

## The Art of Statistics Learning from Data.

David Spiegelhalter (2020)

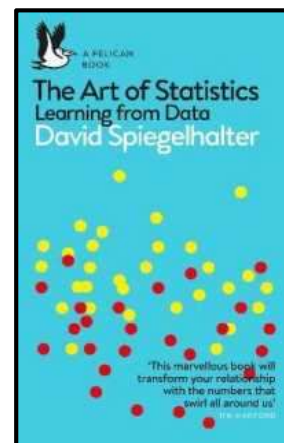
ISBN: 978-024-125-876-7.

คมกริช รุมดอน<sup>1,\*</sup>

Komgrit Rumdon <sup>1,\*</sup>

<sup>1</sup> หอสมุดจอห์น เอฟ. เคนเนดี สำนักวิทยบริการ มหาวิทยาลัยสงขลานครินทร์ ประเทศไทย; John F. Kennedy Library, Office of Academic Resources, Prince of Songkla University, Thailand

\* Corresponding author email: komgrit.r@psu.ac.th



สังคมในยุคปัจจุบันมีการนำข้อมูลและสารสนเทศมาใช้ประกอบการตัดสินใจในชีวิตประจำวัน ทั้งในส่วนของการตัดสินใจเกี่ยวกับการดำเนินธุรกิจ การศึกษา และการประกอบอาชีพ ทั้งนี้ การที่จะนำข้อมูลมาใช้ประกอบการตัดสินใจได้นั้น จำเป็นอย่างยิ่งที่จะต้องมียุทธศาสตร์ทางด้านการวิเคราะห์ ทักษะการใช้เครื่องมือการใช้เทคโนโลยีสารสนเทศหรือโปรแกรมทางสถิติในการวิเคราะห์ข้อมูล ทักษะความรู้ทางด้านวิทยาศาสตร์ข้อมูล ทักษะการรู้สารสนเทศ และองค์ความรู้พื้นฐานเกี่ยวกับสถิติและเงื่อนไขทางสถิติ เพื่อนำไปสู่การตัดสินใจที่มีประสิทธิภาพบนพื้นฐานของเหตุและผลและหลักการทางอัลกอริทึม

สำหรับหนังสือเรื่อง The Art of Statistics Learning from Data แต่งโดย David Spiegelhalter เป็นหนึ่งในหนังสือที่มีเนื้อหาโดดเด่นทางด้านสถิติ เหมาะสำหรับบุคคลที่มีทักษะพื้นฐานและความสนใจทางด้านคณิตศาสตร์ประยุกต์ วิทยาการคำนวณ และวิทยาการข้อมูล โดยเนื้อหาในภาพรวมครอบคลุมในประเด็นด้านสัดส่วนข้อมูลโดยการจำแนกประเภทและอัตราเปอร์เซ็นต์ การสรุปและนำเสนอตัวเลข การกำหนดประชากรและกลุ่มตัวอย่าง การใช้เหตุและผล การสร้างโมเดลความสัมพันธ์โดยใช้การถดถอย การใช้อัลกอริทึมและการทำนาย การกำหนดค่าความเชื่อมั่น ความน่าจะเป็น การทดสอบสมมติฐาน และการประเมินและเลือกใช้ข้อมูลจากสถิติ นอกจากนี้ หนังสือเรื่อง The Art of Statistics Learning from Data ของ David Spiegelhalter ยังมีการยกประเด็นตัวอย่างเกี่ยวกับการเลือกใช้สถิติให้เหมาะสมกับข้อมูล การเลือกใช้กราฟให้เหมาะสมกับประเภทของข้อมูล และการเลือกกลุ่มตัวอย่างที่เหมาะสม เพื่อให้ได้รับสารสนเทศที่ดีจากการวิเคราะห์ข้อมูลด้วยสถิติที่เหมาะสม ทำให้ผู้อ่านรู้สึกสนุก เข้าใจง่าย ด้วยการยกตัวอย่างที่หลากหลาย ซึ่งจะช่วยให้ผู้อ่านเกิดความเข้าใจในเนื้อหาและกลวิธีทางสถิติที่เหมาะสม และสามารถนำไปใช้ปฏิบัติได้จริง

