

**บทความวิชาการ****วิธีการป้องกันและจัดการข้อมูลสูญหายสำหรับการวิเคราะห์ข้อมูลด้านสาธารณสุข**ศักดิ์สิน สิมสินธุ์<sup>1</sup>, ส.ม.**Received:** May 27, 2021**Revised:** June 30, 2021**Accepted:** July 10, 2021**บทคัดย่อ**

ข้อมูลสูญหายเป็นปัญหาที่สามารถพบได้ในทุกรูปแบบการวิจัย ส่งผลทำให้ขนาดตัวอย่างลดลงและลดความสามารถในการเป็นตัวแทนของกลุ่มตัวอย่าง ส่งผลต่อความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ สำหรับการทดสอบสมมติฐานมีผลทำให้อำนาจการทดสอบทางสถิติลดลง เพิ่มความคลาดเคลื่อนชนิดที่ 1

การเพิกเฉยและจัดการข้อมูลสูญหายไม่เหมาะสมสามารถนำสู่ข้อสรุปของผลการศึกษาที่ไม่ถูกต้อง ระดับความรุนแรงของผลกระทบขึ้นอยู่กับขนาดและประเภทของข้อมูลสูญหาย วิธีแก้ปัญหาคือการจัดการข้อมูลสูญหายที่ดีที่สุดคือ การออกแบบระเบียบวิธีวิจัย อย่างไรก็ตามในทางปฏิบัติเป็นการยากที่จะควบคุมการเกิดข้อมูลสูญหายได้ ดังนั้น เมื่อมีข้อมูลสูญหายเกิดขึ้นในชุดข้อมูล การพิจารณาเลือกใช้วิธีการทางสถิติที่เหมาะสมในการจัดการข้อมูลสูญหายจะสามารถช่วยลดปัญหาได้ ปัจจุบันการจัดการข้อมูลสูญหายได้มีการอภิปรายเพิ่มมากขึ้น แต่ในทางปฏิบัติยังไม่มีเปลี่ยนแปลงมากนักยังคงมีการเข้าใจผิดและเลือกวิธีการจัดการที่ไม่เหมาะสมกับชุดข้อมูล ดังนั้นการจัดการข้อมูลสูญหายเป็นเรื่องสำคัญ บทความวิชาการนี้จึงได้ทบทวนสาเหตุ ประเภท และวิธีการจัดการข้อมูลสูญหาย ข้อมูลนี้สามารถใช้เพื่อเป็นแนวทางให้นักวิจัยใช้ในการจัดการเมื่อเกิดข้อมูลสูญหายขึ้นในชุดข้อมูล

**คำสำคัญ:** การจัดการข้อมูลสูญหาย ข้อมูลสูญหาย วิธีการแทนค่าข้อมูลสูญหาย

<sup>1</sup> อาจารย์ สาขาวิชาอนามัยชุมชน คณะสาธารณสุขศาสตร์ มหาวิทยาลัยนครสวรรค์

\* ผู้รับผิดชอบบทความ: ss.simsin@gmail.com

## The method of the prevention and handling of missing data for analysis of public health data

Saksin Simsin<sup>1,\*</sup> M.P.H.

---

### ABSTRACT

Missing data is the trouble that can be encountered in all research. It can reduce the sample size and ability to represent the sample, leading to bias in parameter estimation. For statistical hypothesis testing, it can cause decreasing the power of the test and increasing type 1 error.

The neglected, as well as improper handling of missing data, can lead to invalid conclusions. The severity of the impact depended on the size and the type of missing data. The best solution to this problem was to design the research protocol. However, in practice, it was difficult to control. Therefore, when missing data occur in the data set, selecting an appropriate statistical method for managing the data can help mitigate the problem. In the present, there is an increasing discussion about how to handle missing data, but there has not been much change in practice. Miscomprehension, and improperly selecting handle methods also found. Therefore, handling missing data is important. This article reviews the causes, types and techniques for handling missing data. This information could be used as a guide for researchers in the handling of missing data in the dataset.

**Keywords:** Handling of missing data, Missing data, Imputation of missing data

---

<sup>1</sup> Lecturer, Division of Community Health, Faculty of Public Health, Naresuan University

