สารบัญ

แด่นักสถิติจากทุกหนทุกแห่ง
ที่มีนิสัยช่างใส่ใจรายละเอียด
ใจกว้าง และซื่อตรงอย่างน่ารักน่าเอ็นดู
รวมทั้งความปรารถนาที่จะใช้ข้อมูลอย่างดีที่สุดเท่าที่เป็นไปได้

The Art of Statistics

Learning from Data

|

David Spiegelhalter

แปลโดย

สุนันทา วรรณสินธ์ เบล

บทนำ

ตัวเลขอธิบายตัวเองไม่ได้ เราต้องพูดแทน เราเป็นผู้ให้ความหมายแก่ตัวเลข

— เนต ซิลเวอร์ (Nate Silver), *สัญญาณและคลื่นแทรก (The Signal and the Noise)*[1]

ทำไมเราจึงจำเป็นต้องมีสถิติ

ฮาโรลด์ ชิปแมน (Harold Shipman) เป็นฆาตกรต้องโทษที่สังหารคนมากที่สุดในบริเตน แม้ว่าประวัติและบุคลิกของเขาไม่เหมือนฆาตกรต่อเนื่องทั่วไปก็ตาม เขาเป็นแพทย์เวชศาสตร์ครอบครัวผู้สุขุมนุ่มนวลในเขตชานเมืองแมนเชสเตอร์ ระหว่างปี 1975-1998 เขาฉีดยาฝิ่นเกินขนาดให้คนไข้อย่างน้อย 215 รายซึ่งส่วนใหญ่เป็นผู้สูงอายุ ในที่สุดเขาก็พลาดพลั้งเมื่อเขาปลอมแปลงพินัยกรรมของเหยื่อรายหนึ่งเพื่อให้ตนได้รับเงินมรดกบางส่วน ลูกสาวของเหยื่อซึ่งเป็นนักกฎหมายเกิดสงสัยขึ้นมา และจากการวิเคราะห์ตรวจสอบคอมพิวเตอร์ของชิปแมนก็พบว่าเขาแก้ไขประวัติคนไข้ภายหลังให้ดูเหมือนว่าคนไข้ป่วยหนักกว่าที่เป็นจริง เขาได้ชื่อว่าเป็นคนเปิดรับเทคโนโลยีใหม่ก่อนใครๆ แต่เขาไม่ชำชองมากพอจะตระหนักว่ามีการบันทึกเวลา (time-stamp) ไว้ทุกครั้งที่เขาปรับเปลี่ยนข้อมูล (กรณีนี้ชี้ให้เห็นว่าตัวอย่างข้อมูลที่ดีนั้นมีความหมายแฝง)

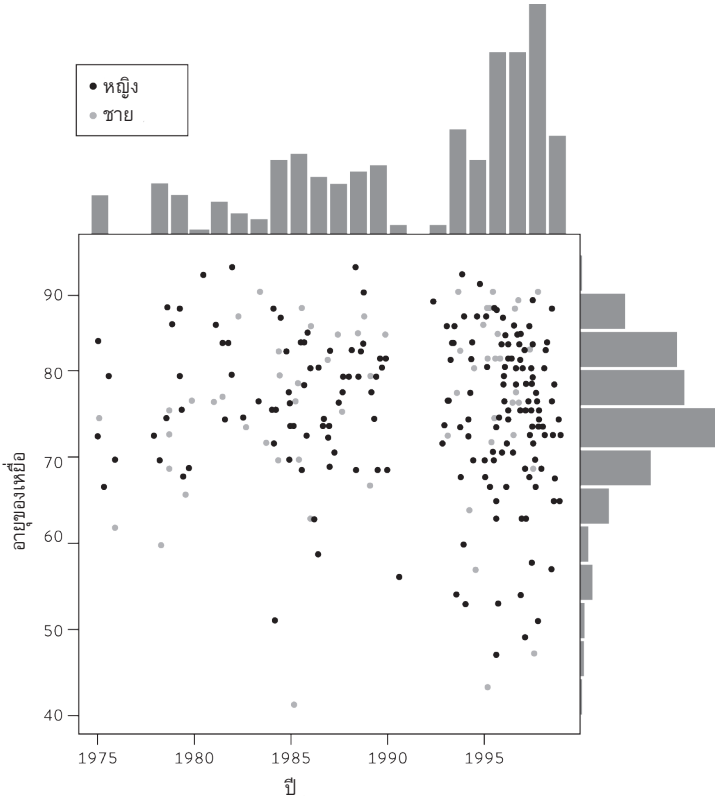
จากการขุดศพผู้ป่วยทั้งหมดที่ไม่ได้รับการฌาปนกิจขึ้นมาตรวจ พบว่า 15 รายมีไดอะมอร์ฟีนซึ่งเป็นเฮโรอีนรูปแบบที่ใช้ทางการแพทย์ ในปริมาณที่เป็นอันตรายถึงชีวิต ชิปแมนจึงถูกดำเนินคดีข้อหาฆาตกรรมเหยื่อ 15 รายในปี 1999 แต่เขาเลือกที่จะไม่สู้คดีและไม่เปิดปากพูดแม้แต่คำเดียวในชั้นศาล ศาลพิพากษาว่าเขาผิดจริงและต้องโทษจำคุกตลอดชีวิต รวมทั้งมีการไต่สวนสาธารณะเพื่อศึกษาว่าเขาอาจก่ออาชญากรรมอื่นอีกหรือไม่นอกเหนือจากที่โดนดำเนินคดี และเพื่อหาคำตอบว่าเขาน่าจะถูกจับได้เร็วกว่านี้หรือไม่ ผมเป็นหนึ่งในนักสถิติที่ได้รับเชิญไปให้ข้อมูลในการไต่สวนสาธารณะครั้งนั้น ซึ่งสรุปว่าเขาฆ่าคนไข้ไป 215 ราย อย่างแน่นอน และอาจมีเพิ่มอีก 45 รายด้วย[2]

หนังสือเล่มนี้จะเน้นการใช้**สถิติศาสตร์*** **(statistical science)** เพื่อตอบคำถามที่ผุดขึ้นเมื่อเราต้องการเข้าใจโลกให้ดียิ่งกว่าเดิม ผมได้เน้นคำถามเหล่านี้ไว้โดยล้อมกรอบสีเทา คำถามแรกที่ผุดขึ้นโดยธรรมชาติหากต้องการทำความเข้าใจพฤติกรรมของชิปแมนคือ

