confidence bands for a beta distribution with applications to credit risk management

Seksan Kiatsupaibul [a,*], Anthony J. Hayter [b], Sarunya Somsong [a]

[a] Department of Statistics, Chulalongkorn University, Bangkok, Thailand
[b] Department of Business Information and Analytics, University of Denver, Denver, USA

HIGHLIGHTS

- An exact confidence band for loss given default distribution in credit risk management is proposed.
- The approach based on a multiple comparison technique for a beta distribution.
- The resulting technique can be employed to rigorously stress test loss given default estimate with a limited data set.
- Estimating the loss given default of global events, the proposed methodology provides a sharp yet more conservative estimates when comparing with those provided by the regulator standard.

ARTICLE INFO

Article history:
Received November 2016
Received in revised form April 2017
Accepted 17 May 2017
Available online 26 May 2017

Keywords:
Credit risk
Loss given default
Beta distribution
Multiple comparison
Confidence band

ABSTRACT

Incorporating statistical multiple comparisons techniques with credit risk measurement, a new methodology is proposed to construct exact confidence sets and exact confidence bands for a beta distribution. This involves simultaneous inference on the two parameters of the beta distribution, based upon the inversion of Kolmogorov tests. Some monotonicity properties of the distribution function of the beta distribution are established which enable the derivation of an efficient algorithm for the implementation of the procedure. The methodology has important applications to financial risk management. Specifically, the analysis of loss given default (LGD) data are often modeled with a beta distribution. This new approach properly addresses model risk caused by inadequate sample sizes of LGD data, and can be used in conjunction with the standard recommendations provided by regulators to provide enhanced and more informative analyses.

1. Introduction

Practitioners in risk management necessarily employ a wide array of statistical estimation techniques in their day-to-day activities. However, statistical inferences which quantify the reliability of an adopted statistical model based upon limited data have long been downplayed. For example, a 99% value at risk is regularly computed and shown on a risk management report, even though its confidence interval has generally been ignored. With current advances in multiple comparisons techniques, risk managers should now become equipped with novel statistical inference tools that enable them to easily incorporate statistical inferences into their standard risk management procedures. In this article, we introduce these ideas through multiple comparisons on the beta distribution model with applications to credit risk measurement.

The beta distribution is an important probability distribution whose range is defined on the interval of real numbers between 0 and 1. It may be employed to describe the probabilistic behavior of system responses as a percentage, or to describe the rate of occurrence of an event, for example. Beta distributions also arise in measurements resulting from basic stochastic processes and order statistics. In addition, the beta distribution is the conjugate prior of the binomial likelihood (see, for example, Feller, 1971; Johnson et al., 1995).

The probability density function of the *standard beta distribution* is

$$\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$$

where $0 < x < 1$ and $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx$ is the beta function, which depends upon two positive shape parameters a and b. Therefore, inferences on this distribution require simultaneous inferences of these two parameters, which is the objective of this paper.

* Corresponding author.
E-mail address: seksan@cbs.chula.ac.th (S. Kiatsupaibul).

http://dx.doi.org/10.1016/j.insmatheco.2017.05.006
0167-6687/

Recent advances in the construction of simultaneous confidence bands for distribution functions have focused on developing exact bands (Frey, 2008; Kiatsupaibul and Hayter, 2015) rather than using the traditional asymptotic bands (Cheng and Iles, 1983; Bickel and Freedman, 1981; Bickel and Krieger, 1989). Parametric exact confidence bands for distributions with multiple parameters have been developed based on nonparametric statistics, such as Kolmogorov statistics. The Weibull distribution (Hayter and Kiatsupaibul, 2013) and the gamma distribution (Hayter and Kiatsupaibul, 2014) are two examples. Exact inferences are possible from the principle of inverse hypothesis testing to construct a confidence set for the parameters at a specified confidence level $1-\alpha$. Confidence bands for the distribution function then follow readily from a mapping of the confidence set into the distribution function space.

This paper shows how exact confidence sets and exact confidence bands can be constructed for the beta distribution based on this methodology. As part of this process, a monotonicity property of the beta distribution function with respect to its parameters is established. This property proves useful in the efficient construction of the confidence set for the parameters, and consequently for the construction of the confidence bands for the distribution.

Confidence bands for distribution functions have important applications to risk management. For example, when a Weibull distribution is adopted as a parametric model for a mortality distribution of interest to an insurer, its confidence band (Hayter and Kiatsupaibul, 2013) can be employed to measure the risk of the portfolio of the life insurance products. In addition, when the arrival times of abnormal internet connections are modeled by a gamma distribution, its confidence band (Hayter and Kiatsupaibul, 2014) provides information about the risk of the internet security system.

Furthermore, in the area of credit risk management the beta distribution is recommended by some standards as a model for the Loss Given Default (LGD). The LGD is a crucial factor in calculating the loss in an event of a credit default, along with the default rate and the exposure at default (see, for example, Gupton et al., 1997; Duffie and Singleton, 2003; Altman, 2008; Frontczak and Rostek, 2015; Wei and Yuan, 2016). However, since there is uncertainty in fitting the LGD model, financial institutions are recommended by the regulators to perform stress tests, and a particularly rigorous way of stress testing the LGD model is to use the confidence band of the model distribution function. Therefore, an important application of the confidence band methodology proposed in this paper can be found in this process of credit risk management, and an example is provided in this paper.

This paper is organized as follows. In Section 2 the proposed methodology is described by first deriving a monotonicity property of the distribution function as functions of its parameters. An algorithm to construct the exact confidence set and confidence band for the beta distribution is then provided. In Section 3 examples of the confidence band construction are given for both simulated data sets and a real data set. In the real data example it is shown how to construct a confidence band for the LGD distribution and a discussion is provided of its application in credit risk management. Finally, a summary is provided in Section 4.

2. Methodology

In this section the proposed methodology for the construction of confidence sets and confidence bands for a beta distribution is provided. The theoretical development is discussed in Section 2.1, and algorithms are outlined in Section 2.2.

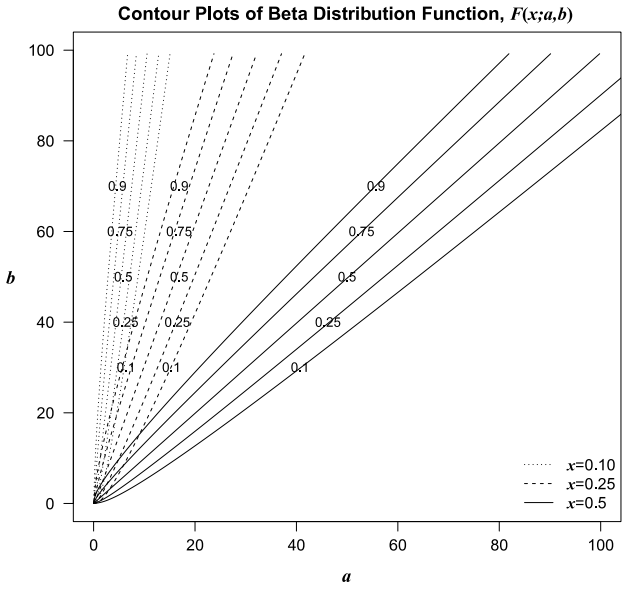


Fig. 1. Some contours of the distribution function $F(x; a, b)$.

2.1. Theoretical development

Consider a vector of n independent identically distributed random variables $\boldsymbol{X} = (X_1, \ldots, X_n)$ having the beta distribution function $F(x; a, b)$, with positive shape parameters a and b, given by

$$F(x; a, b) = \frac{\int_0^x f(t; a, b)\, dt}{\int_0^1 f(t; a, b)\, dt} \tag{1}$$

where

$$f(x; a, b) = x^{a-1}(1-x)^{b-1}.$$

Examples of the distribution function contours at various x on the parameter (a, b) plane are shown in Fig. 1, and notice that $F(x; a, b) = 1 - F(1-x; b, a)$. The expectation of this distribution is $a/(a+b)$, which remains unchanged if the parameters a and b are both scaled by the same quantity, and it can be seen that the contours are reasonably straight lines extending out from the origin $a = b = 0$.

Consider the acceptance set $A(a, b)$ defined by

$$A(a, b) = \left\{ \boldsymbol{X} : \sup_x |G_{\boldsymbol{X}}(x) - F(x; a, b)| \le d_{\alpha,n} \right\}, \tag{2}$$

where $G_{\boldsymbol{X}}(x)$ is the empirical cumulative distribution function of a sample of $\boldsymbol{X}$ and $d_{\alpha,n}$ is the Kolmogorov critical point. It follows from Kolmogorov's test that there is a probability of exactly $1-\alpha$ that the observed $\boldsymbol{X}$ will fall within the acceptance set of the true parameter values. Moreover, Eq. (2) can be written as

$$\frac{i}{n} - d_{\alpha,n} \le F(X_{(i)}; a, b) \le \frac{i-1}{n} + d_{\alpha,n}, \quad i = 1, \ldots, n, \tag{3}$$

where $X_{(1)} \le \cdots \le X_{(n)}$ are the ordered values of $X_1, \ldots, X_n$. Consequently, for an observed value $\boldsymbol{X}$, a $100(1-\alpha)\%$ confidence set $K_\alpha(\boldsymbol{X})$ for the beta parameters is the set of pairs (a, b) for which (3) is satisfied.

It should be noted that the confidence set $K_\alpha(\boldsymbol{X})$ may be empty. This occurs when there are no values of a and b that satisfy Eq. (3), and it alerts the experimenter to the fact that the data should not be modeled with a beta distribution. In other words, the confidence set $K_\alpha(\boldsymbol{X})$ is empty if there is no beta distribution consistent with the data according to Kolmogorov's test.

The following theorem and corollary are useful for the construction of the algorithms in Section 2.2 for finding the solutions to Eq. (3). Specifically, Theorem 1 states that the beta distribution function at a given quantile, also known as *the regularized incomplete beta function* or *the incomplete beta function ratio* (see Johnson et al., 1995; Johnson et al., 2005), is a monotone function in each of the parameters. This implies that, with one parameter fixed, the confidence set with respect to the other parameter is a connected interval. This fact helps in the derivation of an efficient algorithm to construct the desired confidence set of the parameters.

Theorem 1. *Let $F(x; a, b)$ be the cumulative distribution function of the beta distribution given in* Eq. (1)*. Then for any $0 < x < 1$ and fixed $b > 0$, $g(a) = F(x; a, b)$ is a continuous nonincreasing function in $a > 0$. In addition, for any $0 < x < 1$ and fixed $a > 0$, $h(b) = F(x; a, b)$ is a continuous nondecreasing function in $b > 0$. Furthermore,*

$$\lim_{a\to\infty} g(a) = 0 = \lim_{b\to 0} h(b) \quad \text{and} \quad \lim_{a\to 0} g(a) = 1 = \lim_{b\to\infty} h(b).$$

Proof. For fixed $b > 0$, it is first shown that $g(a_1) \geq g(a_2)$ for any $0 < a_1 < a_2$. Observe that

$$s(x) = \frac{f(x; a_1, b)}{f(x; a_2, b)} = x^{(a_1 - a_2)}$$

is decreasing in x. Now define

$$u(x) = \frac{\int_0^x f(t; a_1, b)\, dt}{\int_0^x f(t; a_2, b)\, dt} = \frac{\int_0^x f(t; a_2, b) s(t)\, dt}{\int_0^x f(t; a_2, b)\, dt}$$

which is a weighted average of the decreasing function $s(x)$. Therefore, for $0 < x_1 < x_2$,

$$u(x_1) \geq s(x_1) \geq \frac{\int_{x_1}^{x_2} f(t; a_2, b) s(t)\, dt}{\int_{x_1}^{x_2} f(t; a_2, b)\, dt}.$$

Observe further that

$$u(x_2) = w \cdot u(x_1) + (1 - w) \cdot \left(\frac{\int_{x_1}^{x_2} f(t; a_2, b) s(t)\, dt}{\int_{x_1}^{x_2} f(t; a_2, b)\, dt} \right),$$

where

$$0 \leq w = \frac{\int_0^{x_1} f(t; a_2, b)\, dt}{\int_0^{x_2} f(t; a_2, b)\, dt} \leq 1,$$

so that $u(x_2)$ is a convex combination of $u(x_1)$ and the smaller term. Hence, $u(x_1) \geq u(x_2)$ which implies that for $0 \leq x \leq 1$,

$$\frac{\int_0^x f(t; a_1, b)\, dt}{\int_0^x f(t; a_2, b)\, dt} = u(x) \geq u(1) = \frac{\int_0^1 f(t; a_1, b)\, dt}{\int_0^1 f(t; a_2, b)\, dt}.$$

Multiplying both sides of this inequality by $\left(\int_0^x f(t; a_2, b)\, dt \Big/ \int_0^1 f(t; a_1, b)\, dt \right)$ gives the required result $g(a_1) \geq g(a_2)$. Furthermore, it is clear that g is continuous in $a > 0$ since f is continuous in $a > 0$.