\* Corresponding author: ss.simsin@gmail.com

## บทนำ

การแพทย์และสาธารณสุขในปัจจุบันได้ให้ความสำคัญในการแก้ปัญหาสุขภาพโดยใช้หลักฐานเชิงประจักษ์ที่ได้จากบทความวิจัยที่ตีพิมพ์ในวารสารทางด้านการแพทย์และสาธารณสุขมากยิ่งขึ้น ประกอบกับมีความตื่นตัวของบุคลากรทางด้านการแพทย์และสาธารณสุขในการทำวิจัยเพื่อแก้ไขปัญหาสุขภาพและพัฒนาคุณภาพการบริการเป็นจำนวนมาก จึงทำให้มีการทำวิจัยและใช้บทความวิจัยอย่างกว้างขวาง เพื่อให้ได้รายงานวิจัยและผลการศึกษานี้ที่ถูกต้องสามารถนำไปใช้ประโยชน์ได้มากที่สุด สิ่งหนึ่งที่สำคัญ คือ การได้มาของข้อมูลที่ต้องการและเพียงพอที่จะทำการวิเคราะห์สรุปผลวิจัย แต่ในขั้นตอนการปฏิบัติจริงกลับพบว่าในกระบวนการเก็บรวบรวมข้อมูลจากกลุ่มตัวอย่างมักเกิดปัญหาข้อมูลสูญหาย (Missing data) ซึ่งหมายถึง การที่ค่าสังเกตของหน่วยตัวอย่างที่สนใจศึกษาได้ขาดหายไป ไม่สามารถทราบค่าได้ โดยที่ค่านั้นควรจะทราบค่าได้ หากวิธีการที่ใช้ในการรวบรวมข้อมูลหรือในการวัดค่ามีประสิทธิภาพดีและเหมาะสม (Prasitwattanaseree & Prasitwattanaseree, 2009) ดังนั้น จึงทำให้เกิดปัญหาข้อมูลไม่เพียงพอ สำหรับการวิเคราะห์ทางสถิติเพื่อตอบคำถามและสรุปผลวิจัย แม้ในการศึกษาที่มีการออกแบบระเบียบวิธีวิจัยและควบคุมเป็นอย่างดี

ปัญหาการเกิดข้อมูลสูญหายนั้นสามารถพบได้ในทุกรูปแบบงานวิจัย ทำให้ขนาดตัวอย่างลดลงลดความสามารถในการเป็นตัวแทนของกลุ่มตัวอย่าง นอกจากนั้นยังทำให้ค่าส่วนเบี่ยงมาตรฐานเพิ่มขึ้นส่งผลให้ช่วงความเชื่อมั่นกว้าง เกิดความเอนเอียงและความคลาดเคลื่อนในการประมาณพารามิเตอร์ สำหรับการวิเคราะห์ทางสถิติเพื่อทดสอบสมมติฐานมีผลทำให้อำนาจการทดสอบทางสถิติ (Power of the test) ลดลงเพิ่มความคลาดเคลื่อนชนิดที่ 1 (Type 1 error) ทำให้การวิเคราะห์ทางสถิติซับซ้อนมากยิ่งขึ้นและอาจนำไปสู่การสรุปผลของการวิจัยที่ไม่ถูกต้อง โดยที่ระดับความรุนแรงของผลกระทบขึ้นอยู่กับขนาดของข้อมูลสูญหายและประเภทของข้อมูลสูญหาย (Srisuttiyakom, 2014)

ข้อมูลสูญหายส่งผลโดยตรงต่อผลลัพธ์การวิจัยแต่นักวิจัยส่วนใหญ่กลับเพิกเฉยหรือมีวิธีการจัดการข้อมูลสูญหายไม่เหมาะสม โดยจากการศึกษาของ Karahalios, Baglietto, Carlin, English, & Simpson (2012) ที่ทำการศึกษาเกี่ยวกับการรายงานและวิธีการจัดการข้อมูลสูญหายในการศึกษาระยะยาวที่มีการวัดและเก็บข้อมูลซ้ำกรณีที่มีขนาดตัวอย่างในการศึกษา มากกว่า 1,000 คน ของบทความวิจัยที่ตีพิมพ์ในวารสารที่อยู่ในฐานข้อมูล PubMed ในระหว่างเดือนมกราคม ค.ศ. 2000 ถึง เดือนธันวาคม ค.ศ. 2009 พบว่ามีบทความวิจัยที่คัดเข้าตามเกณฑ์จำนวน 82 บทความ ในจำนวนนี้มีถึง 14 บทความ (ร้อยละ 17.1) ที่ไม่รายงานรายละเอียดเกี่ยวกับวิธีการจัดการข้อมูลสูญหายและในจำนวน 68 บทความ (ร้อยละ 82.9) ที่รายงานเกี่ยวกับวิธีการจัดการข้อมูลสูญหาย มีมากถึง 54 บทความ (ร้อยละ 79.4) ที่เลือกวิธีการจัดการข้อมูลสูญหายโดยการตัดกลุ่มข้อมูลสูญหายทิ้งซึ่งการตัดกลุ่มข้อมูลสูญหายทิ้งทำให้ขนาดตัวอย่างในการวิเคราะห์ลดลง ส่งผลโดยตรงต่อการประมาณค่าและการทดสอบสมมติฐานทางสถิติ สามารถนำไปสู่ข้อสรุปของผลการศึกษาที่ไม่ถูกต้องและคลาดเคลื่อนจากความเป็นจริง แม้ในปัจจุบันจะมีการกล่าวถึงวิธีการในการจัดการข้อมูลสูญหายมากขึ้น แต่ในทางปฏิบัติยังไม่มีการเปลี่ยนแปลงมากนัก บทความวิชาการนี้จึงได้ทบทวนสาเหตุและแนวทางป้องกันการเกิดข้อมูลสูญหาย ประเภทของข้อมูลสูญหาย ตลอดจนวิธีการจัดการข้อมูลสูญหาย เพื่อเป็นแนวทางให้นักวิจัยใช้ในการจัดการเมื่อเกิดข้อมูลสูญหายขึ้นในชุดข้อมูล