ฮาโรลด์ ชิปแมน ฆ่าคนประเภทไหน และพวกเขาตายเมื่อไร

การไต่สวนสาธารณะให้ข้อมูลอายุ เพศ และวันตายของเหยื่อ ภาพประกอบ 0.1 แสดงข้อมูลนี้ในรูปแบบภาพที่ค่อนข้างละเอียด โดยแสดงให้เห็นการกระจายตัวของอายุเหยื่อเมื่อเทียบกับวันตาย เฉดสีของจุดระบุว่าเหยื่อเป็นเพศชายหรือหญิง ทับซ้อนกับแผนภูมิแท่งซึ่งแสดงกลุ่มอายุ (แบ่งเป็นแท่งละ 5 ปี) และปีที่เกิดเหตุบนแกนทั้งสองด้าน

* ข้อความ**ที่ใช้อักษรตัวหนา**มีคำอธิบายในประมวลศัพท์ท้ายเล่ม ซึ่งมีทั้งคำนิยามพื้นฐานและคำจำกัดความทางเทคนิค



ภาพประกอบ 0.1

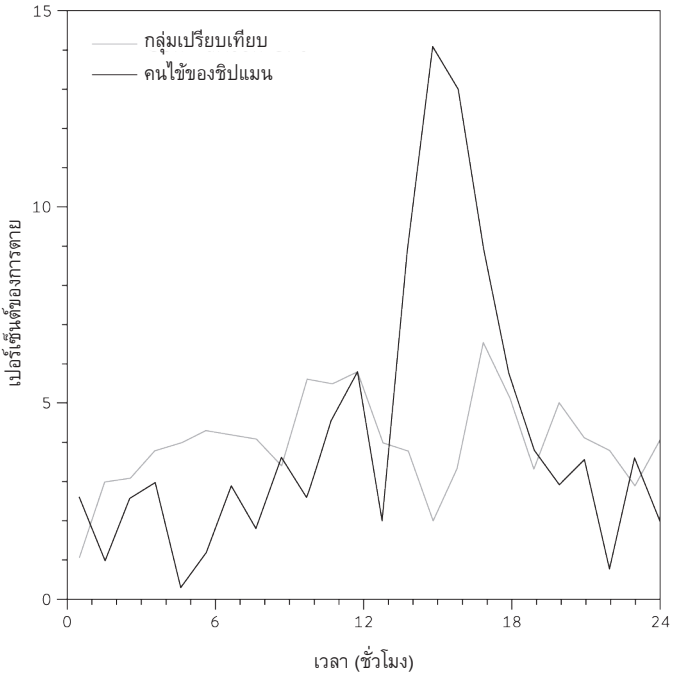
แผนผังการกระจายแสดงถึงอายุและปีที่เสียชีวิตของคนไข้ 215 ราย ซึ่งได้รับการยืนยันว่าเป็นเหยื่อของฮาโรลด์ ชิปแมน แผนภูมิแท่งเสริมแกนด้านข้างระบุรูปแบบของกลุ่มอายุและปีที่เขาลงมือฆาตกรรม

เราได้ข้อสรุปบางประการจากการพิจารณาภาพนี้ ในแผนผังมีจุดสีดำมากกว่าสีขาว เท่ากับว่าเหยื่อของชิปแมนส่วนใหญ่เป็นผู้หญิง แผนภูมิแท่งด้านขวาของภาพแสดงให้เห็นว่าเหยื่อส่วนใหญ่อยู่ในช่วงอายุ 70-80 กว่าๆ แต่เมื่อดูการกระจายตัวของจุด เราจะเห็นว่าแม้ในตอนแรก

เหยื่อทุกรายเป็นผู้สูงอายุ แต่ก็มีเหยื่ออายุน้อยผุดแซมขึ้นบ้างเมื่อเวลาผ่านไป ขณะที่แผนภูมิแท่งด้านบนแสดงให้เห็นชัดเจนว่าช่วงประมาณปี 1992 ไม่เกิดการฆาตกรรมเลย ปรากฏว่าก่อนหน้านั้นชิปแมนทำงานในคลินิกร่วมกับแพทย์คนอื่น เป็นไปได้ว่าเขารู้สึกว่ามีคนสงสัยและจับตามอง ต่อมาเขาจึงออกมาตั้งคลินิกของตัวเองเป็นเอกเทศ และหลังจากนั้นจำนวนเหตุฆาตกรรมก็เพิ่มขึ้น ดังที่เห็นได้จากแท่งแผนภูมิด้านบน

การวิเคราะห์ข้อมูลเหยื่อซึ่งได้รับการระบุตัวตนจากการไต่สวนสาธารณะทำให้เกิดคำถามเพิ่มเติมเกี่ยวกับวิธีการฆาตกรรม มีหลักฐานทางสถิติจากข้อมูลช่วงเวลาการตายของผู้ที่คาดว่าเป็นเหยื่อของเขาตามที่บันทึกในมรณบัตร ภาพประกอบ 0.2 เป็นกราฟเส้นเปรียบเทียบช่วงเวลาที่คนไข้ของชิปแมนเสียชีวิตกับเวลาตายของคนไข้ที่สุ่มตัวอย่างจากแพทย์คนอื่นในท้องที่เดียวกัน ไม่ต้องวิเคราะห์โดยละเอียดก็เห็นรูปแบบได้ชัดเจน ข้อสรุปนั้นเรียกได้ว่า “ตำตา” เพราะมันเด่นชัดมาก คนไข้ของชิปแมนมีแนวโน้มสูงมากที่จะตายตอนบ่ายอ่อนๆ

ข้อมูลไม่สามารถบอกเราได้ว่า *ทำไม* พวกเขาจึงตายในช่วงเวลานั้น แต่การสืบค้นต่อเผยว่าเขาไปเยี่ยมไข้ตามบ้านหลังมื้อกลางวัน ซึ่งเป็นช่วงเวลาที่เขามักได้อยู่กับคนไข้สูงอายุตามลำพัง ชิปแมนจะแนะนำให้คนไข้ฉีดยาซึ่งเขาบอกว่าจะช่วยให้สบายตัวขึ้น แต่ความจริงเป็นไดอะมอร์ฟีนในปริมาณที่อันตรายถึงชีวิต หลังจากคนไข้สิ้นใจอย่างสงบต่อหน้าเขา เขาจะเปลี่ยนประวัติทางการแพทย์ของคนไข้เพื่อให้ดูเหมือนว่าเป็นการตายด้วยเหตุธรรมชาติ ท่านผู้หญิงเจเน็ต สมิธ (Janet Smith) ประธานการไต่สวนสาธารณะกล่าวในเวลาต่อมาว่า “ถึงตอนนี้ดิฉันยังรู้สึกว่ามันเป็นเรื่องสะเทือนขวัญเหนือคำบรรยาย ทั้งเหนือคำบรรยาย เกินคาด และไม่อาจจินตนาการได้ ว่าเขาเล่นบทแพทย์ผู้ห่วงใยดูแลคนไข้วันแล้ววันเล่าทั้งที่พกอาวุธสังหารติดกระเป๋า ... ซึ่งเขาจะหยิบออกมาใช้อย่างไม่สะทกสะท้าน”