It is now shown that $\lim_{a\to\infty} g(a) = 0$ for each $0 < x < 1$. For a fixed $x \in (0, 1)$,

$$\begin{aligned} g(a) &= \frac{\int_0^x t^{a-1}(1-t)^{b-1}\, dt}{\int_0^1 t^{a-1}(1-t)^{b-1}\, dt} \\ &\leq \frac{\int_0^x t^{a-1}(1-t)^{b-1}\, dt}{\int_x^1 t^{a-1}(1-t)^{b-1}\, dt} \\ &\leq \frac{\max\{1, (1-x)^{b-1}\} \int_0^x t^{a-1}\, dt}{\min\{1, (1-x)^{b-1}\} \int_x^1 t^{a-1}\, dt} \\ &= \frac{\max\{1, (1-x)^{b-1}\} x^a}{\min\{1, (1-x)^{b-1}\}(1-x^a)}. \end{aligned}$$

Since $0 < x < 1$, the last term goes to 0 so that $g(a) \to 0$ as $a \to \infty$.

Now it is shown that $\lim_{a\to 0} g(a) = 1$ for each $0 < x < 1$. Define $\bar{g}(a) = 1 - g(a)$. Then for a fixed $x \in (0, 1)$,

$$\begin{aligned} \bar{g}(a) &= \frac{\int_x^1 t^{a-1}(1-t)^{b-1}\, dt}{\int_0^1 t^{a-1}(1-t)^{b-1}\, dt} \\ &\leq \frac{\int_x^1 t^{a-1}(1-t)^{b-1}\, dt}{\int_0^x t^{a-1}(1-t)^{b-1}\, dt} \\ &\leq \frac{\max\{1, (1-x)^{b-1}\} \int_x^1 t^{a-1}\, dt}{\min\{1, (1-x)^{b-1}\} \int_0^x t^{a-1}\, dt} \\ &= \frac{\max\{1, (1-x)^{b-1}\}(1-x^a)}{\min\{1, (1-x)^{b-1}\} x^a}. \end{aligned}$$

Now as $a \to 0$ the last term goes to 0, and hence $\bar{g}(a) \to 0$ and $g(a) \to 1$.

The analogous properties of $h(b)$ can be obtained similarly, or by recognizing that $F(x; a, b) = 1 - F(1 - x; b, a)$. □

The following corollary is useful for finding the solutions to Eq. (3).

Corollary 1. *For $0 < X_{(1)} \leq \cdots \leq X_{(n)} < 1$ and fixed $b > 0$, there exists an interval $[l_a, r_a]$ with $l_a > 0$ such that a satisfies* Eq. (3) *iff $a \in [l_a, r_a]$. In the same way, for a fixed $a > 0$ there exists an interval $[l_b, r_b]$ with $l_b > 0$ such that b satisfies* (3) *iff $b \in [l_b, r_b]$.*

Proof. For fixed $b > 0$ and fixed $i \in \{1, \ldots, n\}$, by Theorem 1 there exists $0 < l_{a,i} < r_{a,i}$ such that a satisfies

$$\frac{i}{n} - d_{\alpha,n} \leq F(X_{(i)}; a, b) \leq \frac{i-1}{n} + d_{\alpha,n} \tag{4}$$

iff $a \in [l_{a,i}, r_{a,i}]$. In fact, the interval $[l_{a,i}, r_{a,i}]$ is the inverse mapping of the interval $[\frac{i}{n} - d_{\alpha,n}, \frac{i-1}{n} + d_{\alpha,n}]$ under $g(a) = F(x_i; a, b)$, which is a bounded continuous nonincreasing function. Then

$$[l_a, r_a] = \bigcap_{i=1}^{n} [l_{a,i}, r_{a,i}]$$

which may be empty. In fact, $l_a = \max_i\{l_{a,i}\}$ and $r_a = \min_i\{r_{a,i}\}$ when the first of these is smaller than the second (and the interval is empty otherwise). For fixed $a > 0$ the interval $[l_b, r_b]$ can be found in a similar way. □

2.2. Algorithms

This section contains an algorithm for obtaining the confidence set from Eq. (3). R-code is available from the authors to implement this algorithm.

Corollary 1 implies that the cross-section of the desired confidence set $K_\alpha(\boldsymbol{X})$ at a given b is a connected interval of a (and similarly, at a given a is a connected interval of b). Therefore, the confidence set construction can be reduced to finding these two interval end points at each value of b, which can easily be performed by, for example, the bisectioning method. Algorithm 1 illustrates a possible version of the confidence set construction procedure.

In Algorithm 1, define the left and the right cross-sectional end points at b as l_b and r_b as in Corollary 2.2, so that

$$l_b = \arg\min_a\{(a, b) : (a, b) \text{ satisfies Eq. (3)}\},$$

$$r_b = \arg\max_a\{(a, b) : (a, b) \text{ satisfies Eq. (3)}\}.$$

Algorithm 1 Confidence Set Construction

Require:

A data set of size $0 < X_{(1)} \leq \cdots \leq X_{(n)} < 1$.
An initial pair of parameters (a_0, b_0) that satisfies equation (3).
Confidence level, $100(1-\alpha)\%$; precision parameter, $\Delta > 0$; cross-sectional tolerance, $\epsilon > 0$; maximum of parameter b, denoted by $b_{\max}$.

Set $j = 0$. Given b_j, find l_{b_j} and r_{b_j}.
while $r_{b_j} - l_{b_j} \geq \epsilon$ and $b_j + \Delta \leq b_{\max}$ **do**
  Set $j \leftarrow j + 1$. Let $b_j = b_{j-1} + \Delta$. Find l_{b_j} and r_{b_j}.
end while
Let $j^+ = j$, set $j = -1$ and $b_{-1} = b_0 - \Delta$. Given b_j, find l_{b_j} and r_{b_j}.
while $r_{b_j} - l_{b_j} \geq \epsilon$ and $b_j - \Delta > 0$ **do**
  Set $j \leftarrow j - 1$. Let $b_j = b_{j-1} - \Delta$. Find l_{b_j} and r_{b_j}.
end while
return The confidence set, $K_\alpha(\boldsymbol{X})$, defined by the sequences of left end points and right end points:

$$\{(l_{j^-}, b_{j^-}), (l_{j^-+1}, b_{j^-+1}), \ldots, (l_{j^+}, b_{j^+})\},$$

$$\{(r_{j^-}, b_{j^-}), (r_{j^-+1}, b_{j^-+1}), \ldots, (r_{j^+}, b_{j^+})\}.$$

Once the confidence set $K_\alpha(\boldsymbol{X})$ has been constructed, the confidence band of the distribution function is the mapping into the distribution function space of all distribution functions with respect to all parameter pairs in $K_\alpha(\boldsymbol{X})$. Specifically, for each x the confidence band of the distribution function at x has a lower bound $F_l(x)$ and an upper bound $F_u(x)$, where

$$F_l(x) = \inf\{F(x; a, b) : (a, b) \in K_\alpha(\boldsymbol{X})\},$$

and

$$F_u(x) = \sup\{F(x; a, b) : (a, b) \in K_\alpha(\boldsymbol{X})\}.$$

For each x, finding $F_l(x)$ and $F_u(x)$ is ostensibly an optimization problem in two dimensions. However, by the monotonicity of the distribution function with respect to the beta parameters given in Theorem 1, it is known that $F_l(x)$ and $F_u(x)$ can be found from the boundary of $K_\alpha(\boldsymbol{X})$. Therefore, once the confidence set has been constructed, the confidence band construction entails only a one dimensional optimization problem.

It is sometimes more convenient to construct the confidence band of the distribution function from the quantile function. In this case, for each probability $0 < q < 1$, the lower bound, $F_l^{-1}(q)$ and the upper bound $F_u^{-1}(q)$ of the quantile function are

$$F_l^{-1}(q) = \inf\{F^{-1}(x; a, b) : (a, b) \in K_\alpha(\boldsymbol{X})\},$$

and

$$F_u^{-1}(q) = \sup\{F^{-1}(x; a, b) : (a, b) \in K_\alpha(\boldsymbol{X})\}.$$

Since F_l^{-1} and F_u^{-1} form the same confidence band as those from F_l and F_u, the values of $F_l^{-1}(q)$ and $F_u^{-1}(q)$ can also be found on the boundary of $K_\alpha(\boldsymbol{X})$. Notice that since the confidence set has an exact confidence level of $1-\alpha$, the confidence bands also have an exact confidence level of $1-\alpha$.

3. Examples

In this section examples of the confidence set and confidence band construction are given simulated data sets in Section 3.1, and for a real data set in Section 3.2. In the real data example it is shown how to construct a confidence band for the Loss Given Default (LGD) distribution and a discussion is provided of its application in credit risk management.

Table 1
The simultaneous 95% confidence lower bounds and upper bounds for some key quantities of the beta distribution constructed from the simulated data sets of sizes 25, 50, 75 and 100 with parameters ($a = 2$, $b = 5$).

Parameter	True value	Size 25		Size 50	
		LCB	UCB	LCB	UCB
a	2.000	0.504	30.766	0.932	15.518
b	5.000	1.213	107.213	2.474	47.844
$a/(a+b)$ (mean)	0.286	0.179	0.296	0.224	0.296
$F^{-1}(0.01)$	0.027	0.000	0.147	0.003	0.133
$F^{-1}(0.25)$	0.161	0.051	0.211	0.097	0.215
$F^{-1}(0.50)$ (median)	0.264	0.159	0.279	0.200	0.286
$F^{-1}(0.75)$	0.389	0.237	0.489	0.273	0.416
$F^{-1}(0.99)$	0.706	0.311	0.959	0.380	0.838
Parameter	True value	Size 75		Size 100	
		LCB	UCB	LCB	UCB
a	2.000	1.042	8.730	1.150	6.171
b	5.000	2.646	25.126	2.863	17.513
$a/(a+b)$ (mean)	0.296	0.248	0.312	0.250	0.308
$F^{-1}(0.01)$	0.027	0.005	0.109	0.007	0.090
$F^{-1}(0.25)$	0.161	0.111	0.224	0.121	0.216
$F^{-1}(0.50)$ (median)	0.264	0.231	0.299	0.239	0.297
$F^{-1}(0.75)$	0.389	0.300	0.424	0.312	0.424
$F^{-1}(0.99)$	0.706	0.447	0.830	0.489	0.816

3.1. Simulated data

Observations were simulated from a beta distribution with parameters $a = 2$ and $b = 5$ and with sample sizes n equal to 25, 50, 75 and 100. In each case the methodology described in Section 2 was used to construct a 95% confidence set for the parameter pair (a, b), which are shown in Fig. 2. These confidence sets were then used to form 95% confidence bands for the cumulative distribution function, which are shown in Fig. 3. Table 1 contains confidence bounds for some key quantities of the beta distribution which are obtained from the confidence sets in Fig. 2.

It can be seen from Fig. 2 that the confidence sets possess a needle shape, which is not surprising considering the contour plots shown in Fig. 2, and the fact that the expectation of the distribution is $a/(a+b)$, which remains unchanged if the parameters a and b are both scaled by the same quantity, as discussed at the beginning of Section 2.1. As expected, the confidence sets become dramatically smaller when the sample size becomes larger, as shown by the areas given in Fig. 2. Each confidence set contains the true parameter values $a = 2$ and $b = 5$.

It can be seen from Fig. 3 that, as expected, the confidence bands becomes narrower when the sample size grows larger. In all cases the band is widest at high quantiles, narrowest at mid quantiles, and slightly wider again at low quantiles. Observe that even though the confidence set areas in Fig. 2 become much smaller when the sample size increases, there is not such a dramatic change in the confidence band areas. The true distribution function with parameter values $a = 2$ and $b = 5$ is shown by the dashed lines in Fig. 3, and in all cases it stays within the confidence bands (since the true parameter values are contained within the confidence sets in Fig. 2). An estimated distribution function corresponding to the maximum likelihood estimates of the parameters is also shown by the dotted lines in Fig. 3. This estimate can be quite far from the true distribution function, although it also stays within the confidence bands.