## สาเหตุและแนวทางการป้องกันการเกิดข้อมูลสูญหาย

เมื่อพิจารณาถึงสาเหตุของการเกิดข้อมูลสูญหายพบว่ามีสาเหตุมาจาก 2 แหล่งหลัก คือ เกิดจากระเบียบวิธีวิจัยและจากกลุ่มตัวอย่างหรือผู้ให้ข้อมูล

1. การเกิดข้อมูลสูญหายจากระเบียบวิธีวิจัยและกระบวนการเก็บรวบรวมข้อมูล กระบวนการทำวิจัย

เริ่มจากนักวิจัยตั้งคำถามวิจัยหรือกำหนดประเด็นการวิจัย การที่นักวิจัยตั้งคำถามวิจัยที่ต้องการศึกษาไว้หลากหลายประเด็น ทำให้ต้องสร้างเครื่องมือวิจัยเพื่อเก็บรวบรวมข้อมูลให้ครอบคลุมทุกประเด็นคำถาม ทำให้มีจำนวนข้อคำถามมากและมีความซับซ้อนส่งผลให้ผู้ให้ข้อมูลต้องใช้เวลามากในการตอบคำถามทุกข้อ จึงอาจทำให้เกิดความเบื่อและเหนื่อยล้าในการตอบคำถามไม่สามารถให้ข้อมูลได้ครบทุกประเด็น เกิดเป็นข้อมูลขาดหาย (Sookcharoen, 2015) นักวิจัยควรตั้งคำถามวิจัยหรือกำหนดประเด็นที่ต้องการศึกษาให้ชัดเจนเพียง 2-3 ประเด็น เพื่อที่จะได้มีข้อคำถามในเครื่องมือวิจัยไม่มาก ลดความซับซ้อนของข้อคำถาม และใช้เวลาตอบไม่นานจนเกินไป นอกจากนี้ในการสร้างข้อคำถามไม่ควรใช้คำศัพท์เฉพาะ เพราะจะทำให้ผู้ให้ข้อมูลไม่เข้าใจคำถาม อาจจะไม่ตอบและส่งผลให้เกิดข้อมูลขาดหายได้ เมื่อสร้างเครื่องมือวิจัยแล้ว ก่อนที่จะนำเครื่องมือวิจัยไปใช้ในการเก็บรวบรวมข้อมูลจริงควรมีการตรวจสอบคุณภาพเพื่อปรับให้ข้อคำถามมีความตรงเชิงเนื้อหา (Content validity) และความเชื่อมั่น (Reliability) ในขั้นตอนการลงพื้นที่เพื่อเก็บรวบรวมข้อมูล หากนักวิจัยไม่ได้เป็นผู้เก็บข้อมูลเองควรมีการสอนหรือให้คำแนะนำผู้ช่วยวิจัยถึงวิธีการใช้เครื่องมือวิจัยก่อน และเมื่อได้รับการตอบกลับจากผู้ให้ข้อมูล นักวิจัยต้องมีการตรวจสอบความถูกต้องครบถ้วนทุกข้อคำถามก่อนที่จะนำเข้าสู่ข้อมูลสู่โปรแกรมวิเคราะห์ข้อมูลทางสถิติ การนำเข้าสู่ข้อมูลสู่โปรแกรมวิเคราะห์ข้อมูลทางสถิติเป็นอีกหนึ่งขั้นตอนที่มีความสำคัญ ซึ่งเป็นอีกสาเหตุหนึ่งที่ทำให้เกิดข้อมูลสูญหายอันเนื่องมาจากความผิดพลาดของผู้นำเข้าสู่ข้อมูล ดังนั้น นักวิจัยจึงควรตรวจสอบความถูกต้องของข้อมูลหลังจากนำเข้าสู่ข้อมูลและก่อนทำการวิเคราะห์ข้อมูลเสมอ (Pedersen, Mikkelsen, Cronin-Fenton, Kristensen, Pham, Pedersen et al., 2017)