ภาพประกอบ 0.2
เวลาที่คนไข้ของฮาโรลด์ ชิปแมน เสียชีวิต เทียบกับคนไข้ของแพทย์คนอื่นๆ ในย่านนั้น ไม่จำเป็นต้องวิเคราะห์สถิติซับซ้อนก็เห็นรูปแบบได้ชัดเจน

เขากล้าเสี่ยงในระดับหนึ่ง เนื่องจากหากมีการชันสูตรพลิกศพเพียงรายเดียวก็จะเปิดโปงเขาได้ แต่ด้วยอายุของคนไข้และสาเหตุการตายที่ดูเหมือนเป็นเหตุธรรมชาติ จึงไม่มีการชันสูตรศพแม้แต่รายเดียว นอกจากนี้ยังไม่เคยมีคำอธิบายถึงเหตุผลเบื้องหลังฆาตกรรมเหล่านี้ เขาไม่ให้การในศาล ไม่เคยปริปากพูดถึงการกระทำผิดของเขากับใคร แม้แต่กับครอบครัวของตัวเอง และสุดท้ายก็ฆ่าตัวตายในคุก ซึ่งประจวบเหมาะกับเวลาที่ภรรยาจะได้รับบำเหน็จบำนาญของเขาพอดี

เราอาจมองว่างานสำรวจตรวจค้นและการคำนวณซ้ำๆ แบบนี้เป็นสถิติเชิง “นิติวิทยาศาสตร์” (“forensic” statistics) ซึ่งในกรณีนี้จริงตามตัวอักษร ไม่ต้องใช้คณิตศาสตร์ ไม่ต้องอาศัยทฤษฎี แค่มองหารูปแบบอันอาจนำไปสู่คำถามที่น่าสนใจยิ่งขึ้น เราระบุรายละเอียดความผิดของชิปแมนได้จากหลักฐานของแต่ละกรณี แต่การวิเคราะห์ข้อมูลเช่นนี้ช่วยให้เกิดความเข้าใจโดยรวมว่าเขาก่ออาชญากรรมอย่างไร

ในบทที่ 10 เราจะดูว่าการวิเคราะห์สถิติอย่างเป็นกิจจะลักษณะจะช่วยให้จับตัวชิปแมนได้เร็วขึ้นหรือไม่* ก่อนที่จะไปถึงจุดนั้น เรื่องราวของชิปแมนแสดงให้เห็นถึงศักยภาพมหาศาลของการใช้ข้อมูลเพื่อช่วยให้เราเข้าใจโลกและใช้วิจารณญาณได้ดีขึ้น นี่คือหัวใจของสถิติศาสตร์

เปลี่ยนโลกเป็นข้อมูล

การใช้สถิติวิเคราะห์อาชญากรรมของฮาโรลด์ ชิปแมน กำหนดให้เราต้องก้าวถอยหลังออกมามองภาพกว้างของเรื่องน่าเศร้ายาวเป็นหางว่าวซึ่งเขาเป็นผู้กระทำ เราต้องย่นย่อรายละเอียดส่วนตัวอันมีเอกลักษณ์เฉพาะเกี่ยวกับชีวิตและการตายของผู้คน ให้กลายเป็นข้อเท็จจริงและตัวเลขชุดหนึ่งซึ่งสามารถนับและวาดเป็นกราฟได้ เมื่อมองแวบแรกอาจดูเย็นชาและไร้ความเป็นมนุษย์ แต่หากเราจะใช้สถิติศาสตร์เพื่อทำความเข้าใจโลก เราต้องแปลงประสบการณ์ในชีวิตแต่ละวันของเราให้เป็นข้อมูล ซึ่งหมายถึงการแยกประเภทและติดป้ายเหตุการณ์ต่างๆ บันทึกค่า วิเคราะห์ผล จนถึงการสื่อสารข้อสรุป

แต่เพียงแค่การแยกประเภทและติดป้ายก็อาจสร้างความท้าทายใหญ่หลวง ลองใคร่ครวญคำถามง่ายๆ ต่อไปนี้ ซึ่งคนที่ใส่ใจสิ่งแวดล้อมน่าจะสนใจ

* เฉลย: ทำได้แน่นอน

ในโลกนี้มีต้นไม้ทั้งหมดกี่ต้น

ก่อนจะคิดว่าเราจะหาคำตอบได้อย่างไร เราต้องทำความเข้าใจประเด็นที่ค่อนข้างพื้นฐานก่อนว่า “ต้นไม้” คืออะไร คุณอาจรู้สึกว่าแค่เห็นมันคุณก็รู้ได้ทันทีว่าเป็นต้นไม้ แต่การตัดสินของคุณอาจผิดแผกแตกต่างจากคนอื่นที่มองว่ามันเป็นแค่พุ่มไม้หรือกอไม้ ดังนั้น ในการแปลงประสบการณ์เป็นข้อมูล เราต้องเริ่มด้วยคำนิยามที่เคร่งครัดชัดเจน

ปรากฏว่าคำนิยามอย่างเป็นทางการของ “ต้นไม้” คือ พืชที่มีลำต้นเป็นไม้และมีเส้นผ่านศูนย์กลางที่ความสูงระดับอก (diameter at breast height - DBH) ใหญ่พอสมควร กรมป่าไม้สหรัฐอเมริการะบุว่าพืชต้องมี DBH มากกว่า 5 นิ้ว (12.7 เซนติเมตร) จึงจะนับว่าเป็นต้นไม้ แต่หน่วยงานส่วนใหญ่ใช้เกณฑ์ว่าต้นไม้ต้องมี DBH 10 ซม. (4 นิ้ว)