3.2. Loss given default (LGD) data

In credit risk management the loss from a default to a loan can be estimated by the following simple relationship

$$\text{Expected Loss} = \text{PD} \times \text{LGD} \times \text{Exposure}, \tag{5}$$

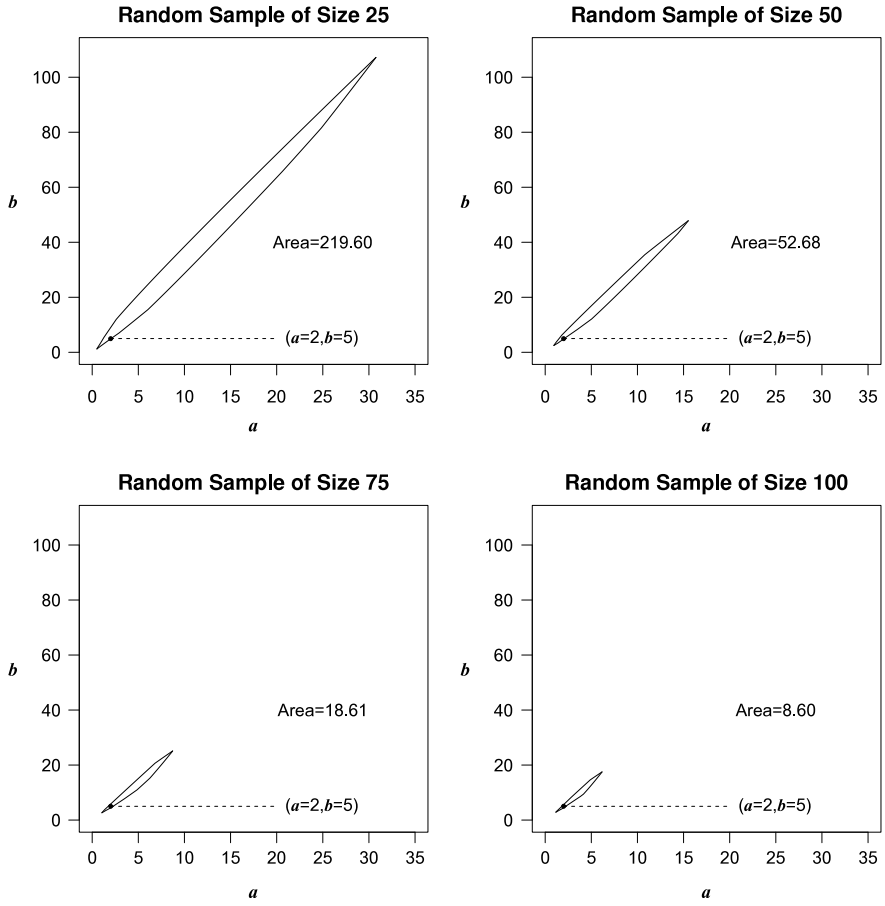


Fig. 2. The exact 95% confidence sets for the beta distribution parameters (a, b) constructed from simulated data sets of sizes 25, 50, 75 and 100 with parameters $(a = 2, b = 5)$.

where PD is the probability of default, LGD is the loss given default, and Exposure is the outstanding value of the loan at the time of the estimation (see, for example, Duffie and Singleton, 2003; Altman, 2008). In practice, the LGD is often modeled as a random quantity, and its expectation is substituted into Eq. (5). To guard against risk in economic downturns, the BIS Basel II (2004) financial regulations suggested that financial institutions perform stress tests on their LGD. Moreover, the Federal Reserve System of the United States recommends that a downturn LGD of the form

$$\text{LGD in Downturn} = 0.08 + 0.92E[\text{LGD}]$$

is employed where $E[LGD]$ is the expected loss given default (Altman, 2008).

J.P. Morgan's *CreditMetrics*^{TM} recommends a beta distribution for LGD (Gupton et al., 1997), which provides a motivation for the work in this paper. This recommendation can provide more information about an LGD other than merely its expected value. With a distribution specified for the LGD, risk managers are equipped with a more powerful tool to assess and manage risk in different economic climates. Of course, the distribution of the LGD cannot be known with certainty, and so the methodology developed in this paper can be used to make inferences about the LGD.

Neither the downturn LGD suggested by the US Federal Reserve System nor the beta model suggested by *CreditMetrics*^{TM} account for the model risk caused by an inadequate sample size. Therefore, it is proposed that an upper confidence band (UCB) fitted to a data set of LGD values experienced by financial institutions can be employed in conjunction with the recommendations already in place. This will be a rigorous stress test model for credit risk management purposes because this confidence band aggregates the information regarding the recovery risk and the model risk. In addition, it is a distribution function in its own right that can conveniently be applied to a wide range of risk calculations.

As a demonstration of the new suggested approach using the methodology developed in this paper, Moody's Default & Recovery Database was used to obtain a data set of defaults during the 10-year period from 2006 to 2015 whose recovery rate was greater than zero and less than one. This data set contains 249 events. The recovery rate of an event was computed from the weighted average of the discounted settlements corresponding to the instruments of the event. The LGD was then calculated as one minus the recovery rate.

Confidence sets and confidence bands at 95% and 99% confidence levels based on the proposed methodology were constructed from the data set of 249 LGD events, and they are shown in Fig. 4. Some key quantities obtained from the confidence set and the confidence bands are given in Table 2. Notice that the confidence sets in Fig. 4 is not empty, which confirms that this data set of LGD events can be modeled with a beta distribution.

From Fig. 4, it is natural that the 99% confidence set is larger than and covers the 95% confidence set. As a result, the 99% confidence band is also wider than the 95% confidence band. Observe that the area of a 99% confidence band is $0.1520/0.1212 \approx 1.25$ times that of 95% confidence band. This ratio can be considered as the average ratio between the 99% and 95% confidence intervals of a quantile of the estimated beta distribution. Recall that the width of a 99% confidence interval formed by a normal distribution is $z_{0.005}/z_{0.025} = 1.31$ times that of the corresponding 95% confidence interval. Therefore, the ratio from the newly proposed methodology is not very far from the ratio from a regular normal approximation (1.25 versus 1.31). However, in this case, the new method

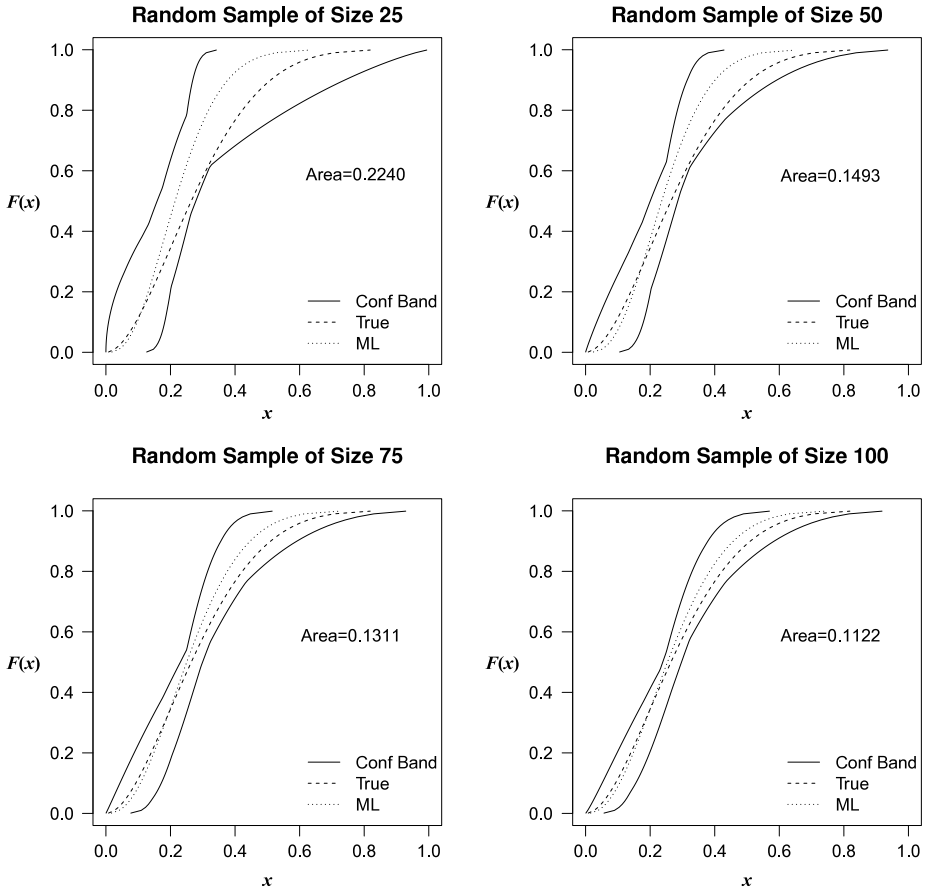


Fig. 3. The exact 95% confidence bands for the cumulative distribution function constructed from simulated data sets of sizes 25, 50, 75 and 100 with parameters $(a = 2, b = 5)$.

Table 2
Some key quantities of the beta distribution fitted to the 249 LGD events. The point estimates are computed using the maximum likelihood estimates of the parameters. The lower confidence bounds (LCB) and the upper confidence bounds (UCB), at 95% and 99% confidence levels, are computed using the new proposed methodology.

Parameter	Point estimates	95% Conf. Interval		99% Conf. Interval	
	Max likelihood	LCB	UCB	LCB	UCB
a	1.1611	0.7487	2.0453	0.6695	2.3580
b	1.3215	0.8665	2.4215	0.7715	2.7915
$a/(a+b)$ (mean)	0.4677	0.4211	0.5057	0.4095	0.5173
Downturn LGD[a]	0.5103	0.4674	0.5453	0.4568	0.5559
$F^{-1}(0.01)$	0.0146	0.0025	0.0533	0.0014	0.0672
$F^{-1}(0.25)$	0.2417	0.1801	0.3113	0.1642	0.3286
$F^{-1}(0.50)$ (median)	0.4579	0.4002	0.5074	0.3859	0.5222
$F^{-1}(0.75)$	0.6864	0.6065	0.7436	0.5917	0.7622
$F^{-1}(0.99)$	0.9732	0.9087	0.9935	0.8900	0.9961

[a] Downturn LGD is defined as $0.08 + 0.92a/(a+b)$.

comparatively provides a little more aggressive 99% confidence interval.

In Table 2 the point estimates can be viewed as standard *CreditMetrics*^{TM} estimates. The point estimate of the mean 0.4677 is computed from the mean formula of the beta distribution, and is almost the same as the sample mean 0.4634. Consequently, the downturn LGD suggested by the US Federal Reserve System would become slightly lower (0.5064) if the mean was replaced by the sample mean.

The UCB column contains the new estimates based on the proposed upper confidence band. The mean in the UCB column is computed from the maximum of $a/(a+b)$ over all pairs (a, b) in the confidence set $K_\alpha(\boldsymbol{X})$ shown in the left panel of Fig. 4. It is then substituted into the downturn LGD formula to obtain the UCB of the downturn LGD. The values for the quantiles in the UCB column are taken from the upper confidence band shown in the right panel of Fig. 4.

The UCB values are naturally larger than the point estimates, with the discrepancies reflecting the model risk caused by the finiteness of the sample size. This implies that the UCB approach is more conservative, and hence more appropriate for a comprehensive risk management program that aims to take into consideration model risk. It is therefore proposed that this new UCB approach be employed in conjunction with the methodology developed in this paper. The discrepancies between the UCB estimates and the point estimates will become smaller as the analysis is based on larger data sets.

4. Summary

A new methodology has been proposed to construct an exact confidence set and exact confidence bands for a beta distribution. This involves simultaneous inference on the two parameters of the beta distribution, based upon the inversion of Kolmogorov tests.

It has been shown that the distribution function of the beta distribution is a monotone function with respect to either of its parameters, and this has enabled the derivation of an efficient algorithm for the confidence set and confidence band constructions. The methodology has been demonstrated with simulated data sets of different sample sizes.