2. การเกิดข้อมูลสูญหายจากกลุ่มตัวอย่างหรือผู้ให้ข้อมูลอาจเกิดเนื่องจากผู้ให้ข้อมูลไม่เข้าใจวัตถุประสงค์ของการวิจัย ไม่เข้าใจข้อคำถามและมีเวลาจำกัด ทำให้ตอบข้อมูลไม่ครบ ไม่สมบูรณ์ หรือไม่สามารถให้ข้อมูลอย่างต่อเนื่องในกรณีที่เป็นการศึกษาระยะยาว นอกจากนี้การที่ผู้ให้ข้อมูลไม่เข้าใจวัตถุประสงค์

ของการวิจัยที่ชัดเจน อาจจะทำให้เกิดความสงสัยวิตกกังวลและกลัวที่จะให้ข้อมูล โดยเฉพาะบุคคลหรือกลุ่มตัวอย่างที่เปราะบาง อาจกลัวการถูกเปิดเผยข้อมูล ทำให้ไม่ประสงค์ที่จะให้ข้อมูล ดังนั้นเพื่อลดโอกาสเกิดข้อมูลสูญหาย นักวิจัยควรอธิบายและสอบถามความสมัครใจในการให้ข้อมูลก่อนเก็บข้อมูลจากกลุ่มตัวอย่าง (Sookcharoen, 2015)

### ประเภทของข้อมูลสูญหาย

การเข้าใจถึงประเภทและลักษณะของข้อมูลสูญหายมีความสำคัญในการพิจารณาเลือกแนวทางในการจัดการข้อมูลสูญหาย ซึ่งโดยทั่วไปสามารถจำแนกข้อมูลสูญหายออกเป็น 3 ประเภท (Jakobsen, Gluud, Wetterslev, & Winkel, 2017; Dziura, Post, Zhao, Fu, & Peduzzi, 2013; Suwannasaen, 2018) ดังนี้

**1. Missing completely at random** เป็นรูปแบบของข้อมูลสูญหายที่เกิดขึ้นอย่างสุ่ม การขาดหายของค่าสังเกตของตัวแปรที่สนใจศึกษาไม่ขึ้นกับค่าสังเกตของตัวแปรนั้นๆ และตัวแปรอื่นๆ ในฐานข้อมูล นั่นคือค่าความน่าจะเป็นที่จะเกิดการขาดหายแต่ละค่าสังเกตของตัวแปรที่สนใจศึกษาจะมีค่าเท่ากันและเป็นอิสระจากค่าสังเกตอื่นๆ ในฐานข้อมูล ข้อมูลสูญหายแบบ Missing completely at random มีสาเหตุเกิดจากระเบียบวิธีวิจัยและกระบวนการเก็บรวบรวมข้อมูลซึ่งอาจเกิดจากเครื่องมือวิจัยหรืออุปกรณ์ในการวัดมีข้อบกพร่อง ตัวอย่างขาดการติดตามรักษา หรือการนำเข้าสู่ข้อมูลไม่ถูกต้อง

**2. Missing at random** เป็นรูปแบบของข้อมูลสูญหายซึ่งไม่ได้เกิดขึ้นอย่างสุ่มจากค่าสังเกตทั้งหมดแต่เกิดขึ้นอย่างสุ่มภายใต้เงื่อนไขบางค่าสังเกตของตัวแปรตัวอื่นๆ ในฐานข้อมูลซึ่งไม่ได้เป็นตัวแปรที่เกิดข้อมูลสูญหาย นั่นคือค่าความน่าจะเป็นที่จะเกิดข้อมูลสูญหายจะเท่ากันภายในตัวแปรที่ถูกกำหนดเงื่อนไขโดยข้อมูลที่สมบูรณ์ของตัวแปรอื่นๆ ในฐานข้อมูล

**3. Not missing at random** เป็นรูปแบบของข้อมูลสูญหายซึ่งไม่ได้เกิดขึ้นอย่างสุ่ม แต่การเกิดข้อมูลสูญหายขึ้นอยู่กับข้อมูลสูญหายในตัวแปรเดียวกัน และรวมถึงตัวแปรอื่นๆ ในฐานข้อมูล (Srisuttiyakorn, 2014)

ตัวอย่างเช่น การศึกษาเกี่ยวกับรายได้ของบุคคลจากการรวบรวมข้อมูลพบว่า บุคคลที่มีรายได้สูงมากมักจะปฏิเสธที่จะตอบคำถามเกี่ยวกับรายได้และค่าใช้จ่ายมากกว่าบุคคลที่มีรายได้ปานกลาง