ทว่าเราไม่สามารถเดินทางไปทั่วโลกเพื่อวัดขนาดพืชที่มีลำต้นตันทุกต้นแล้วนับจำนวนที่ผ่านเกณฑ์นี้ นักวิจัยที่ศึกษาปัญหานี้จึงต้องใช้วิธีที่ดำเนินการได้จริง ขั้นแรกพวกเขาศึกษาอาณาเขตที่มีภูมิประเทศเดียวกันซึ่งเรียกว่าชีวนิเวศ (biome) และนับจำนวนต้นไม้โดยเฉลี่ยที่พบต่อพื้นที่หนึ่งตารางกิโลเมตร จากนั้นพวกเขาใช้ภาพถ่ายดาวเทียมเพื่อประเมินพื้นที่ทั้งหมดของโลกที่จัดอยู่ในชีวนิเวศแต่ละประเภท สร้างโมเดลทางสถิติที่ซับซ้อน และในที่สุดก็ได้ยอดรวมโดยประมาณจำนวน 3.04 ล้านล้านต้น (3,040,000,000,000) บนโลกใบนี้ ซึ่งดูเหมือนมีจำนวนมาก แต่พวกเขาคาดว่าโลกเคยมีต้นไม้มากกว่านี้ถึงสองเท่า*[3]

* ตัวเลขนี้รายงานพร้อมขอบเขตความคลาดเคลื่อนที่หนึ่งแสนล้านต้น ซึ่งหมายความว่านักวิจัยมั่นใจว่าตัวเลขแท้จริงอยู่ระหว่าง 2.94-3.14 ล้านล้านต้น (ผมยอมรับว่าตัวเลขนี้อาจแม่นยำเกินจริงเพราะมีข้อตกลงหลายประการในการสร้างโมเดลนี้) พวกเขายังประเมินว่าในแต่ละปีมีการตัดต้นไม้หนึ่งหมื่นห้าพันล้านต้น (15,000,000,000) และโลกสูญเสีย 46% ของต้นไม้ทั้งหมดนับตั้งแต่อารยธรรมมนุษย์เริ่มขึ้น

ถ้าหน่วยงานต่างๆ เห็นไม่ตรงกันเรื่องคำนิยามของต้นไม้ ก็ไม่น่าประหลาดใจที่แนวคิดซึ่งคลุมเครือกว่านี้จะจำกัดความได้ยากยิ่งยกตัวอย่างสุดโต่ง คำนิยามอย่างเป็นทางการของ “การว่างงาน” ในสหราชอาณาจักรถูกปรับเปลี่ยนอย่างน้อย 31 ครั้งระหว่างปี 1979-1996[4] คำนิยามของผลิตภัณฑ์มวลรวมภายในประเทศ (GDP) ก็ถูกปรับเปลี่ยนอยู่ตลอดเวลา เช่น ในปี 2014 การค้ายาผิดกฎหมายและการค้าประเวณีถูกเพิ่มเข้าไปนับรวมใน GDP ของสหราชอาณาจักร ซึ่งค่าประมาณใช้ตัวเลขจากแหล่งข้อมูลฉีกแนว เช่น จากเว็บไซต์ Punternet ที่รีวิวและให้คะแนนบริการทางเพศ รวมถึงให้ข้อมูลราคาสำหรับกิจกรรมต่างๆ[5]

แม้กระทั่งความรู้สึกส่วนตัวลึกๆ ของเราก็สามารถเปลี่ยนเป็นรหัสและนำมาวิเคราะห์เชิงสถิติได้ เมื่อสิ้นปีงบประมาณ 2017 คนในสหราชอาณาจักรจำนวน 150,000 คนได้รับเชิญให้เข้าร่วมประเมินในการสำรวจ “โดยรวมแล้ว เมื่อวานนี้คุณมีความสุขเพียงใด”[6] จากระดับ 0-10 คำตอบโดยเฉลี่ยอยู่ที่ 7.5 สูงกว่าปี 2012 ที่เท่ากับ 7.3 ซึ่งผลที่ได้อาจเกี่ยวเนื่องกับการฟื้นตัวทางเศรษฐกิจนับตั้งแต่ภาวะเศรษฐกิจตกต่ำในปี 2008 คะแนนต่ำสุดพบมากในกลุ่มอายุระหว่าง 50-54 ปี และสูงสุดในกลุ่มอายุ 70-74 ปี ซึ่งเป็นรูปแบบปกติสำหรับประชากรในสหราชอาณาจักร*

การวัดค่าความสุขถือเป็นเรื่องยาก ขณะที่การตัดสินว่าคนหนึ่งๆ ยังมีชีวิตอยู่หรือตายแล้วยังน่าจะง่ายกว่า ดังที่จะได้เห็นจากตัวอย่างในหนังสือเล่มนี้ การอยู่รอดและการตายเป็นประเด็นที่พบบ่อยในสถิติศาสตร์ ทว่าในสหรัฐอเมริกาแต่ละรัฐอาจมีนิยามความตายเชิงกฎหมายต่างกัน และแม้ว่าจะประกาศใช้รัฐบัญญัติจัดระเบียบการกำหนดนิยามของความตาย (Uniform Determination of Death Act) ในปี 1981 เพื่อเป็นต้นแบบร่วมกัน แต่ก็ยังมีข้อแตกต่างเล็กน้อยอยู่ ใครคนหนึ่งซึ่งถูกประกาศยืนยันว่าตายแล้วในรัฐแอละบามา อย่างน้อยโดยหลักการ ก็อาจไม่ใช่

* ซึ่งมอบความหวังแก่ผม ถ้าผมเป็นเหมือนคนโดยเฉลี่ย

คนตายตามกฎหมายเมื่อข้ามเส้นแบ่งรัฐเข้าไปยังฟลอริดา ซึ่งระบุว่าการยืนยันการตายต้องกระทำโดยแพทย์ทรงคุณวุฒิสองคน[7]

ตัวอย่างเหล่านี้แสดงให้เห็นว่าสถิติสร้างขึ้นจากฐานวิจารณญาณเสมอ และผิดมหันต์ที่จะคิดว่าทุกแง่มุมซับซ้อนของประสบการณ์ส่วนตัวสามารถแปลงเป็นรหัสที่ไม่คลุมเครือ และจัดวางบนตารางสเปรดชีตหรือซอฟต์แวร์อื่นได้ แม้ว่าการนิยาม การนับจำนวน รวมทั้งการวัดค่าลักษณะของเราเองและโลกรอบตัวเราเป็นเรื่องท้าทาย แต่ทั้งหมดนี้ก็เป็นเพียงสารสนเทศและเป็นจุดเริ่มต้นสู่การทำความเข้าใจโลกอย่างแท้จริง