Moreover, the methodology has been applied to an important problem in financial risk management. For the analysis of loss

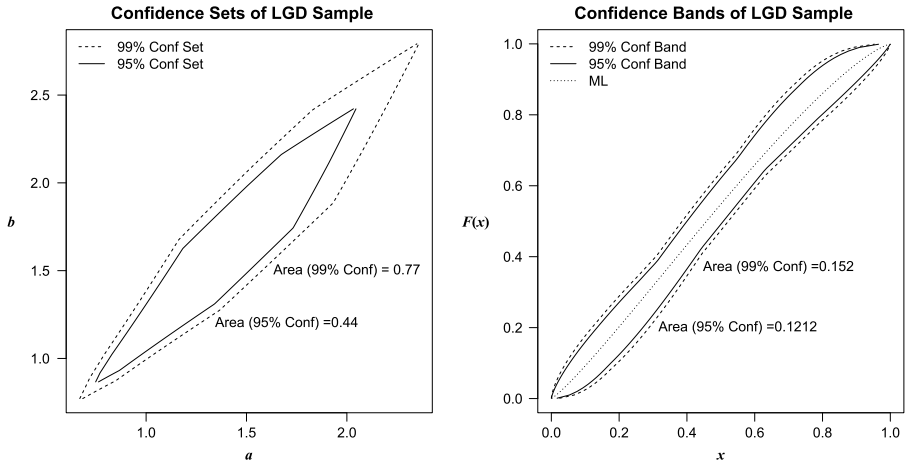


Fig. 4. The exact 95% and 99% confidence sets and confidence bands for the beta distribution fitted to the data set of 249 LGD events during the period 2006–2015 obtained from Moody's Default & Recovery Database.

given default (LGD) data, it has been proposed that the methodology be employed to calculate upper confidence bounds (UCB) for the quantities of interest. This new approach properly addresses model risk caused by inadequate sample sizes of LGD data, and can be used in conjunction with the standard recommendations provided by regulators to provide enhanced and more conservative analyses.

Acknowledgments

This work was supported by the Thailand Research Fund and Chulalongkorn University [RSA5780005].

References

Altman, E.I., 2008. Default recovery rates and lgd in credit risk modelling and practice: An updated review of the literature and empirical evidence. In: Jones, S., Hensher, D.A. (Eds.), Advances in Credit Risk Modelling and Corporate Bankruptcy Prediction. Cambridge University Press, Cambridge, pp. 175–206.

Bickel, P.J., Freedman, D.A., 1981. Some asymptotic theory of the bootstrap. Ann. Statist. 9, 1196–1217.

Bickel, P.J., Krieger, A.M., 1989. Confidence bands for a distribution function using the bootstrap. J. Amer. Statist. Assoc. 84 (405), 95–100.

Cheng, R., Iles, T., 1983. Confidence bands for cumulative distribution function of continuous random variables. Technometrics 25 (1), 77–86.

Duffie, D., Singleton, K.J., 2003. Credit Risk: Pricing, Measurement, and Management. Princeton University Press, Princeton.

Feller, W., 1971. An Introduction To Probability Theory and Its Applications, Vol. 2, second ed. Wiley, New York.

Frey, J.C., 2008. Optimal distribution-free confidence bands for a distribution function. J. Statist. Plann. Inference 138 (10), 3086–3098.

Frontczak, R., Rostek, S., 2015. Modeling loss given default with stochastic collateral. Econ. Modell. 44, 162–170.

Gupton, G., Finger, C., Bhatia, M., 1997. Credit Metrics™ Technical Document. JPMorgan & Co., New York.

Hayter, A.J., Kiatsupaibul, S., 2013. Exact inferences for a weibull model. Qual. Eng. 25 (2), 175–180.

Hayter, A.J., Kiatsupaibul, S., 2014. Exact inferences for a gamma distribution. Qual. Technol. 46 (2), 140–149.

Johnson, N.L., Kemp, A.W., Kotz, S., 2005. Univariate Discrete Distributions, third ed.. Wiley, New York.

Johnson, N.L., Kotz, S., Balakrishnan, N., 1995. Continuous Univariate Distributions, Vol. 2, second ed. Wiley, New York.

Kiatsupaibul, S., Hayter, A.J., 2015. Recursive confidence band construction for an unknown distribution function. Biom. J. 57 (1), 39–51.

Wei, L., Yuan, Z., 2016. The loss given default of a low-default portfolio with weak contagion. Insurance Math. Econom. 66, 113–126.

ภาคผนวก 3

บทความวิจัยปีที่ 3

Hayter, A.J., P. Yang and S. Kiatsupaibul. 2017. Win-probabilities for comparing two Weibull distributions. *Quality Technology and Quantitative Management*. 14, 1-18.

QUALITY TECHNOLOGY & QUANTITATIVE MANAGEMENT, 2017
VOL. 14, NO. 1, 1–18
http://dx.doi.org/10.1080/16843703.2016.1189197

Taylor & Francis
Taylor & Francis Group

Win-probabilities for comparing two Weibull distributions

A. J. Hayter[a1], P. Yang[b] and S. Kiatsupaibul[c]

[a]Department of Business Information and Analytics, University of Denver, Denver, CO, USA; [b]Department of Statistics, The Chinese University of Hong Kong, Hong Kong, China; [c]Department of Statistics, Chulalongkorn University, Bangkok, Thailand

ABSTRACT

This paper considers the problem of comparing two processes or treatments which are each modelled with a Weibull distribution. Win-probabilities are considered, which compare potential single future observations from each of the two treatments. This information can be useful in helping decide which of the two treatments to adopt, and can be combined with other factors relevant to a practitioner such as the availabilities, costs and side-effects of the two treatments. A methodology employing joint confidence sets is developed which not only allows estimation and confidence interval construction for the win-probabilities, but at the same guaranteed confidence level also tests whether Weibull distributions are appropriate for the data, identifies any common Weibull distributions for the two processes and also provides individual inferences for the two Weibull distributions. Examples are given to illustrate the implementation and application of this methodology, for which R computer code is available from the authors. This methodology can be extended to different models such as other two-parameter and three-parameter Weibull models, and to the comparison of three or more Weibull distributions.

ARTICLE HISTORY
Accepted 10 May 2016
Published online 19 July 2016

KEYWORDS
Weibull distribution; failure times; win-probabilities; joint confidence sets; confidence intervals; two-sample problem

1. Introduction

Let

$$f(x; a, \lambda) = \frac{a}{\lambda}\left(\frac{x}{\lambda}\right)^{a-1} e^{-(x/\lambda)^a}$$

be the probability density function of a Weibull distribution with parameters a and λ, with a cumulative distribution function

$$F(x; a, \lambda) = 1 - e^{-(x/\lambda)^a},$$

for $x \geq 0$, with $a > 0$ and $\lambda > 0$. This distribution has received considerable attention in the reliability literature, and most standard approaches to making inferences with this model involve graphical methods or approximate theoretical methods (see, for example, Abernethy, 2006; Lawless, 2003; Rinne, 2008).

Consider the two sample problem with independent data X_{1i}, $1 \leq i \leq n_1$, from a Weibull distribution with parameters a_1 and λ_1, and independent data X_{2i}, $1 \leq i \leq n_2$,

CONTACT S. Kiatsupaibul seksan@cbs.chula.ac.th
[1]Department of Statistics, Chulalongkorn University, Bangkok, Thailand.

from a Weibull distribution with parameters a_2 and λ_2, with all four parameters being unknown. The comparison of these two Weibull distributions based on the two samples is a difficult problem, and the methodologies available in the literature generally rely on asymptotic approximate arguments with some assumptions about the parameters.

For example, Schafer and Sheffield (1976) discuss how to test the equality of the two scale parameters λ_1 and λ_2 under the assumption that the two shape parameters a_1 and a_2 are equal. More recently, Hudak and Tiryakioglu (2011) show how to compare the shape parameters of two Weibull distributions and discuss applications to the fracture properties of ceramics and metals. Also, Louzada-Neto, Bolfarine, and Rodrigues (2002) provide a good motivation for the related problem of comparing two Weibull regression models with respect to assessing the reliability of manufactured items, and provide a Bayesian solution for accelerated data. Finally, Parsi, Ganjali, and Farsipour (2011) provide approximate methods for this two sample problem based on the asymptotic normality of the maximum likelihood estimators and using bootstrap methods when there is Type-II progressive censoring.

In this paper, the comparison of the two Weibull distributions is based upon the construction of joint confidence sets for the parameters, and the consideration of win-probabilities which are defined as follows. Let X_1^* be a potential future observation from the Weibull distribution with parameters a_1 and λ_1, and let X_2^* be a potential future observation from the Weibull distribution with parameters a_2 and λ_2. Then for $\rho > 0$ a win-probability is defined as

$$W_2(\rho) = P(X_2^* \geq \rho X_1^*) = \int_0^\infty f(x; a_1, \lambda_1)(1 - F(\rho x; a_2, \lambda_2))\mathrm{d}x$$
$$= \int_0^\infty \frac{a_1}{\lambda_1}\left(\frac{x}{\lambda_1}\right)^{a_1-1} e^{-(x/\lambda_1)^{a_1}} e^{-(\rho x/\lambda_2)^{a_2}}\mathrm{d}x.$$

For $\tau = \rho\lambda_1/\lambda_2$ this is

$$W_2(\rho) = \int_0^\infty a_1 x^{a_1-1} e^{-(x^{a_1}+(\tau x)^{a_2})}\mathrm{d}x. \tag{1}$$

Thus, if the two Weibull distributions represent two processes or treatments, and if larger observations are better, then the win-probability $W_2(1)$ provides the probability that a potential future observation from the second treatment will be at least as large as a potential future observation from the first treatment. The values of $W_2(\rho)$ for other values of ρ also provide further information on how much better the observation from the second treatment will be. Also, it is clear that

$$W_1(\rho) = P(X_1^* \geq \rho X_2^*) = 1 - W_2(1/\rho)$$

and if small observations are better then these win-probabilities can similarly be used to compare the two distributions.

The information provided by the win-probabilities can be of direct use in helping decide which of the two treatments to adopt when a practitioner has a choice between the two treatments, and it can be combined with other factors relevant to a practitioner such as the availabilities, costs and side-effects of the two treatments. Discussions of win-probabilities

can also be found in Hayter (2013) for normally distributed data, in Wiwatwattana, Hayter, and Kiatsupaibul (2015) for binomial data, in Hayter (in press) for Poisson data, and in Hayter (2012) for regression models.

Standard approaches to two sample problems generally involve testing the equality of parameters and comparing expectations and quantiles. This can be useful from a policy perspective in the sense that allocating all future observations to the treatment with the largest expectation, say, will guarantee the largest long run average for the future observations. However, as has been discussed, there are no direct procedures available to apply these standard approaches to the situation of two Weibull distributions without resorting to asymptotic arguments and assumptions.

The use of win-probabilities provides a way around this problem, and it also provides more pertinent information from an individual perspective rather than from a policy perspective. This is because an individual will have a potential outcome X_1^* if the first treatment is undertaken, and a potential outcome X_2^* if the second treatment is undertaken. The win-probabilities then provide direct information to the individual concerning what might happen if either treatment is taken, and this information can be combined with other considerations such as the costs, availabilities and side-effects of the two treatments which may be specific to that individual for that particular decision.

Thus, even if it is possible to establish that the first treatment has a larger expectation than the second treatment, say, with larger observations being better, for a particular individual it may be advantageous to take the second treatment if it is more readily available, has a substantially lower cost or if its detrimental side-effects are less. If it is found that $W_2(\rho)$ is not too small for certain values ρ of interest, then it may be judged that the 'penalty' for taking the second treatment due to its smaller expectation is not too severe, and so it may be considered to be the better decision when all aspects are taken into account.

These win-probabilities can be used in any of the reliability areas where the Weibull distribution is typically adopted. Thus, the two treatments may correspond to two manufacturing processes, say, or to different carbon fibres as illustrated in Example 1 in Section 4.2. Also, an application of win-probabilities to compare the costs of two processes is provided in Example 2 in Section 4.2. The win-probabilities can be particularly relevant to medical studies where a patient may be faced with choosing between two medical treatments, and where measurements of interest such as times to recovery or failure can reasonably be modelled with a Weibull distribution.

In the medical setting, increasing attention has been directed recently towards non-inferiority studies (see, for example, Fleming, 2008; Kwong, Cheung, Hayter, and Wen, 2012), where it has been recognized that as long as a new treatment is not worse than a standard treatment by more than a specified non-inferiority margin, then it can be preferable due to other reasons such as cost and availability. The calculation of win-probabilities $W_2(\rho)$ with ρ corresponding to a particular non-inferiority margin can be particularly useful in this case.