## ขั้นตอนการจัดการข้อมูลสูญหาย

วิธีการในการจัดการข้อมูลสูญหายที่ดีที่สุดคือการป้องกันการขาดหายของค่าสังเกตด้วยการออกแบบระเบียบวิธีวิจัยและกระบวนการเก็บรวบรวมข้อมูลที่ดี (Wisniewski, Leon, Otto, & Trivedi, 2006) แต่การปฏิบัติในสถานการณ์จริงเป็นเรื่องยากที่จะควบคุมไม่ให้เกิดข้อมูลสูญหาย ถึงแม้ว่านักวิจัยจะมีกระบวนการป้องกันย่อมเกิดข้อมูลขาดหายได้ เมื่อมีข้อมูลสูญหายเกิดขึ้นมีข้อเสนอแนะในการจัดการกับข้อมูลสูญหายดังนี้

### ขั้นตอนที่ 1 การตรวจสอบความสมบูรณ์ของข้อมูล

การตรวจสอบความสมบูรณ์ของข้อมูลและการระบุข้อมูลสูญหายเป็นขั้นแรกของการจัดการข้อมูลสูญหาย ซึ่งสามารถจัดได้เป็น 2 กลุ่มหลัก

1. ข้อมูลสูญหายที่เกิดจากข้อจำกัดที่สามารถอธิบายเหตุผลได้ เช่น ไม่สามารถติดตามตัวอย่างได้เนื่องจากการโยกย้ายถิ่นฐานหรือเสียชีวิตก่อนสิ้นสุดการศึกษา ข้อมูลสูญหายจะถูกจัดให้เป็นข้อมูลสูญหายแบบสามารถละเอียดได้ ซึ่งเป็นข้อมูลขาดหายที่ไม่ต้องดำเนินการใดๆ

2. ข้อมูลสูญหายที่เกิดจากกระบวนการวิจัยและการเก็บรวบรวมข้อมูล เช่น การที่ผู้ให้ข้อมูลตอบข้อคำถามไม่ครบ การไม่ควบคุมการเก็บข้อมูลให้ครบถ้วน การลงรหัสหรือการบันทึกข้อมูลผิดพลาด ข้อมูลสูญหายจะถูกจัดให้เป็นข้อมูลขาดหายที่ไม่สามารถละเอียดได้ ซึ่งเป็นข้อมูลสูญหายที่สามารถแก้ไขให้เป็นข้อมูลสมบูรณ์ได้ นักวิจัยต้องพิจารณาดำเนินการแก้ไข ไม่สามารถละเอียดได้

### ขั้นตอนที่ 2 การพิจารณาถึงขนาดของข้อมูลสูญหาย

ขนาดของข้อมูลสูญหายมีผลโดยตรงต่อคุณภาพการวิเคราะห์ทางสถิติและผลลัพธ์การวิจัย นักวิจัยต้อง

วิเคราะห์ข้อมูลเพื่อตรวจสอบสัดส่วนของตัวแปรที่มีข้อมูลสูญหาย และจำนวนหน่วยตัวอย่างที่ขาดหายไปในแต่ละตัวแปร โดยมีหลักในการพิจารณาจัดการดังนี้ (Srisuttiyakorn, 2014; Dong & Peng, 2013)

1. ถ้าข้อมูลสูญหายเกิดขึ้นน้อยกว่าร้อยละ 5 ของขนาดตัวอย่างทั้งหมดวิธีการที่แนะนำในการจัดการข้อมูล คือ การตัดข้อมูลแบบวิธี Listwise deletion ซึ่งเป็นวิธีการตัดหน่วยตัวอย่างที่มีค่าข้อมูลสูญหายทั้งหมด การใช้วิธีการแทนค่าสูญหายอาจทำให้เกิดความเอนเอียงในการประมาณค่าพารามิเตอร์และส่งผลต่อข้อสรุปของผลการศึกษา (Jakobsen et al., 2017; Prasitwattanaseree & Prasitwattanaseree, 2009; Srisuttiyakorn, 2014; Dong & Peng, 2013)

2. ถ้าข้อมูลสูญหายเกิดขึ้นระหว่างร้อยละ 5.0 ถึงร้อยละ 10.0 ของขนาดตัวอย่างทั้งหมด ให้พิจารณาว่าข้อมูลสูญหายนั้นเกิดขึ้นในตัวแปรอิสระหรือตัวแปรตาม ถ้าข้อมูลสูญหายนั้นเกิดขึ้นในตัวแปรอิสระ อาจพิจารณาเลือกใช้ตัวแปรอิสระอื่นที่สมบูรณ์และมีเนื้อหาทางทฤษฎีรองรับสามารถใช้อธิบายเพื่อตอบคำถามวิจัยแทนในการวิเคราะห์แทนได้ หรืออาจจะพิจารณาตัดแบบ Listwise deletion เช่น นักวิจัยอาจพิจารณาใช้ตัวแปรนำหน้าในการวิเคราะห์และอธิบายแทนตัวแปรต้นที่มีมวลกายได้ในกรณีที่ตัวแปรต้นมีมวลกายมีข้อมูลสูญหาย ในกรณีที่ข้อมูลสูญหายเกิดขึ้นในตัวแปรตามพิจารณาตัดแบบ Listwise deletion (Srisuttiyakorn, 2014; Dong & Peng, 2013)