หนังสือเล่มนี้มีการเรียบเรียงเนื้อหาด้วยบทต่าง ๆ ที่มีความสอดคล้องกัน โดยเริ่มต้นตั้งแต่สัดส่วนข้อมูลโดยการจำแนกประเภทและอัตราเปอร์เซ็นต์ ซึ่งเป็นการปูพื้นฐานให้แก่ผู้อ่านและปิดท้ายด้วยการสรุปเนื้อหาในภาพรวมที่จะช่วยให้ผู้อ่านมองเห็นภาพรวมของหนังสือตั้งแต่เริ่มต้นจนถึงการสรุปเนื้อหา ทำให้ผู้อ่านไม่เกิดความสับสน โดยเนื้อหาภายในเล่ม ประกอบด้วยประเด็นเนื้อหาหลัก 13 ประเด็น ดังนี้

ประเด็นที่ 1 และ 2 เป็นการนำเสนอจำนวนและสัดส่วนของชุดข้อมูล เช่น การวิเคราะห์ข้อมูล เหตุการณ์หนึ่งว่าจะเกิดขึ้นหรือไม่ จะใช้ข้อมูลทวิภาค (Binary data) เนื่องจากผลลัพธ์จะมีเพียงสองค่า ซึ่งอาจสรุปข้อมูลด้วยความถี่และค่าร้อยละ นอกจากนี้การนำเสนอข้อมูลแบบจัดกลุ่ม การเลือกใช้แผนภูมิวงกลม จะทำให้รู้สึกสับสนและเปรียบเทียบขนาดของพื้นที่ข้อมูลแต่ละกลุ่มยาก ซึ่งหากเลือกใช้แผนภูมิแท่งจะทำให้เข้าใจและเปรียบเทียบง่ายมากกว่า ทั้งนี้ในส่วนของการนำเสนอตัวเลขด้วยแผนภาพแต่ละแบบมีข้อดีที่แตกต่างกัน เช่น แผนภูมิสตรีปจะแสดงจุดข้อมูลทุกจุด แผนภาพกล่องและเส้น ซึ่งจะสรุปด้วยสายตาได้รวดเร็ว และฮิสโทแกรมจะเห็นการแจกแจงข้อมูลได้ดี เห็นได้ชัดว่ามีหลากหลายวิธีที่ใช้ศึกษาข้อมูลจำนวนมาก แต่การแสดงผลด้วยภาพที่ดีจะต้องมีสารสนเทศที่เชื่อถือได้ และนำเสนอข้อมูลในรูปแบบกราฟหรืออินโฟกราฟิกได้อย่างสวยงาม แต่ไม่ควรสำคัญกว่าความชัดเจนและความลุ่มลึกของข้อมูล เนื้อหาในประเด็นดังกล่าวจะช่วยอธิบายและเปรียบเทียบให้ผู้อ่านได้เห็นถึงความแตกต่างของการใช้กราฟได้ เช่น การนำเสนอข้อมูลจากแผนภูมิแท่งจะทำให้ผู้อ่านเข้าใจง่ายและเปรียบเทียบข้อมูลได้ดีกว่า แต่ขณะเดียวกันแผนภูมิวงกลมไม่สามารถบอกได้ถึงความมากมายของข้อมูลแต่ละชุดข้อมูล ทั้งนี้ในประเด็นข้างต้นมีการแสดงตัวอย่างประกอบเพื่อนำไปสู่การสร้างความเข้าใจให้แก่ผู้อ่านเพิ่มเติมด้วย

ประเด็นที่ 3 การกำหนดประชากรและกลุ่มตัวอย่างให้เหมาะสม กล่าวถึงกระบวนการเปลี่ยนข้อมูลดิบที่ได้จากแบบสอบถามให้เป็นค่ากล่าวอ้างเกี่ยวกับการศึกษาพฤติกรรมของมนุษย์ ประกอบไปด้วย 4 ขั้นตอน เริ่มตั้งแต่ขั้นตอนการนำข้อมูลดิบโดยการตั้งข้อตกลงเบื้องต้นเพื่อไปสู่ขั้นตอนกลุ่มตัวอย่าง และการนำกลุ่มตัวอย่างที่กำหนดไว้ไปสู่ขั้นตอนประชากรที่ศึกษา จากนั้นนำประชากรที่ศึกษาไปศึกษาความเป็นไปได้ที่จะเข้าร่วมการสำรวจเพื่อนำไปสู่การกำหนดประชากรเป้าซึ่งเป็นขั้นตอนสุดท้าย ซึ่งในประเด็นที่ 3 นี้ ผู้เขียนใช้วิธีการเขียนเชิงชี้แนะพร้อมกับการใช้เทคนิคตั้งคำถามกับผู้อ่านเพื่อนำไปสู่การสร้างแรงจูงใจในการอ่านในประเด็นต่อไป