การใช้ข้อมูลเป็นแหล่งความรู้มีข้อจำกัดหลักๆ สองประการ ประการแรกคือ เราไม่มีทางวัดค่าสิ่งที่สนใจจริงๆ ได้อย่างสมบูรณ์ การถามว่าเมื่อสัปดาห์ที่ผ่านมาคนมีความสุขเพียงใดจากระดับ 0-10 นั้นแทบไม่สามารถสรุปครอบคลุมสุขภาวะทางอารมณ์ของคนทั้งชาติได้ ประการที่สองคือ สิ่งใดก็ตามที่เราเลือกวัดจะแตกต่างไปตามสถานที่ บุคคล รวมทั้งช่วงเวลา และปัญหาคือการกลั่นกรองความรู้ที่มีความหมายจาก**ความเปลี่ยนแปร (Variability)** ทั้งหมดซึ่งดูเหมือนไร้รูปแบบ

สถิติศาสตร์ได้เผชิญหน้ากับความท้าทายคู่นี้และรับบทตัวเอกในความพยายามที่จะเข้าใจโลกในเชิงวิทยาศาสตร์ สถิติมอบพื้นฐานสำหรับการตีความข้อมูลซึ่งไม่มีวันสมบูรณ์แบบ เพื่อที่จะแยกแยะความสัมพันธ์สำคัญๆ จากความเปลี่ยนแปรพื้นฐานที่ทำให้เราทุกคนแตกต่างกัน แต่โลกเปลี่ยนแปลงอยู่เสมอ พร้อมกับที่คำถามและแหล่งข้อมูลใหม่ๆ ผุดขึ้น สถิติศาสตร์จึงต้องเปลี่ยนแปลงเช่นกัน

เรานับและวัดอยู่เสมอ แต่สาขาวิชาสถิติสมัยใหม่เริ่มต้นขึ้นในช่วงทศวรรษ 1650 เมื่อแบลส ปาสกาล (Blaise Pascal) และปิแยร์ เดอ แฟร์มา (Pierre de Fermat) ทำความเข้าใจความน่าจะเป็นอย่างถูกต้องเป็นครั้งแรก ดังที่เราจะได้เห็นในบทที่ 8 สาขาสถิติก้าวหน้าได้อย่างรวดเร็วเมื่อมีพื้นฐานทางคณิตศาสตร์ที่มั่นคงเช่นนี้ในการรับมือกับ

ความเปลี่ยนแปร ทฤษฎีความน่าจะเป็นมอบพื้นฐานหนักแน่นในการคำนวณบำนาญและเบี้ยเงินรายปีเมื่อใช้ร่วมกับข้อมูลเกี่ยวกับอายุขัยของคน เกิดการปฏิวัติดาราศาสตร์เมื่อนักวิทยาศาสตร์เข้าใจว่าทฤษฎีความน่าจะเป็นสามารถใช้รับมือกับความเปลี่ยนแปรในการวัดค่าต่างๆ ได้อย่างไร ผู้คนสมัยวิกตอเรียที่คลั่งไคล้ข้อมูลเริ่มเก็บข้อมูลเกี่ยวกับร่างกายมนุษย์ (รวมทั้งเรื่องอื่นทุกเรื่อง) และก่อรากฐานความเชื่อมโยงแน่นแฟ้นระหว่างการวิเคราะห์ทางสถิติ กับพันธุศาสตร์ ชีววิทยา และการแพทย์ ต่อมาในศตวรรษที่ 20 สถิติเน้นคณิตศาสตร์ยิ่งขึ้น และนับว่าเป็นโชคร้ายของนักศึกษาและนักสถิติหลายคนที่สถิติถูกประยุกต์ใช้เป็นชุดเครื่องมือทางสถิติคละแบบโดยอัตโนมัติ เครื่องมือหลายชิ้นตั้งชื่อตามนักสถิติสติเฟื่องผู้ชอบโต้แย้งซึ่งเราจะได้รู้จักต่อไปในหนังสือเล่มนี้

มุมมองทั่วไปที่เห็นว่าสถิติเป็น “ชุดเครื่องมือ” เบื้องต้นชุดหนึ่งกำลังเผชิญความท้าทายขนานใหญ่ ข้อแรก เราอยู่ในยุค**วิทยาศาสตร์ข้อมูล (data science)** ซึ่งเก็บรวบรวมชุดข้อมูลขนาดใหญ่และซับซ้อนจากแหล่งทั่วไป เช่น การตรวจสอบสัญญาณเข้าออกภายในการใช้งานเครือข่าย (traffic monitor หรือ network monitor) ข้อความที่โพสต์ในโซเชียลมีเดียและการซื้อของผ่านอินเทอร์เน็ต และนำมาใช้เป็นพื้นฐานสำหรับนวัตกรรมทางเทคโนโลยี เช่น การวางแผนการเดินทางที่ดีที่สุด โฆษณาแบบเจาะกลุ่มเป้าหมาย หรือระบบการแนะนำสินค้า เราจะกล่าวถึง**อัลกอริธึม (algorithms)** ที่มีพื้นฐานจาก “**บิ๊กดาต้า**” **(big data)** ในบทที่ 6 การฝึกอบรมด้านสถิติถือเป็นองค์ประกอบสำคัญหนึ่งสำหรับการเป็นนักวิทยาศาสตร์ข้อมูลมากยิ่งขึ้นทุกวัน รวมทั้งทักษะการบริหารจัดการข้อมูล การเขียนโปรแกรม และพัฒนาอัลกอริธึม ตลอดจนความรู้เกี่ยวกับด้านนี้โดยตรงด้วย

ความท้าทายอีกอย่างหนึ่งต่อมุมมองดั้งเดิมของสถิติมาจากการที่จำนวนการวิจัยเชิงวิทยาศาสตร์เพิ่มสูงขึ้นอย่างมาก โดยเฉพาะอย่างยิ่งในสาขาวิทยาศาสตร์ชีวการแพทย์และสังคมศาสตร์ นอกจากนี้ยังมีความกดดันให้ตีพิมพ์งานวิจัยในวารสารที่จัดอยู่ในอันดับสูง เรื่องนี้นำไปสู่ความ

คลางแคลงใจเกี่ยวกับความน่าเชื่อถือของวรรณกรรมทางวิทยาศาสตร์ (scientific literature) บางส่วน ซึ่งอ้างว่านักวิจัยคนอื่นไม่สามารถนำ “การค้นพบ” หลายชิ้นมาทำซ้ำได้ เช่น ข้อโต้แย้งที่ยังดำเนินต่อไปว่าภาษากายที่องอาจซึ่งรู้จักแพร่หลายในชื่อ “ท่าแห่งอำนาจ” (power pose) นั้นอาจกระตุ้นให้เกิดการเปลี่ยนแปลงฮอร์โมนและอื่นๆ[8] การใช้วิธีการทางสถิติพื้นฐานอย่างผิดๆ เป็นต้นเหตุของสิ่งที่เรียกว่า วิกฤตการทำซ้ำหรือการถ่ายซ้ำ (replication) ในวิทยาศาสตร์