Kundu and Gupta (2006) provide a nice discussion of the estimation of the win-probability $W_2(\rho)$ when $\rho = 1$ and when it is assumed that the two shape parameters a_1 and a_2 are equal. They show how to obtain an approximate maximum likelihood estimate of the win-probability in this setting, and they derive approximate confidence intervals based on asymptotic normality arguments, bootstrap methods and also using a Bayesian approach.

Kundu and Gupta motivate the problem with respect to a standard engineering context concerning the mechanical reliability of a system, in which the first treatment corresponds to the stress that a system is subjected to, while the second treatment corresponds to its strength. Thus, $W_2(1)$ is the probability that the system's strength is larger than the stress to which it is subjected, in which case there will be no failure. This problem is also considered by Lin and Ke (2013) for general location-scale distributions with progressive Type-II censored data.

The approach taken in this paper is to consider inferences on win-probabilities $W_2(\rho)$ for general values of ρ, and without the assumption of equal shape parameters. Furthermore, the objective is to obtain confidence intervals for the win-probabilities that guarantee a nominal confidence level through the construction of confidence sets for the two sets of parameters with a guaranteed confidence level.

Moreover, a practitioner will usually be interested in a range of questions for this two-sample problem, starting with an assessment of whether the two data-sets can actually be modelled with Weibull distributions, progressing to whether there is a difference between the two distributions and leading to questions about the magnitude of such a difference if it exists. In addition, if one or both of the treatments are selected for further use, then individual inferences will be required on one or both of the distributions. Usually these questions are tackled separately with distinct error rates, which make it difficult to assess the overall confidence level of the complete statistical analysis.

The methodology presented in this paper through the construction of confidence sets for the two sets of parameters allows all of these questions to be answered with a guaranteed specified overall simultaneous confidence level. The use of a procedure such as this which addresses the multiplicity of the range of questions of interest to the practitioner has been discussed in Hayter (2014) with respect to normally distributed data.

The layout of this paper is as follows. Section 2 discusses the estimation of the win-probabilities, while the general methodology which in particular allows confidence interval construction for the win-probabilities is considered in Section 3. Some examples and illustration of these methodologies are provided in Section 4, and Section 5 contains a summary.

2. Estimation of the win-probabilities

It can be seen from Equation (1) that the win-probability $W_2(\rho)$ depends upon a_1, a_2 and $\tau = \rho\lambda_1/\lambda_2$. To provide an indication of this dependence, Figures 1–3 provide contour plots of $W_2(\rho)$ for $\tau = 1, 2$ and 3 (for $\tau = 1/2$ and $1/3$ the contour plots can be obtained from those for $\tau = 2$ and 3 by switching a_1 and a_2 and subtracting the contour values from 1). Also, Figure 4 provides an illustration of how the win-probabilities depend upon the value of ρ for several sets of parameter values. The R code to calculate the win-probabilities, which is available from the authors, uses R's standard numerical methodologies to evaluate the integral in Equation (1).

For given data-sets X_{1i}, $1 \le i \le n_1$ and X_{2i}, $1 \le i \le n_2$, the win-probability $W_2(\rho)$ for a particular value ρ of interest can be estimated by evaluating Equation (1) with estimates of the parameters

$$\hat{W}_2(\rho) = \int_0^\infty \hat{a}_1 x^{\hat{a}_1 - 1} e^{-(x^{\hat{a}_1} + (\hat{\tau}x)^{\hat{a}_2})} \mathrm{d}x,$$

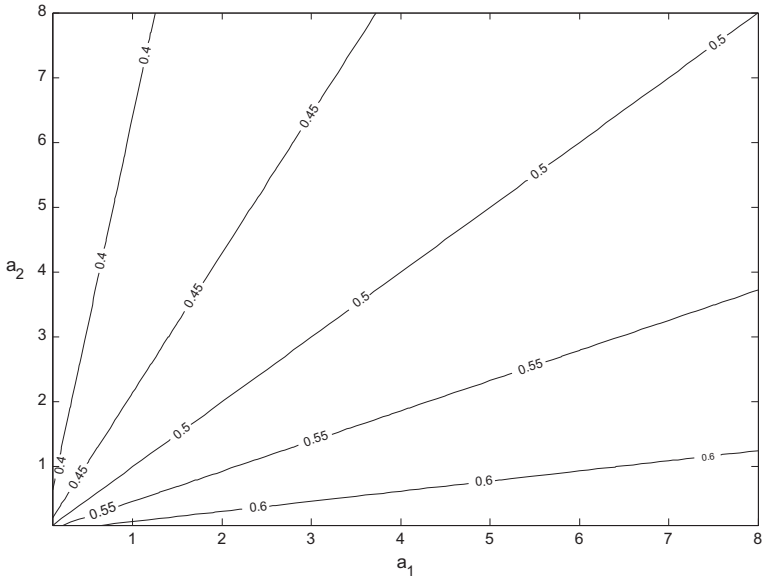


Figure 1. Contour plot of the win-probability $W_2(\rho)$ for $\tau = 1$.

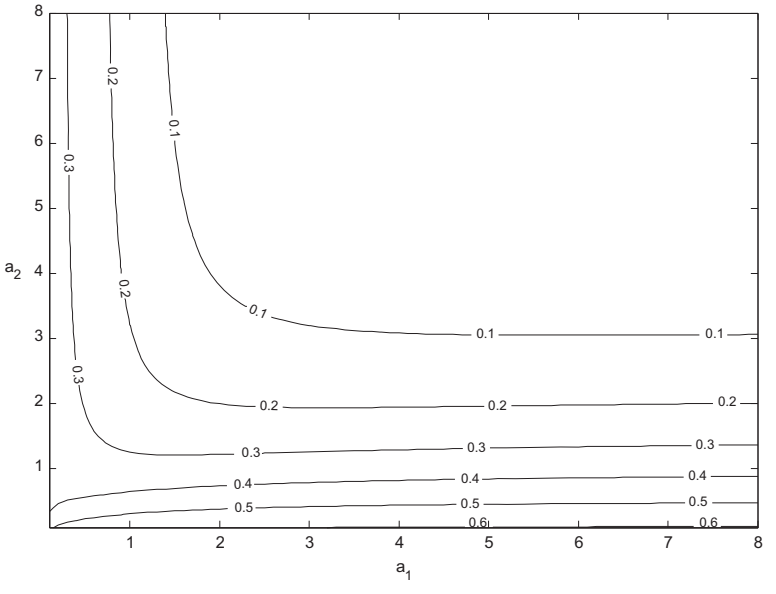


Figure 2. Contour plot of the win-probability $W_2(\rho)$ for $\tau = 2$.

where $\hat{\tau} = \rho\hat{\lambda}_1/\hat{\lambda}_2$. These parameter estimates may, for example, be obtained by maximum likelihood, as discussed by Balakrishnan and Kateri (2008).

Specifically, for data X_{1i}, $1 \leq i \leq n_1$, from a Weibull distribution with parameters a_1 and λ_1, the derivatives of the log-likelihood function l are

Table 1. Averages from $N = 100$ simulations of the maximum likelihood estimates (with standard deviations in brackets) and $1 - \alpha = 0.95$ confidence intervals for the win-probabilities.

	$a_1 = 2\ \lambda_1 = 1\quad a_2 = 3\ \lambda_2 = 1$								
$n_1 = n_2$	$W_2(0.5) = 0.869$			$W_2(1) = 0.527$			$W_2(2) = 0.193$		
20	0.874	(0.056)	(0.534,0.997)	0.534	(0.092)	(0.193,0.862)	0.194	(0.062)	(0.010,0.526)
40	0.868	(0.037)	(0.630,0.989)	0.522	(0.060)	(0.273,0.766)	0.183	(0.045)	(0.029,0.417)
70	0.869	(0.028)	(0.693,0.975)	0.521	(0.050)	(0.336,0.712)	0.186	(0.038)	(0.058,0.364)
100	0.866	(0.025)	(0.723,0.964)	0.525	(0.037)	(0.365,0.684)	0.192	(0.027)	(0.074,0.338)
200	0.871	(0.019)	(0.771,0.944)	0.527	(0.031)	(0.415,0.642)	0.191	(0.022)	(0.104,0.292)
	$a_1 = 2\ \lambda_1 = 1\quad a_2 = 3\ \lambda_2 = 0.6$								
$n_1 = n_2$	$W_2(0.5) = 0.634$			$W_2(1) = 0.261$			$W_2(2) = 0.077$		
20	0.648	(0.091)	(0.290,0.932)	0.263	(0.083)	(0.032,0.615)	0.077	(0.043)	(0.001,0.381)
40	0.642	(0.068)	(0.389,0.873)	0.266	(0.058)	(0.072,0.508)	0.079	(0.029)	(0.004,0.280)
70	0.636	(0.041)	(0.444,0.818)	0.259	(0.041)	(0.106,0.445)	0.076	(0.022)	(0.010,0.223)
100	0.638	(0.039)	(0.474,0.788)	0.262	(0.034)	(0.127,0.412)	0.077	(0.017)	(0.015,0.197)
200	0.635	(0.024)	(0.520,0.744)	0.260	(0.023)	(0.162,0.366)	0.076	(0.012)	(0.026,0.157)
	$a_1 = 2\ \lambda_1 = 1\quad a_2 = 3\ \lambda_2 = 1.4$								
$n_1 = n_2$	$W_2(0.5) = 0.945$			$W_2(1) = 0.719$			$W_2(2) = 0.331$		
20	0.947	(0.031)	(0.635,1.000)	0.730	(0.087)	(0.374,0.972)	0.339	(0.098)	(0.073,0.695)
40	0.944	(0.024)	(0.734,0.998)	0.715	(0.060)	(0.457,0.923)	0.325	(0.061)	(0.114,0.574)
70	0.946	(0.016)	(0.803,0.996)	0.720	(0.042)	(0.530,0.888)	0.329	(0.044)	(0.160,0.517)
100	0.942	(0.014)	(0.826,0.992)	0.709	(0.035)	(0.549,0.853)	0.318	(0.035)	(0.174,0.472)
200	0.945	(0.009)	(0.869,0.986)	0.719	(0.025)	(0.606,0.823)	0.332	(0.029)	(0.229,0.443)

$$\frac{\partial l}{\partial a_1} = \frac{n_1}{a_1} - n_1 \log \lambda_1 + \sum_{i=1}^{n_1} \log X_{1i} + \lambda_1^{-a_1} \log \lambda_1 \sum_{i=1}^{n_1} X_{1i}^{a_1} - \lambda_1^{-a_1} \sum_{i=1}^{n_1} X_{1i}^{a_1} \log X_{1i},$$

and

$$\frac{\partial l}{\partial \lambda_1} = -\frac{a_1 n_1}{\lambda_1} + a_1 \lambda_1^{-a_1 - 1} \sum_{i=1}^{n_1} X_{1i}^{a_1}.$$

Setting these derivatives to be zero gives $\lambda_1 = \left(\frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i}^{a_1}\right)^{\frac{1}{a_1}}$ and

$$\frac{n_1}{a_1} + \sum_{i=1}^{n_1} \log X_{1i} - \frac{n_1 \sum_{i=1}^{n_1} X_{1i}^{a_1} \log X_{1i}}{\sum_{i=1}^{n_1} X_{1i}^{a_1}} = 0.$$

A method such as Newton–Raphson can be used to solve this latter equation to give $\hat{a}_1$, which can then be used to obtain $\hat{\lambda}_1$. In a similar way the estimates $\hat{a}_2$ and $\hat{\lambda}_2$ can be obtained from the data X_{2i}, $1 \le i \le n_2$.

Indications of the accuracy of these maximum likelihood estimates of the win-probabilities are given by Table 1 and Figure 6 which are discussed in the subsequent sections.