ประเด็นที่ 4 ถึง 9 ผู้เขียนกล่าวถึงความเป็นเหตุเป็นผลว่าแนวคิดเชิงสถิติที่เกี่ยวกับความน่าจะเป็นกับสถิติ และค่าประมาณและความเชื่อมั่นที่ไม่ชี้ชัด เมื่ออยากรู้อะไรเป็นสาเหตุของอะไร ผู้เขียนใช้วิธีแนะนำให้ใช้วิธีการทดลองหรือเฟ้นสุ่มเพื่อหาข้อเท็จจริงที่เกิดขึ้น นอกจากนี้ยังมีการกล่าวถึงการสร้างโมเดลความสัมพันธ์โดยใช้การถดถอยเพื่อนำไปสู่การพยากรณ์ปัจจัยหรือความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตาม รวมไปถึงการสร้างสมการหรือโมเดลคณิตศาสตร์สำหรับการทำนายความน่าจะเป็นบนพื้นฐานของหลักเหตุและผลผ่านค่าสถิติของการวิเคราะห์ขั้นสูง ทั้งนี้ ผู้เขียนยังเชื่อมโยงแนวคิดเชิงสถิติความเป็นเหตุและผลโดยการสร้างโมเดลความสัมพันธ์เข้ากับหลักการทางด้านวิทยาการคอมพิวเตอร์ด้วยการใช้อัลกอริทึมและการเรียนรู้ของเครื่อง (Machine Learning) ในการจำแนกประเภทข้อมูลเพื่อเรียนรู้แบบมีผู้สอน (Supervised learning) และการเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning) ในการทำนายหรือการแนะนำ จากประเด็นดังกล่าวแสดงให้เห็นว่าผู้เขียนให้ความสำคัญเกี่ยวกับการเชื่อมโยงเนื้อหาระหว่างประเด็นตั้งแต่แนวคิดการให้เหตุผลไปสู่การประยุกต์ใช้โมเดลที่เห็นถึงความสัมพันธ์และความน่าจะเป็นของข้อมูลและ

สารสนเทศที่ได้จากการวิเคราะห์ แต่ทั้งนี้ในทัศนคติของผู้วิจารณ์ให้ข้อสังเกตว่า ผู้อ่านจำเป็นที่จะต้องมีความรู้ และทักษะพื้นฐานทางด้านวิทยาศาสตร์ข้อมูล (Data Science) ประกอบด้วยซึ่งจะทำให้เกิดความเข้าใจ มากขึ้น

ประเด็นที่ 10 การทดสอบสมมติฐาน เมื่อเราสร้างกราฟเพื่อวิเคราะห์และหาข้อสรุปผ่านโมเดล คณิตศาสตร์ที่เหมาะสมแล้ว จะนำไปสู่การทดสอบสมมติฐาน ซึ่งผู้เขียนได้ยกตัวอย่างการนำสถิติมาใช้ในการ ทดสอบ เช่น ทางทดสอบแบบทางเดียว การทดสอบแบบสองทาง และการทดสอบความสัมพันธ์ไคส นอกจากนี้ยังมีการนำทฤษฎีเนย์แมน-เพียร์สัน เพื่อนำมาพิจารณาทางเลือกที่เป็นไปได้หลังการทดสอบ สมมติฐานอย่างมีหลักการ

ประเด็นที่ 11 ทฤษฎีของเบส์ ผู้เขียนแสดงให้เห็นถึงการใช้นิเวศเพื่อนำมาศึกษาความน่าจะเป็นแบบมี เงื่อนไขโดยพิจารณาจากข้อมูลที่เกี่ยวข้องหรือมีความสัมพันธ์กับเหตุการณ์เพื่อนำมาศึกษาความน่าจะเป็นที่จะ เกิดขึ้น นอกจากนี้ยังสามารถนำมาใช้ทดสอบสมมติฐานของการวิจัย ซึ่งแสดงให้เห็นว่าผู้เขียนใช้เทคนิค กระบวนการขั้นสูงให้แก่ผู้อ่าน ซึ่งเป็นองค์ประกอบหนึ่งของกระบวนการวิจัยในระดับต่าง ๆ