3. ถ้าข้อมูลสูญหายเกิดขึ้นระหว่างร้อยละ 10.0 ถึงร้อยละ 20.0 ของขนาดตัวอย่างทั้งหมด นักวิจัยควรพิจารณาแทนที่ค่าข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณ (Imputation) (Srisuttiyakorn, 2014)

4. ถ้าการสูญหายเกิดขึ้นมากกว่าร้อยละ 20.0 ของขนาดตัวอย่างทั้งหมด วิธีการแก้ไขปัญหามูลสูญหายต่างๆ จะมีประสิทธิภาพต่ำ นักวิจัยอาจจะตัดสินใจตัดตัวแปรที่เกิดการสูญหายออกจากการวิเคราะห์หรือพิจารณาเก็บข้อมูลเพิ่มเติมเพื่อทดแทนข้อมูลสูญหายหรือจำนวนตัวอย่างที่ถูกตัดออก แต่การเก็บข้อมูลเพิ่มเติมจะทำให้มีค่าใช้จ่ายเพิ่ม และเสียเวลาในการเก็บรวบรวมข้อมูล (Srisuttiyakorn, 2014)

### ขั้นตอนที่ 3 การพิจารณาเลือกวิธีการจัดการข้อมูลสูญหายตามความเหมาะสม

เป็นขั้นตอนที่นักวิจัยพิจารณาเลือกวิธีการจัดการข้อมูลสูญหายตามความเหมาะสมกับชุดข้อมูลโดยมีวิธีการจัดการข้อมูลสูญหายจะกล่าวถึงต่อไป

#### วิธีการจัดการข้อมูลสูญหาย

1. การตัดกลุ่มข้อมูลสูญหายทิ้ง ซึ่งเป็นวิธีที่ง่ายและสะดวกที่สุด โดยนักวิจัยสามารถทำการวิเคราะห์เฉพาะข้อมูลที่มีความสมบูรณ์เท่านั้น สามารถแบ่งวิธีการจัดการและการวิเคราะห์ข้อมูลได้ 3 แบบ ดังนี้

1.1 วิธี Listwise deletion หรือ Complete case analysis เป็นวิธีการจัดการข้อมูลสูญหายโดยการตัดหน่วยตัวอย่างที่มีค่าสังเกตขาดหายไปจากการวิเคราะห์ทั้งหมด ไม่สนใจข้อมูลสูญหายที่เกิดขึ้นคงเหลือไว้เฉพาะหน่วยตัวอย่างที่มีข้อมูลที่สมบูรณ์และจะทำการวิเคราะห์ข้อมูลจากข้อมูลเฉพาะส่วนที่สมบูรณ์เท่านั้น เป็นวิธีการจัดการข้อมูลสูญหายที่ง่ายที่สุดเหมาะสมในกรณีที่มีข้อมูลสูญหายมีจำนวนน้อยกว่าร้อยละ 5 ของขนาดตัวอย่างทั้งหมด (Jakobsen et al., 2017; Suwannasaen, 2018) วิธีการนี้มักถูกกำหนดให้ใช้เป็นวิธีการหลักสำหรับจัดการกับข้อมูลสูญหายในโปรแกรมวิเคราะห์ทางสถิติต่างๆ ไป หากไม่เจาะจงเลือกใช้วิธีการอื่นในการจัดการกับข้อมูลสูญหายข้อเสียของวิธีนี้ คือ เมื่อตัดข้อมูลสูญหายออกไปทำให้ขนาดตัวอย่างลดลง เช่น ในการศึกษาหนึ่งที่คัดตัวอย่างเข้าในการศึกษาจำนวน 1,000 คน มีตัวแปรที่ทำการเก็บรวบรวมข้อมูลจำนวน 20 ตัวแปร โดยที่แต่ละตัวแปรที่มีข้อมูลสูญหายร้อยละ 5.0 ในกรณีแบบนี้ถ้าจะต้องตัดหน่วยตัวอย่างออกจากการวิเคราะห์ถึง 640 คน และมีข้อมูลที่สมบูรณ์เพียง 360 คน