3. Confidence interval construction for the win-probabilities

Confidence intervals for the win-probabilities $W_2(\rho)$ can be derived from confidence sets for a_1, a_2 and τ. In order to construct confidence intervals that guarantee a nominal

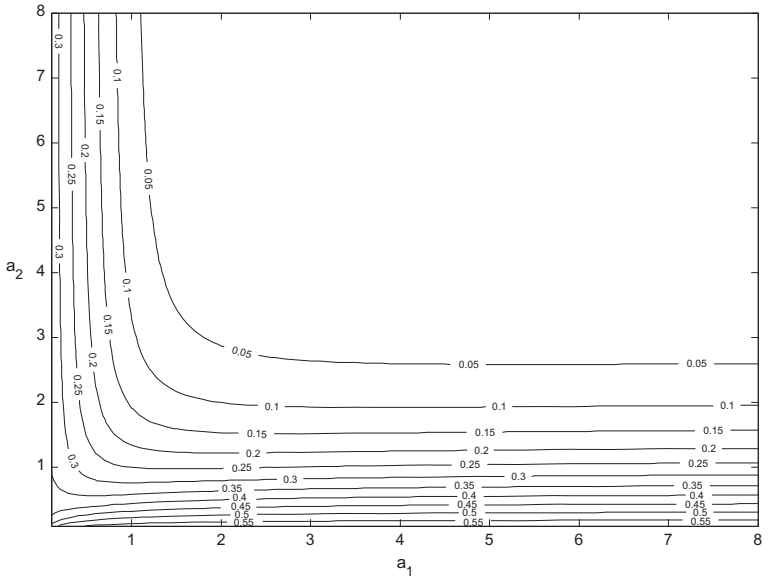


Figure 3. Contour plot of the win-probability $W_2(\rho)$ for $\tau = 3$.

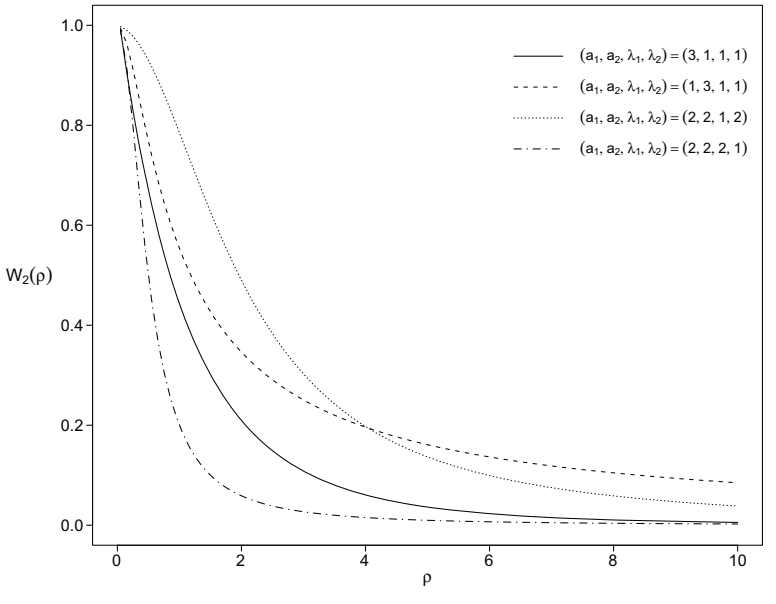


Figure 4. Dependence of the win-probability $W_2(\rho)$ on ρ for various parameter values.

confidence level of $1-\alpha$, the method of Hayter and Kiatsupaibul (2013) can be employed which derives a confidence set for the two parameters of a Weibull distribution with a guaranteed confidence level.

This method can be used to generate a confidence set for a_1 and λ_1 based on the data X_{1i}, $1 \le i \le n_1$, with a confidence level of $1-\alpha_1$, and a confidence set for a_2 and λ_2 based on the data X_{2i}, $1 \le i \le n_2$, with a confidence level of $1-\alpha_2$. If $1-\alpha = (1-\alpha_1)(1-\alpha_2)$

then these confidence sets can be used to provide confidence intervals for $W_2(\rho)$ with a guaranteed confidence level of $1-\alpha$.

In addition to providing confidence intervals for the win-probabilities, these two confidence sets for the parameters can also be used to indicate whether the data-sets fit Weibull distributions, since as explained by Hayter and Kiatsupaibul the confidence sets will be empty if there are no plausible Weibull distributions that fit the data. Furthermore, the intersection of the two confidence sets also indicates whether there are any common Weibull distributions which fit the two data-sets, and if so what their parameter values are. Finally, the confidence sets can also be used to provide individual confidence bounds on the two distribution functions, and all of these inferences are provided with a guaranteed overall simultaneous confidence level of $1-\alpha=(1-\alpha_1)(1-\alpha_2)$.

For the Weibull parameterization used in this paper, the method of Hayter and Kiatsupaibul can be employed as follows. For order statistics $X_{1(1)} \leq \ldots \leq X_{1(n_1)}$ from a Weibull distribution with parameters a_1 and λ_1, the $1-\alpha_1$ confidence set for these parameters is given by

$$-d_{\alpha_1,n_1}+\frac{i}{n_1} \leq 1-e^{-(X_{1(i)}/\lambda_1)^{a_1}} \leq d_{\alpha_1,n_1}+\frac{i-1}{n_1}$$

for $1 \leq i \leq n_1$, where d_{α_1,n_1} is the Kolmogorov critical point. Equivalently,

$$l_{1i} \leq a_1 \log X_{1(i)} - a_1 \log \lambda_1 \leq u_{1i}$$

for $1 \leq i \leq n_1$, where

$$l_{1i}=\log\left(\max\left\{0,-\log\left(\frac{n_1-i}{n_1}+d_{\alpha_1,n_1}\right)\right\}\right)$$

and

$$u_{1i}=\log\left(-\log\left(\max\left\{0,\frac{n_1-i+1}{n_1}-d_{\alpha_1,n_1}\right\}\right)\right).$$

Consequently, the values of a_1 satisfy

$$\max\left\{\frac{l_{1i}}{a_1}-\log X_{1(i)}, 1 \leq i \leq n_1\right\} \leq \min\left\{\frac{u_{1i}}{a_1}-\log X_{1(i)}, 1 \leq i \leq n_1\right\} \tag{2}$$

and for these values of a_1 the parameter λ_1 satisfies

$$e^{-\min\{u_{1i}/a_1-\log X_{1(i)}, 1\leq i\leq n_1\}} \leq \lambda_1 \leq e^{-\max\{l_{1i}/a_1-\log X_{1(i)}, 1\leq i\leq n_1\}}. \tag{3}$$

Therefore, the confidence set can be constructed by first finding the upper and lower bounds for a_1 from Equation (2). After this confidence interval for a_1 has been obtained, it is then straightforward to calculate the bounds on λ_1 corresponding to each value of a_1 from Equation (3). The confidence set for a_2 and λ_2 based on the data X_{2i}, $1 \leq i \leq n_2$, with a confidence level of $1-\alpha_2$ can be constructed in a similar manner.

A confidence interval for $W_2(\rho)$ with a guaranteed confidence level of $1-\alpha=(1-\alpha_1)(1-\alpha_2)$ can be derived from these confidence sets as follows. Notice that $W_2(\rho)$ is monotonically decreasing in λ_1 and monotonically increasing in λ_2, so that $W_2(\rho)$ is maximized by searching for the maximum value of

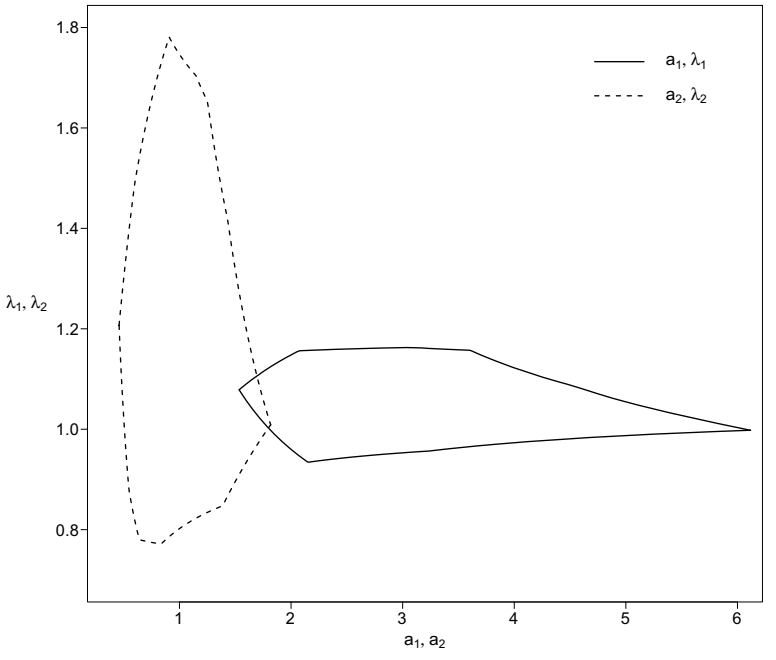


Figure 5. Confidence sets with individual confidence levels of $\sqrt{0.95}$ for simulated data-sets with $n_1 = n_2 = 60$ from Weibull distributions with parameters $a_1 = 3$ and $\lambda_1 = 1$, and $a_2 = 1$ and $\lambda_2 = 1$.

$$\int_0^{\infty} a_1 x^{a_1-1} e^{-\left(x^{a_1}+\left(\rho x \lambda_1^{\min}/\lambda_2^{\max}\right)^{a_2}\right)} \mathrm{d}x$$

over the confidence intervals for a_1 and a_2, and $W_2(\rho)$ is minimized by searching for the minimum value of

$$\int_0^{\infty} a_1 x^{a_1-1} e^{-\left(x^{a_1}+\left(\rho x \lambda_1^{\max}/\lambda_2^{\min}\right)^{a_2}\right)} \mathrm{d}x$$

over the confidence intervals for a_1 and a_2, where

$$\lambda_1^{\max} = e^{-\max\{l_{1i}/a_1 - \log X_{1(i)}, 1 \le i \le n_1\}}$$

$$\lambda_1^{\min} = e^{-\min\{u_{1i}/a_1 - \log X_{1(i)}, 1 \le i \le n_1\}}$$

$$\lambda_2^{\max} = e^{-\max\{l_{2i}/a_2 - \log X_{2(i)}, 1 \le i \le n_2\}}$$

and

$$\lambda_2^{\min} = e^{-\min\{u_{2i}/a_2 - \log X_{2(i)}, 1 \le i \le n_2\}}.$$

It can be noted that the Kolmogorov critical point, required for this methodology, has been extensively tabulated. In addition, it can be conveniently obtained from the R package using the routine 'kolmim', which uses the algorithm proposed by Marsaglia, Tsang, and Wang (2003) with an improvement by Luis Carvalho. Another efficient set of codes to compute the Kolmogorov critical point in C and Java has been provided by Simard and L'Ecuyer (2011).

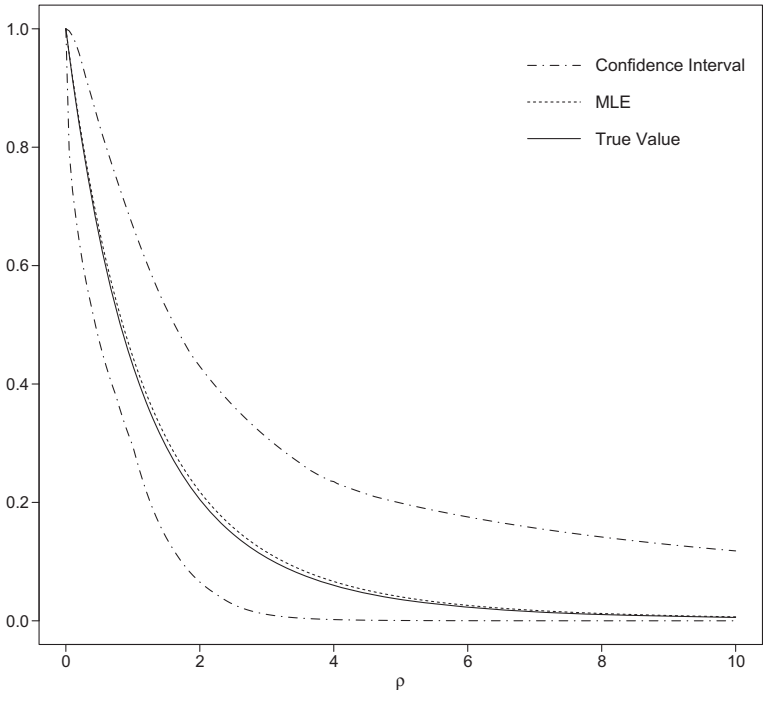


Figure 6. Estimates and confidence intervals for the win-probabilities $W_2(\rho)$ with a confidence level of 0.95 for simulated data-sets with $n_1 = n_2 = 60$ from Weibull distributions with parameters $a_1 = 3$ and $\lambda_1 = 1$, and $a_2 = 1$ and $\lambda_2 = 1$.