ประเด็นที่ 12 และ 13 เป็นการนำเสนอตัวอย่างความผิดพลาดในการใช้สถิติกับกลุ่มตัวอย่างและการ วิเคราะห์ข้อมูล และการเลือกรูปแบบการนำเสนอสารสนเทศและสถิติที่ไม่เหมาะสม เช่น การเลือกกลุ่ม ตัวอย่างที่สะดวกมากกว่ากลุ่มที่เป็นตัวแทนประชากรจริง การออกแบบกลุ่มตัวอย่างที่มีขนาดเล็กเกินไป การ ปรับแต่งข้อมูลที่ทำให้ผลลัพธ์ที่ออกมาไม่เป็นจริง และการนำเสนอข้อเท็จจริงหรือข้อค้นพบที่ไม่เหมาะสม ซึ่ง จะทำให้เกิดกระแสวิพากษ์วิจารณ์ในสังคม โดยไม่สนใจข้อมูลที่ดี นอกจากนี้ผู้เขียนยังกล่าวถึงการประเมินและ เลือกใช้ข้อมูลสถิติที่ดี โดยพิจารณาความน่าเชื่อถือของตัวเลข ความน่าเชื่อถือของแหล่งข้อมูล และการ ตีความหมาย ท้ายที่สุดผู้เขียนยังให้ข้อเสนอแนะที่จำเป็นและพึงกระทำเกี่ยวกับการนำเสนอข้อมูลทั้งท้ายไว้ว่า “การคำนึงถึงจริยศาสตร์ข้อมูล เช่น ไม่บิดเบือนข้อมูล ใช้สถิติที่เหมาะสม เลือกรูปแบบตัวอย่างที่เป็นประชากร จริง ความเป็นส่วนตัวและกรรมสิทธิ์ของข้อมูล รวมถึงการยินยอมให้นำข้อมูลไปเผยแพร่” เป็นหัวใจสำคัญอีก หนึ่งประเด็นที่ผู้อ่านควรตระหนักทั้งก่อนและหลังการนำข้อมูลและสารสนเทศไปใช้เพื่อการนำเสนอ การบอก ต่อ การเผยแพร่ และการประยุกต์ใช้ในชีวิตประจำวัน

จากหนังสือเล่มนี้ โดยสรุปแล้วพบว่า สถิติศาสตร์มีบทบาทสำคัญในชีวิตและเปลี่ยนแปลงอยู่เสมอตาม ปริมาณและความลุ่มลึกของข้อมูลที่เพิ่มขึ้น แต่การศึกษาสถิติไม่เพียงมีผลกระทบต่อสังคมโดยรวมแต่ยังมี ผลกระทบในระดับบุคคลด้วย แต่ทั้งนี้ การที่ผู้อ่านจะสามารถทำความเข้าใจเกี่ยวกับการประยุกต์ใช้สถิติได้ อย่างแม่นยำและมีประสิทธิภาพนั้น ผู้วิจารณ์มีความคิดเห็นว่า จำเป็นอย่างยิ่งที่ควรจะมีความรู้พื้นฐาน ทางด้านคณิตศาสตร์ประยุกต์ วิทยาการคำนวณ รวมถึงวิทยาการคอมพิวเตอร์ มากไปกว่านั้น จำเป็นอย่างยิ่ง ที่จะต้องมีความรู้พื้นฐานของศาสตร์ทางด้านสารสนเทศหรือวิทยาการสารสนเทศด้วย เนื่องจากการมีเครื่องมือ และสถิติที่ดีแล้วนั้นจะต้องควบคู่กับการมีข้อมูล สารสนเทศ และทักษะการประเมินสารสนเทศที่ดีประกอบ

เพิ่มด้วย จึงจะได้สารสนเทศที่มีคุณภาพเพื่อนำไปสู่การประยุกต์ใช้ในชีวิตประจำวัน และประกอบการตัดสินใจที่มีประสิทธิภาพได้เพิ่มมากยิ่งขึ้น

#### เอกสารอ้างอิง

Spiegelhalter, David. (2020). **The Art of Statistics: Learning from Data**. London: Penguin Books Ltd.