1.2 วิธี Pairwise deletion เป็นวิธีการจัดการกับข้อมูลสูญหายสำหรับกรณีที่ทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรคู่ โดยจะทำการวิเคราะห์ข้อมูลจากข้อมูลส่วนที่มีค่าสมบูรณ์ทั้งสองตัวแปร วิธีนี้มีข้อดีในส่วนของการใช้งานข้อมูลอย่างเต็มที่และมีประสิทธิภาพแต่มีความซับซ้อนกว่าวิธี Listwise และเสียเวลามากกว่า (Haukoos & Newgard, 2007; Suwannasaen, 2018)

1.3 วิธี Weighted complete-case analysis เป็นวิธีการที่พัฒนามาเพื่อแก้ไขปัญหาที่เกิดจากวิธี Complete case analysis โดยในขั้นตอนของการวิเคราะห์จะมีการถ่วงน้ำหนักที่แตกต่างกันของข้อมูลระหว่างหน่วยตัวอย่างที่มีข้อมูลสมบูรณ์และไม่สมบูรณ์เพื่อปรับลดค่าความเอนเอียงของผลการศึกษาวิธีการนี้เหมาะสำหรับการศึกษาที่มีขนาดตัวอย่างใหญ่ (Haukoos & Newgard, 2007)

2. การแทนที่ค่าข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณ (Imputation) เป็นวิธีการที่ทำการประมาณค่าของค่าสังเกตที่สูญหายและทำการแทนค่าสังเกตที่สูญหายนั้นด้วยค่าประมาณก่อนที่จะทำการวิเคราะห์ข้อมูลในการประมาณค่าสังเกตที่สูญหายจะอาศัยค่าข้อมูลสมบูรณ์และความสัมพันธ์ระหว่างกลุ่มข้อมูล ซึ่งมี 2 วิธีหลัก ดังนี้ (Pedersen et al., 2017)

2.1 Single imputation เป็นวิธีการที่ไม่ซับซ้อนมากนัก โดยการแทนค่าข้อมูลสูญหายด้วยค่าสถิติค่าหนึ่งที่ได้จากค่าข้อมูลสมบูรณ์ของกลุ่มตัวอย่างก่อนที่จะทำการวิเคราะห์ทางสถิติ การแทนทุกค่าสังเกตที่สูญหายในชุดข้อมูลด้วยค่าสถิติเพียง 1 ค่านั้นมีโอกาสถูกต้องต่ำและไม่สะท้อนความจริงของหน่วยตัวอย่างและชุดข้อมูล นอกจากนั้นยังมีผลทำให้ค่าความแปรปรวนลดลง ซึ่งส่งผลต่อช่วงความเชื่อมั่นและการทดสอบทางสถิติ (Fielding, Fayers, Loge, Jordhoy, & Kaasa, 2006; Haukoos & Newgard, 2007; Dziura et al., 2013) นักวิจัยควรพิจารณาเลือกใช้ด้วยความระมัดระวังให้เหมาะสมแล้วแต่กรณี ซึ่งมีหลายวิธี ดังนี้

2.1.1 Mean and Median imputation เป็นวิธีการแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ยหรือค่ามัธยฐานของตัวแปรที่มีข้อมูลสูญหาย โดยคำนวณค่าเฉลี่ยหรือค่ามัธยฐานจากค่าสังเกตที่ทราบค่าในแต่ละตัวแปร นักวิจัยควรเลือกใช้ค่าเฉลี่ยสำหรับตัวแปรที่มีการแจกแจงปกติและใช้ค่ามัธยฐานสำหรับตัวแปรที่มีการแจกแจงแบบอื่น วิธีนี้เป็นวิธีที่ง่ายต่อการจัดการข้อมูลสูญหายไม่ทำให้ขนาดตัวอย่างลดลงเมื่อเปรียบเทียบกับวิธีการตัดกลุ่มตัวอย่างที่มีข้อมูลสูญหายทิ้ง แต่วิธีนี้ไม่ได้เพิ่มข้อสรุปหรือการสื่อสารข้อมูลใหม่ แต่กลับทำให้เกิดการเอนเอียงในการประมาณค่าพารามิเตอร์

และประมาณค่าความแปรปรวนต่ำกว่าความจริงได้ (Haukoos & Newgard, 2007; Kang, 2013; Bennett, 2001)

2.1.2 Regression imputation เป็นวิธีการแทนค่าข้อมูลสูญหายด้วยค่าทำนายจากสมการถดถอย โดยทำการสร้างสมการถดถอยที่กำหนดให้ตัวแปรที่มีข้อมูลสูญหายเป็นตัวแปรตามจากนั้นใช้ค่าสังเกตที่ทราบค่าในแต่ละตัวแปร ที่เก็บรวบรวมได้จากตัวอย่างแทนค่าในสมการถดถอยเพื่อทำนายค่าของข้อมูลที่สูญหายก่อนที่จะทำการวิเคราะห์ทางสถิติมาตรฐานเพื่อสรุปผลการวิจัย (Haukoos & Newgard, 2007; Kang, 2013; Bennett, 2001)