As an illustration of this methodology, data-sets of size $n_1 = n_2 = 60$ were simulated from Weibull distributions with parameters $a_1 = 3$ and $\lambda_1 = 1$, and with $a_2 = 1$ and $\lambda_2 = 1$. The resulting confidence sets for the two sets of parameters with $1 - \alpha_1 = 1 - \alpha_2 = \sqrt{0.95}$ are shown in Figure 5. The first thing to note from Figure 5 is that neither confidence set is empty, which implies that both data-sets can be modelled with a Weibull distribution, as would be expected. If this methodology is applied and a confidence set is empty, then this is a warning that a Weibull distribution is not appropriate for that data-set.

Furthermore, it is interesting to note from Figure 5 that there is a small non-empty intersection of the two confidence sets. This indicates that it is plausible that the two data-sets could be modelled with a common Weibull distribution, which would have a shape parameter a between 1.53 and 1.82 and a scale parameter λ between 1.00 and 1.11. The confidence sets can also be used to provide confidence bands for the two individual distribution functions, as explained in Hayter and Kiatsupaibul (2013).

Confidence intervals for the win-probabilities $W_2(\rho)$ for all ρ can also be obtained which together with all of these other inferences have an overall simultaneous confidence level of 0.95. Figure 6 shows the true values of the win-probabilities together with their maximum likelihood estimates discussed in Section 2, which can be seen to be very close for these sample sizes.

The confidence intervals are also shown, and for example, despite the difference in the shape parameters, the true value of $W_2(1)$ is close to 0.5 and the confidence interval $W_2(1) \in (0.295, 0.667)$ is obtained. Thus, the inference can be drawn from the data-sets that the probability that a potential future observation from the second treatment will

exceed a potential future observation from the first treatment is at least 0.295 but no more than 0.667.

In addition, the lower confidence interval is equal to 0.5 when $\rho = 0.44$, and so it can be inferred that it is at least as likely as not that a potential future observation from the second treatment will exceed 44% of a potential future observation from the first treatment. Also, the upper confidence interval is equal to 0.5 when $\rho = 1.62$, and so it can be inferred that it is at least as likely as not that a potential future observation from the first treatment will exceed $1/1.62 = 62\%$ of a potential future observation from the second treatment.

4. Examples and illustrations

In this section, some simulations and examples are presented to illustrate the methodology proposed in this paper.

4.1. Simulations

Table 1 shows some simulation results of the estimates of the win-probabilities $W_2(0.5)$, $W_2(1)$ and $W_2(2)$, together with their confidence intervals with a guaranteed confidence level of $1 - \alpha = 0.95$. Three sets of parameter configurations are considered, and the averages of $N = 100$ simulations are presented, together with the sample standard deviations of these 100 values. Individual confidence levels of $\sqrt{0.95}$ were used for the two confidence sets of the parameters.

These results provide an indication of how the confidence interval lengths depend upon the sample sizes, which are taken to be equal for the two treatments. It should be remembered that many other inferences besides the assessment of the win-probabilities are included with this confidence level, as illustrated in the following two examples.

4.2. Examples

Two examples with real data-sets are presented to illustrate the methodologies proposed in this paper.

Example 1: Kundu and Gupta (2006) analyse data on carbon fibre strengths taken from Badar and Priest (1982). In their example, the first sample is $n_1 = 69$ observations of fibre strengths for gauge lengths of 20 mm, while the second sample is $n_2 = 63$ observations of fibre strengths for gauge lengths of 10 mm. They argue that Weibull distributions are appropriate to model the two data-sets when a value of 0.75 has been subtracted from the strengths.

Figure 7 shows the individual $\sqrt{0.95}$ level confidence sets for the Weibull parameters of these two data-sets (with 0.75 subtracted). It is first important to note that neither confidence set is empty, which confirms the analysis of Kundu and Gupta that the two data-sets can be modelled with Weibull distributions. It is also interesting to note that the two confidence sets are disjoint, which establishes that the two data-sets cannot be modelled with a common Weibull distribution. Thus, a hypothesis test that the two treatments have identical Weibull distributions is rejected at size $\alpha = 0.05$.

Figure 8 shows the estimates and confidence intervals for the win-probabilities $W_2(\rho)$, and Table 2 provides some values of these estimates and confidence intervals together with

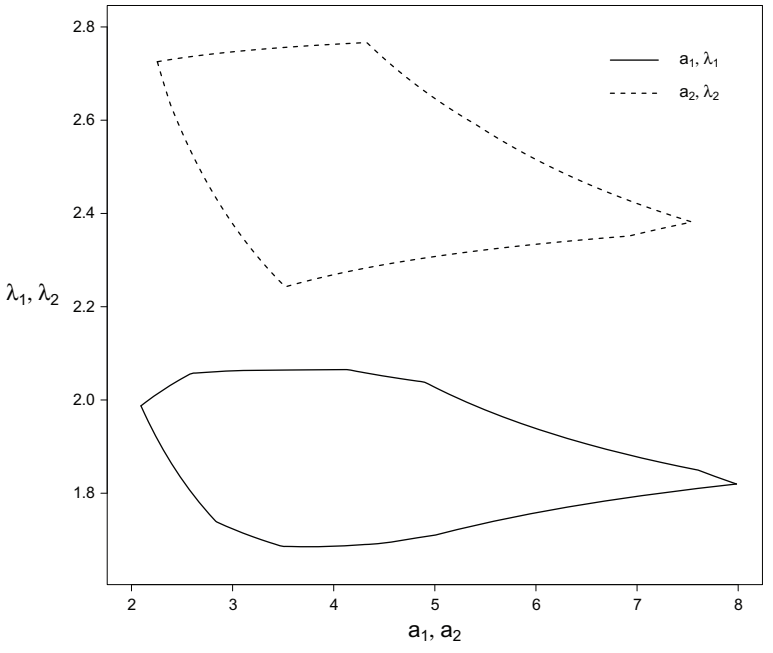


Figure 7. Confidence sets with individual confidence levels of $\sqrt{0.95}$ for the carbon fibre strength data in Example 1.

Table 2. Some estimates and confidence intervals for the carbon fibre strength data in Example 1 with an overall simultaneous confidence level of 0.95.

	Win-probabilities	
$W_2(0.5)$	0.980	(0.903,0.999)
$W_2(1)$	0.765	(0.566,0.907)
$W_2(2)$	0.182	(0.032,0.347)
	Gauge length 20 mm	
Expectation	1.700	(1.517,1.875)
$F^{-1}(0.05)$	0.868	(0.481,1.254)
$F^{-1}(0.25)$	1.360	(1.096,1.581)
$F^{-1}(0.50)$	1.709	(1.518,1.891)
$F^{-1}(0.75)$	2.047	(1.820,2.333)
$F^{-1}(0.95)$	2.501	(2.088,3.356)
	Gauge length 10 mm	
Expectation	2.304	(2.018,2.519)
$F^{-1}(0.05)$	1.191	(0.730,1.607)
$F^{-1}(0.25)$	1.850	(1.562,2.074)
$F^{-1}(0.50)$	2.317	(2.021,2.542)
$F^{-1}(0.75)$	2.766	(2.461,3.150)
$F^{-1}(0.95)$	3.369	(2.755,4.433)

some inferences on the expectations and quantiles of the individual strengths of the two types of carbon fibres. It can be seen that $W_2(1)$ is estimated to be 0.765 with a confidence interval $(0.566, 0.907)$ so that it can be inferred that a 10 mm fibre has at least a 0.566 probability of having a strength greater than a 20 mm fibre. In fact, the confidence interval

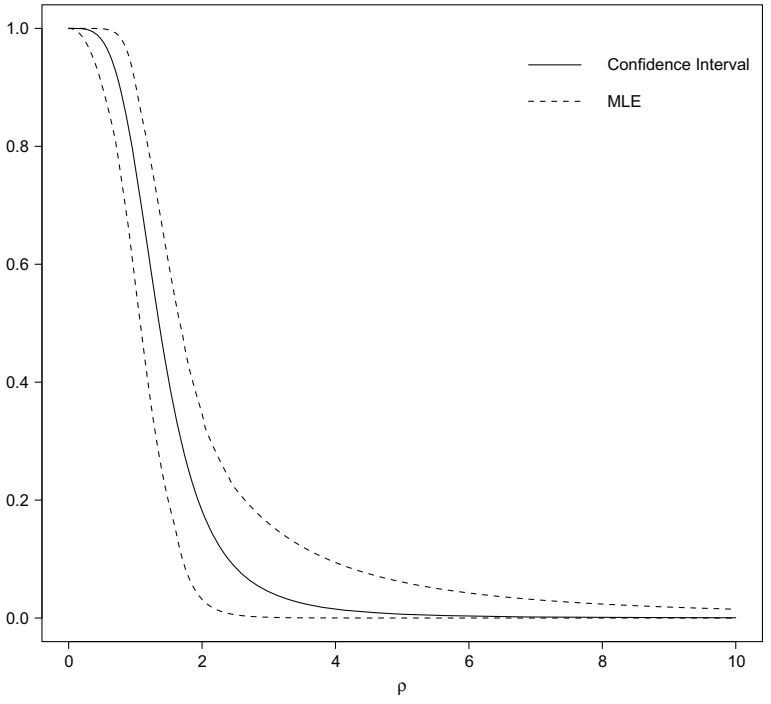


Figure 8. Estimates and confidence intervals for the win-probabilities $W_2(\rho)$ with a confidence level of 0.95 for the carbon fibre strength data in Example 1.

for $W_2(2)$ indicates that there could be a probability of as high as 0.347 that a 10 mm fibre has a strength at least twice that of a 20 mm fibre.

Kundu and Gupta discuss several ways of calculating approximate confidence intervals for $W_2(1)$ under the assumption that the shape parameters are equal (which is not an unreasonable assumption for these data where the maximum likelihood estimates are $\hat{a}_1 = 3.844$, $\hat{\lambda}_1 = 1.880$, $\hat{a}_2 = 3.910$, and $\hat{\lambda}_2 = 2.545$). They report similar estimates of about 0.762 for $W_2(1)$ with approximate 95% confidence intervals of about $(0.700, 0.828)$, which are substantially shorter than the confidence interval given in Table 2. From this perspective Kundu and Gupta's method is preferable.

However, this is balanced by the fact that with the guaranteed nominal confidence level of 0.95, the methodology presented in this paper has also tested whether the data can be modelled with Weibull distributions (this is examined separately by Kundu and Gupta with its own individual error rate), has established that the two data-sets cannot be modelled with a common Weibull distribution, has provided confidence intervals on the win-probabilities $W_2(\rho)$ for all values of ρ, and has provided confidence intervals on the expectations and quantiles of the individual strengths of the two types of carbon fibres, without having to assume that the two shape parameters are equal. The next example considers a situation where it is not reasonable to assume that the two shape parameters are equal.

Example 2: A communication switching machine undergoes cycles of up-time when it functions, and down-time when it is out of order and requires repairing. The following data-sets contain $n_1 = 38$ down-times of the machine and $n_2 = 44$ up-times of the machine

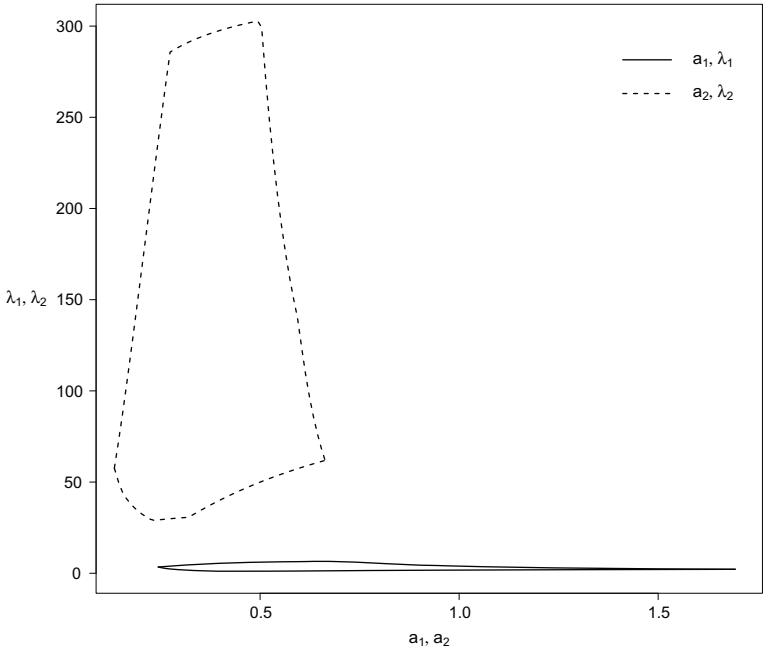


Figure 9. Confidence sets with individual confidence levels of $\sqrt{0.95}$ for the down-times and up-times data in Example 2.

in hours. This original real data-set was extracted from the log file of a communication switching machine operating in Khon Kaen University in Thailand.