เมื่อมีชุดข้อมูลขนาดใหญ่และซอฟต์แวร์วิเคราะห์ที่ใช้งานง่ายมากขึ้น คนอาจคิดว่าการฝึกอบรมเกี่ยวกับวิธีการทางสถิติมีความจำเป็นน้อยลง ความคิดนี้ตื้นเขินอย่างยิ่ง เพราะนอกจากเราจะไม่ได้รับการปลดปล่อยจากความจำเป็นของทักษะทางสถิติแล้ว ฐานข้อมูลที่มีขนาดใหญ่ขึ้นและการศึกษาเชิงวิทยาศาสตร์ที่เพิ่มจำนวนและซับซ้อนยิ่งขึ้นยังทำให้การหาข้อสรุปที่เหมาะสมยากยิ่งขึ้นอีก เมื่อมีข้อมูลมากขึ้น เราย่อมจำเป็นต้องตระหนักมากยิ่งขึ้นว่าที่จริงแล้วหลักฐานนั้นๆ มีคุณค่ามากน้อยเพียงใด

ยกตัวอย่างเช่น การศึกษาชุดข้อมูลจากข้อมูลประจำ (routine data) โดยละเอียดอาจทำให้เกิดการค้นพบผิดพลาด (false discovery) สูงขึ้น ทั้งเนื่องจากความลำเอียงเชิงระบบที่ฝังอยู่ในแหล่งข้อมูล และจากการทำการวิเคราะห์มากมายแต่เลือกรายงานเฉพาะสิ่งที่ดูน่าสนใจที่สุด การกระทำเช่นนี้บางครั้งเรียกว่า “การกรองข้อมูล” (data-dredging) เราควรตระหนักถึงอันตรายของการเลือกรายงาน (selective reporting) ความจำเป็นที่นักวิจัยอิสระสามารถนำคำกล่าวอ้างทางวิทยาศาสตร์มาทำซ้ำ และอันตรายจากการตีความการศึกษานอกเหนือบริบท เพื่อให้สามารถพิจารณ์ผลงานตีพิมพ์ทางวิทยาศาสตร์ โดยเฉพาะอย่างยิ่งรายงานจากสื่อที่มีมากขึ้นซึ่งเราต้องพบเห็นเป็นประจำทุกวัน

ความเข้าใจเหล่านี้สามารถสรุปได้ภายใต้หัวข้อ**ความฉลาดรู้ด้านข้อมูล (data literacy)** ซึ่งกล่าวถึงความสามารถที่จะนำปัญหาในโลกแห่งความเป็นจริงมาวิเคราะห์ทางสถิติ รวมทั้งเข้าใจและพิจารณา

ข้อสรุปใดก็ตามที่คนอื่นเสนอบนพื้นฐานของสถิติ ทว่าการพัฒนาความฉลาดรู้ด้านข้อมูลหมายถึงเราต้องเปลี่ยนวิธีการสอนสถิติเสีย

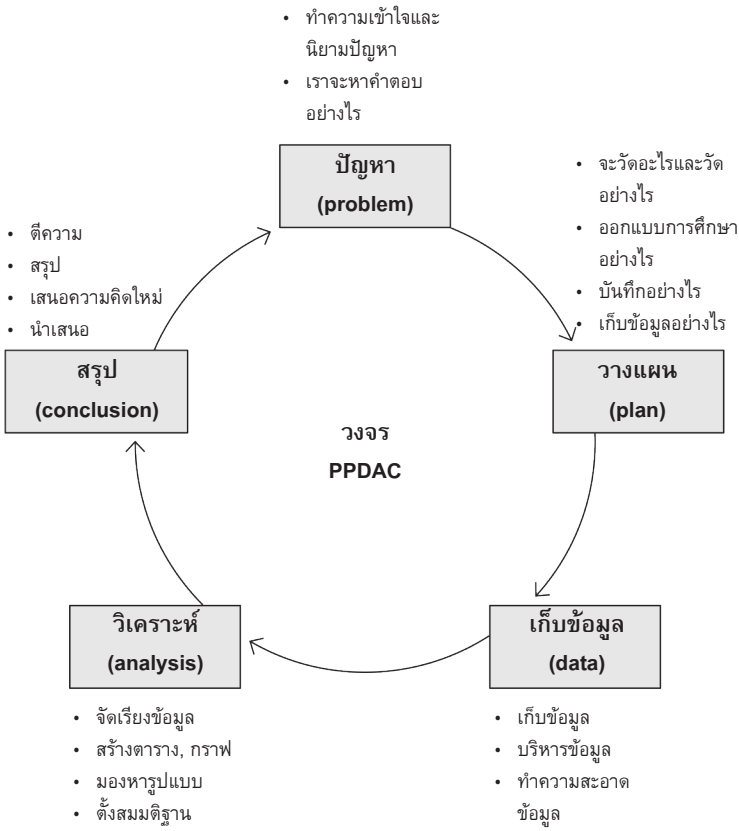
การสอนสถิติ

นักศึกษาหลายรุ่นต้องทนเรียนวิชาสถิติแสนน่าเบื่อซึ่งมีพื้นฐานจากการเรียนรู้เทคนิคชุดหนึ่งเพื่อนำไปประยุกต์ในสถานการณ์ต่าง ๆ โดยเน้นทฤษฎีคณิตศาสตร์มากกว่าความเข้าใจว่าทำไมจึงใช้สูตรคำนวณเหล่านั้นและความท้าทายที่เกิดขึ้นเมื่อพยายามใช้ข้อมูลตอบคำถาม

โชคดีที่เรื่องนี้กำลังเปลี่ยนไป ความจำเป็นของวิทยาศาสตร์ข้อมูลและความฉลาดรู้ด้านข้อมูลผลักดันให้เกิดวิธีการที่ขับเคลื่อนด้วยปัญหา (problem-driven approach) ซึ่งมองว่าการใช้เครื่องมือทางสถิติเฉพาะอย่างนั้นเป็นเพียงองค์ประกอบหนึ่งของวัฏจักรการศึกษาค้นคว้าที่สมบูรณ์ มีการเสนอแผน **PPDAC** เพื่อเป็นวิธีแสดงวงจรการแก้ปัญหาซึ่งเราจะนำมาใช้โดยตลอดในหนังสือเล่มนี้[9] ภาพประกอบ 0.3 แสดงตัวอย่างแผน PPDAC จากนิวซีแลนด์ซึ่งเป็นผู้นำของโลกด้านการเรียนการสอนสถิติในโรงเรียน