Down-times:

6.650	0.233	0.683	0.450	2.417	2.933
3.334	0.867	3.200	0.200	1.233	6.361
0.019	1.683	11.467	2.917	0.093	0.783
0.933	3.951	4.592	28.101	0.0003	0.283
8.167	6.183	0.0006	0.0003	0.133	1.800
7.650	5.533	0.133	10.417	1.117	3.483
0.767	0.550				

Up-times:

1.333	35.783	119.167	233.817	1323.133	68.017
120.967	1100.317	716.017	52.517	476.467	0.050
314.150	12.183	233.050	42.750	149.417	818.383
154.483	67.500	19.333	0.001	0.0008	0.0006
0.0006	1.050	23.617	8.117	0.050	0.002
0.002	0.006	0.003	4.517	243.217	182.417
13.133	305.050	171.050	325.400	138.317	513.083
40.583	2.000				

It is useful to be able to compare the distributions of the down-times and the up-times. Specifically, if the benefits of the machine functioning are a linear function of the up-time, while the repair costs are a linear function of the down-time, then the win-probabilities provide an assessment of the benefit to cost ratios of this communication

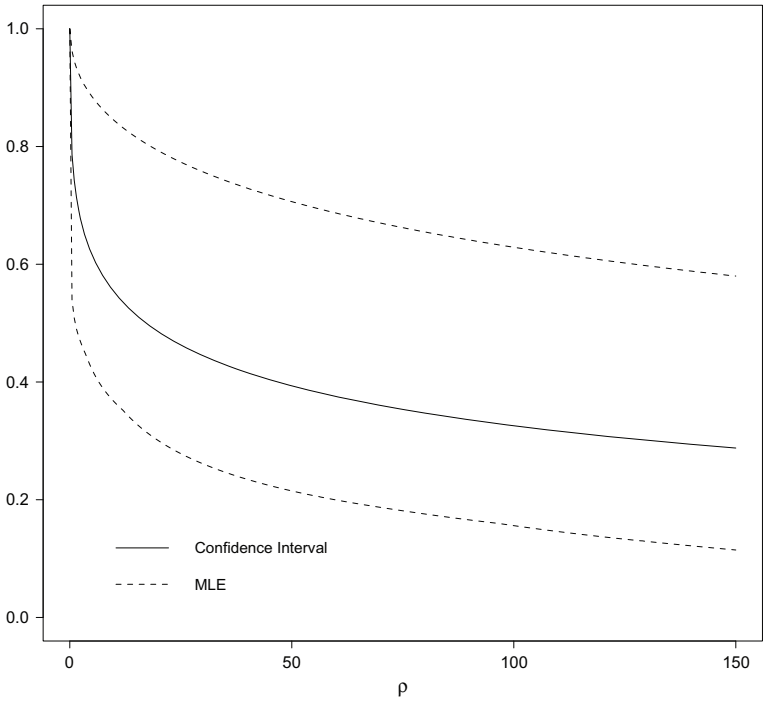


Figure 10. Estimates and confidence intervals for the win-probabilities $W_2(\rho)$ with a confidence level of 0.95 for the down-times and up-times data in Example 2.

switching machine. For example, if the benefit per unit time is twice the repair cost per unit time, then $W_2(0.5)$ is the probability that one cycle of repairing the machine and running it until failure will be economically advantageous.

Figure 9 shows the individual $\sqrt{0.95}$ level confidence sets for the Weibull parameters of these two data-sets, which can be seen to be disjoint and possessing quite different shapes. First of all, the fact that neither confidence set is empty indicates that both data-sets can be modelled with a Weibull distribution. In addition, since the confidence sets are disjoint it can be inferred that the two data-sets cannot be modelled with a common Weibull distribution. In fact, while the up-times are generated from the machine reliability, the down-times depend upon the human repair policy, and so it is quite reasonable to expect that the two distributions would be unequal.

The parameter estimates are $\hat{a}_1 = 0.572$, $\hat{\lambda}_1 = 2.314$, $\hat{a}_2 = 0.324$, and $\hat{\lambda}_2 = 58.276$, and even though the projections of the two confidence sets onto the shape parameter axis have some intersection, the non-intersected parts are much larger. Consequently, it is unreasonable to simplify the analysis by assuming that the two data-sets have identical shape parameters.

Figure 10 shows the estimates and confidence intervals for the win-probabilities $W_2(\rho)$. For example, $W_2(0.5)$ is estimated to be 0.787 with a confidence interval of $(0.540, 0.963)$, so that (when the benefit per unit time is twice the repair cost per unit time) it can be inferred that one cycle of repairing the machine and running it until failure is more likely than not to be economically advantageous, with an estimated chance of about 79%. Also, $W_2(1)$ is estimated to be 0.743 with a confidence interval of $(0.509, 0.945)$, and $W_2(2)$ is

estimated to be 0.693 with a confidence interval of (0.477, 0.924). Thus, it can be inferred that there is at least about a 48% chance that one cycle of repairing the machine and running it until failure will generate benefits of four times the cost, and that this can be estimated to occur with a chance of about 69%.

On the other hand, in this communications setting a shutdown can be catastrophic in terms of cost, and so it may be the case that the benefit per unit time is only 1% of the repair cost per unit time. In this case $W_2(100)$ is estimated to be 0.326 with a confidence interval of (0.156, 0.629) so that there is an estimated chance of only about 33 that one cycle of repairing the machine and running it until failure will be economically advantageous.

5. Summary

This paper has considered the problem of comparing two Weibull distributions. The current statistical literature does not offer much scope for addressing this problem, with available procedures relying on large sample asymptotic results and assumptions about the parameters.

The procedure proposed in this paper allows a thorough investigation of the two Weibull distributions, so that testing whether Weibull distributions are appropriate, identifying whether the two Weibull distributions can be taken to be identical (and if so, identification of the common parameter values), and comparisons of the difference between the two Weibull distributions with win-probabilities are all achieved with a guaranteed overall simultaneous confidence level.

The win-probabilities with different choices of ρ allow an assessment of the practical difference between the two Weibull distributions. They can be particularly useful when a choice must be made between the two distributions and provide information that can be combined with other factors such as the costs, availabilities and side-effects of the two choices. R code is available to implement the methodology discussed in this paper, and it can be requested from the authors.

This methodology can also be applied to making inferences on other parametric models, and in particular to the wide range of additional two-parameter and three-parameter Weibull models, which are discussed in Murthy, Bulmer, and Eccleston (2004a, 2004b) and Nadarajah and Kotz (2008), for example. Extensions can also be made to the comparisons of three or more distributions through the construction of joint confidence sets for the parameters from each of the distributions.

Acknowledgements

We would like to sincerely thank each of the reviewers for their helpful and insightful comments and suggestions that have resulted in a much improved version of this manuscript. We would also like to acknowledge the support of Dr. Kanda Runapongsa Saikaew and her team at the Computer Center of Khon Kaen University, Thailand, for help with the data-set used in Example 2. This work was supported by the Thailand Research Fund and Chulalongkorn University [RSA5780005].

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Dr. Anthony Hayter is a Professor in the Department of Business Information and Analytics at the University of Denver in USA, and has interests in various areas of statistics, reliability, and statistical computing.

Ping Yang is currently a PhD student in Statistics at the Chinese University of Hong Kong, under the supervision of Professor Siu Hung Cheung and Professor Wai-Yin Poon. She received her BS in Statistics from Zhejiang University, China, in 2012. Her research focuses on multiple comparisons.

Dr. Seksan Kiatsupaibul is an Associate Professor in the Department of Statistics at Chulalongkorn University in Thailand, and has interests in statistics, reliability, R programming, finance and business.

References

Abernethy, R. B. (2006). *The new Weibull handbook: Reliability and statistical analysis for predicting life, safety, supportability, risk, cost and warranty claims* (5th ed.). New York, NY: Barringer and Associates.

Badar, M. G., & Priest, A. M. (1982). Statistical aspects of fiber and bundle strength in hybrid composites. In T. Hayashi, K. Kawata, & S. Umekawa (Eds.), *Progress in science and engineering composites* (pp. 1129–1136). Tokyo.

Balakrishnan, N., & Kateri, M. (2008). On the maximum likelihood estimation of parameters of Weibull distribution based on complete and censored data. *Statistics and Probability Letters, 78*, 2971–2975.

Fleming, T. R. (2008). Current issues in non-inferiority trials. *Statistics in Medicine, 27*, 317–332.

Hayter, A. J. (2012). Win-probabilities for regression models. *Statistical Methodology, 9*, 520–527.

Hayter, A. J. (2013). Inferences on the difference between future observations for comparing two treatments. *Journal of Applied Statistics, 40*, 887–900.

Hayter, A. J. (2014). Identifying common normal distributions. *Test, 23*, 135–152.

Hayter, A. J. (in press). Win-probabilities for comparing two Poisson variables. *Communications in Statistics – Theory and Methods*.

Hayter, A. J., & Kiatsupaibul, S. (2013). Exact inferences for a Weibull model. *Quality Engineering, 25*, 175–180.

Hudak, D., & Tiryakioglu, M. (2011). On comparing the shape parameters of two Weibull distributions. *Materials Science and Engineering A, 528*, 8028–8030.

Kundu, D., & Gupta, R. D. (2006). Estimation of $P[Y \leq X]$ for Weibull distributions. *IEEE Transactions on Reliability, 55*, 270–280.

Kwong, K. S., Cheung, S. H., Hayter, A. J., & Wen, M. (2012). Extension of three-arm non-inferiority studies to trials with multiple new treatments. *Statistics in Medicine, 31*, 2833–2843.

Lawless, J. F. (2003). *Statistical models and methods for lifetime data analysis* (2nd ed.). New York, NY: John Wiley.

Lin, C. T., & Ke, S. J. (2013). Estimation of $P(Y \leq X)$ for location-scale distributions under joint progressively Type-II right censoring. *Quality Technology & Quantitative Management, 10*, 339–352.

Louzada-Neto, F., Bolfarine, H., & Rodrigues, J. (2002). Comparing two Weibull models with accelerated data. *Statistics, 36*, 175–184.

Marsaglia, G., Tsang, W. W., & Wang, J. (2003). Evaluating Kolmogorov's distribution. *Journal of Statistical Software, 8*(18), 1–4.

Murthy, D. N. P., Bulmer, M., & Eccleston, J. A. (2004a). Weibull model selection for reliability modelling. *Reliability Engineering and System Safety, 86*, 257–267.

Murthy, D. N. P., Xie, M., & Jiang, R. (2004b). *Weibull models*. New York, NY: John Wiley.

Nadarajah, S., & Kotz, S. (2008). Strength modeling using Weibull distributions. *Journal of Mechanical Science and Technology, 22*, 1247–1254.

Parsi, S., Ganjali, M., & Farsipour, N. S. (2011). Conditional maximum likelihood and interval estimation for two Weibull populations under joint Type-II progressive censoring. *Communications in Statistics – Theory and Methods, 40*, 2117–2135.

Rinne, H. (2008). *The Weibull distribution: A handbook*. New York, NY: CRC Press.

Schafer, R. E., & Sheffield, T. S. (1976). On procedures for comparing two Weibull populations. *Technometrics, 18*, 231–235.

Simard, R., & L'Ecuyer, P. (2011). Computing the two-sided Kolmogorov-Smirnov distribution. *Journal of Statistical Software, 39*, 11.

Wiwatwattana, N., Hayter, A. J., & Kiatsupaibul, S. (2015). Win-probabilities for comparing two binary outcomes. *Communications in Statistics – Simulation and Computation*, in press.