ขั้นตอนแรกของวงจรคือการบ่งชี้ปัญหา การศึกษาสถิติมักเริ่มต้นด้วยคำถามเสมอ เช่น การที่เราตั้งคำถามเกี่ยวกับรูปแบบฆาตกรรมของฮาโรลด์ ชิปแมน หรือจำนวนต้นไม้ในโลก ในบทต่อ ๆ ไปเราจะเน้นศึกษาปัญหาหลากหลาย ไม่ว่าจะเป็นประโยชน์ที่คาดหวังจากการรักษาแบบต่าง ๆ หลังการผ่าตัดมะเร็งเต้านม ไปจนถึงคำถามที่ว่าทำไมชายชราจึงมีใบหูใหญ่

คนมักนึกอยากข้ามขั้นตอนการวางแผนโดยละเอียด คำถามเกี่ยวกับฆาตกรรมของชิปแมนต้องการเพียงการเก็บข้อมูลเกี่ยวกับเหยื่อของเขาให้ได้มากที่สุด แต่พวกที่ต้องนับจำนวนต้นไม้นั้นให้ความใส่ใจละเอียดลออกับคำนิยามที่ชัดเจนและวิธีการวัด เนื่องจากจะสรุปอย่าง



ภาพประกอบ 0.3
วงจรแก้ปัญหา PPDAC เริ่มจากปัญหา วางแผน เก็บข้อมูล วิเคราะห์ ไปจนถึงการสรุปและการนำเสนอ แล้วเริ่มต้นวงจรใหม่อีกครั้ง

มั่นใจได้ก็ต่อเมื่อออกแบบการศึกษาอย่างถูกต้อง น่าเสียดายที่เมื่อเร่งรีบเก็บข้อมูลและเริ่มวิเคราะห์ คนมักมองข้ามความสำคัญของการวางแผนไป

การเก็บข้อมูลที่ดีต้องอาศัยทักษะการจัดการและลงรหัสแบบที่วิทยาศาสตร์ข้อมูลให้ความสำคัญยิ่งขึ้น โดยเฉพาะในกรณีที่ข้อมูลจาก

แหล่งประจำอาจต้องทำความสะอาด <data cleansing คือกระบวนการการเตรียมข้อมูลให้พร้อมสำหรับการวิเคราะห์ โดยตรวจหาข้อมูลที่ไม่สมบูรณ์ ไม่เกี่ยวข้อง ซ้ำซ้อน หรือไม่เหมาะสม จากนั้นก็แก้ไขหรือปรับเปลี่ยน> อย่างถี่ถ้วน เพื่อเตรียมพร้อมสำหรับการวิเคราะห์ ระบบการเก็บข้อมูลอาจมีการเปลี่ยนแปลงเมื่อเวลาผ่านไป อาจมีข้อผิดพลาดที่เห็นได้ชัด หรือปัจจัยอื่นๆ คำว่า "ข้อมูลที่พบ" สื่อความว่าอาจเป็นข้อมูลที่ไร้ระเบียบ เทียบเคียงได้กับสิ่งที่เก็บจากท้องถนน

แต่เดิมขั้นตอนการวิเคราะห์เป็นหัวข้อหลักในวิชาสถิติ และเราจะกล่าวถึงเทคนิคการวิเคราะห์หลากหลายรูปแบบในหนังสือเล่มนี้ แต่บางครั้งสิ่งสำคัญอาจเป็นการแสดงภาพอย่างสื่อความหมาย อย่างในภาพประกอบ 0.1 ประการสุดท้าย กุญแจสำคัญสู่สถิติศาสตร์คือการหาข้อสรุปที่เหมาะสมซึ่งกล่าวถึงข้อจำกัดของหลักฐานอย่างรอบคอบและนำเสนออย่างชัดเจน เหมือนการแสดงกราฟข้อมูลคดีของชิปแมน ข้อสรุปใดก็ตามมักนำไปสู่คำถามเพิ่มเติม วงจรนี้จึงเริ่มขึ้นอีกครั้ง ดังเช่นตอนที่เราเริ่มดูช่วงเวลาของวันที่คนไข้ของชิปแมนเสียชีวิต

แม้ในทางปฏิบัติ วงจร PPDAC อาจไม่ได้เป็นไปตามขั้นตอนอย่างเคร่งครัดดังที่แสดงในภาพประกอบ 0.3 แต่วงจรนี้เน้นย้ำว่าเทคนิคการวิเคราะห์สถิติอย่างเป็นกิจจะลักษณะ เป็นเพียงส่วนหนึ่งของงานที่นักสถิติหรือนักวิทยาศาสตร์ข้อมูลทำ สถิติศาสตร์เป็นมากกว่าคณิตศาสตร์แขนงหนึ่งที่ใช้สูตรคำนวณเฉพาะทางที่นักศึกษาหลายต่อหลายรุ่นต้องทุกข์ทน (อย่างไม่เต็มใจนัก)

หนังสือเล่มนี้

ตอนที่ผมเป็นนักศึกษาในบริเตนเมื่อช่วงปี 1970 มีสถานีโทรทัศน์เพียงสามช่อง มีเครื่องคอมพิวเตอร์ขนาดใหญ่กว่าตู้เสื้อผ้าสองเท่า และสิ่งที่ใกล้เคียงกับวิกิพีเดียที่สุดที่เรามีคืออุปกรณ์มือถือในจินตนาการจาก

นิยายเรื่อง *คู่มือท่องกาแล็กซี่ ฉบับนักโบก* (*Hitchhiker's Guide to the Galaxy*) ของดักลาส อดัมส์ (Douglas Adams) เราจึงต้องหันไปพึ่งพาหนังสือชุดเพลิแกน (Pelican books) เพื่อเพิ่มพูนความรู้ สันหนังสือสีฟ้าอันเป็นเอกลักษณ์ของสำนักพิมพ์นี้จึงวางเด่นอยู่บนชั้นหนังสือของนักศึกษาทุกคน

ตอนนั้นผมศึกษาสถิติอยู่ หนังสือชุดเพลิแกนของผมจึงประกอบไปด้วย *ข้อเท็จจริงจากตัวเลข* (*Facts from Figures*) ของเอ็ม. เจ. โมโรนีย์ (M.J. Moroney, 1951) และ *วิธีปั่นหัวคนด้วยสถิติ* (*How to Lie with Statistics*) ของแดร์เรล ฮัฟฟ์ (Darrell Huff, 1954) หนังสือทรงคุณค่าสองเล่มนี้มียอดขายหลายแสนเล่ม ซึ่งสะท้อนถึงระดับความสนใจเรื่องสถิติและสะท้อนว่าเรามีตัวเลือกไม่มากนักในสมัยนั้น หนังสือคลาสสิกทั้งสองยืนหยัดอย่างสง่างามตลอดระยะเวลา 65 ปีที่ผ่านมา แต่ยุคสมัยปัจจุบันเรียกร้องวิธีการสอนสถิติแบบใหม่ตามหลักการที่ผมกล่าวถึงก่อนหน้านี้

ดังนั้นหนังสือเล่มนี้จึงใช้วิธีแก้ปัญหาในโลกแห่งความเป็นจริงเป็นจุดเริ่มต้นในการนำเสนอความคิดทางสถิติ ความคิดบางอย่างอาจฟังดูเหมือนกำปั้นทุบดิน แต่บางอย่างก็ละเอียดอ่อนและต้องออกแรงสมองบ้าง แม้ว่าไม่ต้องใช้ทักษะคณิตศาสตร์ เมื่อเทียบกับตำราแบบดั้งเดิม หนังสือเล่มนี้เน้นประเด็นทางความคิดมากกว่าเนื้อหาเฉพาะทาง และมีสมการง่าย ๆ เพียงไม่กี่สมการ ประกอบกับประมวลศัพท์เพื่อช่วยทำความเข้าใจ แม้ซอฟต์แวร์จะเป็นส่วนสำคัญของทุกงานในวิทยาศาสตร์ข้อมูลและสถิติ แต่ไม่ใช่จุดสำคัญของหนังสือเล่มนี้ มีการสอนภาษาโปรแกรมที่ใช้แพร่หลายและไม่เสียค่าใช้จ่ายอย่างภาษา R และ Python ด้วย

ทุกคำถามที่ผมล้อมกรอบสีเทาไว้อาจหาคำตอบได้จากการวิเคราะห์สถิติในระดับหนึ่ง แม้จะแตกต่างกันในเรื่องขอบเขต บางคำถามเป็นสมมติฐานเชิงวิทยาศาสตร์ที่สำคัญ เช่น อนุภาคฮิกส์โบซอนมีอยู่จริงหรือไม่ หรือมีหลักฐานน่าเชื่อถือที่สนับสนุนพลังพิเศษหรือสัมผัสที่หกหรือไม่ บางคำถามเกี่ยวกับการดูแลสุขภาพ เช่น โรงพยาบาลที่มีผู้ใช้

บริการมากกว่ามีอัตราการรอดชีวิตของผู้ป่วยสูงกว่าหรือไม่ และการตรวจคัดกรองมะเร็งรังไข่มีประโยชน์หรือไม่ บางครั้งเราต้องการเพียงตัวเลขโดยประมาณ เช่น ความเสี่ยงโรคมะเร็งจากแซนด์วิชเบคอน จำนวนคู่นอนทั้งหมดของคนในบริเตนในช่วงชีวิตของพวกเขา และประโยชน์ของการกินยาลดไขมันในเลือด (สแตติน) ทุกวัน

และบางคำถามใส่ไว้ในเล่มนี้เพียงเพราะเป็นประเด็นน่าสนใจ เช่น การระบุตัวผู้รอดชีวิตจากเรือ *ไททานิก* ที่โชคช่วยที่สุด ฮาโรลด์ ชิปแมนอาจถูกจับได้เร็วกว่านี้ไหม และการประเมินความน่าจะเป็นที่ว่าโครงกระดูกซึ่งพบที่ลานจอดรถในเลสเตอร์เป็นของกษัตริย์ริชาร์ดที่สามจริง ๆ

หนังสือเล่มนี้เขียนขึ้นเพื่อนักศึกษาสถิติที่มองหาหนังสือเบื้องต้นอ่านง่ายที่แนะนำประเด็นพื้นฐาน และผู้อ่านทั่วไปที่อยากได้ความรู้เพิ่มเติมเกี่ยวกับสถิติซึ่งพบเห็นทั้งในที่ทำงานและในชีวิตประจำวัน ผมมุ่งเน้นการใช้สถิติอย่างมีทักษะและละเอียดถี่ถ้วน ตัวเลขอาจดูเหมือนข้อเท็จจริงตายตัวดิ้นไม่ได้ แต่ความพยายามที่จะวัดสิ่งต่าง ๆ ไม่ว่าจะเป็นต้นไม้ ความสุข และความตาย ได้แสดงให้เห็นแล้วว่าเราต้องปฏิบัติต่อตัวเลขอย่างละเอียดอ่อนเพียงใด

สถิติอาจมอบความชัดเจนและกระจ่างแจ้งแก่ปัญหาที่เราเผชิญ แต่เราต่างรู้ดีว่าสถิติอาจถูกนำไปใช้ในทางที่ผิด บ่อยครั้งเพื่อสนับสนุนความเห็นหรือเพียงเพื่อดึงดูดความสนใจ ความสามารถที่จะประเมินความน่าเชื่อถือของข้อกล่าวอ้างเชิงสถิตินับเป็นทักษะสำคัญในโลกสมัยใหม่ และผมหวังว่าหนังสือเล่มนี้อาจช่วยมอบความมั่นใจให้ผู้คนในการตั้งคำถามต่อตัวเลขที่พบเห็นในชีวิตประจำวัน

สรุป

- การแปลงประสบการณ์เป็นข้อมูลนั้นไม่ใช่เรื่องง่าย และข้อมูลสามารถบรรยายโลกได้ในขอบเขตจำกัด
- สถิติศาสตร์มีประวัติศาสตร์ความสำเร็จยาวนาน แต่ศาสตร์นี้กำลังเปลี่ยนไปเมื่อข้อมูลมีปริมาณเพิ่มขึ้น
- ทักษะการใช้วิธีการทางสถิติมีบทบาทสำคัญในอาชีพนักวิทยาศาสตร์ข้อมูล
- การสอนสถิติกำลังเปลี่ยนจากรูปแบบที่เน้นวิธีการทางคณิตศาสตร์ไปอิงการแก้ไขปัญหาทั้งวงจร
- วงจร PPDAC ให้กรอบความคิดที่ใช้งานสะดวก อันประกอบด้วยขั้นตอนการระบุปัญหา วางแผน เก็บข้อมูล วิเคราะห์ สรุป และนำเสนอ
- ความฉลาดรู้ด้านข้อมูลเป็นทักษะสำคัญในโลกสมัยใหม่