2.1.3 Stochastic regression imputation เป็นวิธีการแทนค่าข้อมูลสูญหายด้วยค่าทำนายจากสมการถดถอย แต่เป็นการแก้ไขข้อบกพร่องหรือข้อจำกัดของ Regression imputation ด้วยการใส่ Residual error หรือ Stochastic error term เข้าไปด้วย เพื่อให้มีความแปรปรวนใกล้เคียงความจริงมากขึ้น (Haukoos & Newgard, 2007)

2.1.4 Hot and cold deck imputation เป็นวิธีการพิจารณาเลือกหน่วยตัวอย่างที่มีลักษณะคล้ายคลึงกับหน่วยตัวอย่างที่เกิดข้อมูลสูญหายมากที่สุด จากนั้นแทนค่าข้อมูลที่สูญหายด้วยค่าข้อมูลที่ทราบค่าของหน่วยตัวอย่างที่คล้ายคลึงนั้น ข้อดีของวิธีนี้คือ มีการวัดความคล้ายคลึงกันของข้อมูล (Haukoos & Newgard, 2007; Bennett, 2001)

2.1.5 Last observation carried forward เป็นวิธีการใช้ข้อมูลที่ค่าสังเกตที่ทราบค่าที่วัดได้ในหน่วยตัวอย่างครั้งสุดท้ายก่อนที่จะเกิดข้อมูลสูญหายของหน่วยตัวอย่างในหน่วยตัวอย่างเดียวกันแทนข้อมูลที่สูญหาย (Haukoos & Newgard, 2007; Kang, 2013; Bennett, 2001)

2.1.6 Worst case analysis เป็นวิธีการที่นิยมใช้ในการจัดการข้อมูลสูญหายในงานวิจัยทางคลินิกที่ตัวแปรตามมีข้อมูลสูญหาย โดยจะแทนค่าข้อมูลสูญหายด้วยค่าที่เกิดเหตุการณ์เลวร้ายที่สุดของตัวแปรตาม เช่น การศึกษาผลการรักษามะเร็งปอดด้วยยาเคมีบำบัดชนิดใหม่ ตัวแปรตามในการศึกษา คือ ผลการรักษาที่

วัดค่าได้ 2 ค่า คือ หายหรือเสียชีวิตภายหลังได้รับการรักษา ในกรณีนี้เมื่อตัวแปรตามเกิดข้อมูลสูญหายจะแทนค่าตัวแปรตามด้วยเสียชีวิตภายหลังได้รับการรักษา เทคนิคการจัดการข้อมูลสูญหายด้วยวิธีนี้ไม่ได้ให้ความสำคัญถึงความถูกต้องของค่าที่แทนข้อมูลสูญหาย แต่แสดงให้เห็นว่าข้อมูลสูญหายไม่ส่งผลกระทบต่อคุณภาพผลการศึกษา (Haukoos & Newgard, 2007)

2.2 Multiple imputation (MI) เป็นเทคนิคที่เริ่มใช้ในช่วงต้นปี ค.ศ. 1970 และนิยมใช้มากขึ้นในปัจจุบัน เนื่องจากมีโปรแกรมทางสถิติรองรับวิธีการวิเคราะห์และสามารถใช้เทคนิคนี้ได้กับข้อมูลหลายชนิด เป็นเทคนิคที่จำลองชุดข้อมูลขึ้นมามากกว่า 1 ชุดข้อมูล ก่อนที่จะแทนค่าข้อมูลสูญหายด้วยค่าสถิติในแต่ละชุดข้อมูล แล้วจึงทำการวิเคราะห์ชุดข้อมูลด้วยวิธีการทางสถิติ ก่อนที่จะสรุปรวมผลการวิเคราะห์จากแต่ละชุดข้อมูลมีทั้งหมด 3 ขั้นตอน ดังนี้ (Jakobsen et al., 2017; Sterne, White, Carlin, Spratt, Royston, Kenward et al., 2009)

1) Imputation step เป็นขั้นตอนการสร้างชุดข้อมูลสูญหายที่ประกอบด้วยกลุ่มของตัวแปรในการศึกษาและแทนค่าข้อมูลสูญหายในแต่ละชุดข้อมูลด้วยค่าทำนายจากโมเดลทางสถิติที่เหมาะสมภายใต้การแจกแจงของค่าสังเกตของแต่ละตัวแปรที่ทราบค่าในขั้นตอนของการสร้างชุดข้อมูลนั้น จากการทบทวนเอกสารพบว่า 5 ชุดข้อมูลได้รับคำแนะนำว่าเพียงพอ แต่ 50 ชุดข้อมูลหรือมากกว่าจะดีกว่าในการลดความแปรปรวนของการสุ่มจากกระบวนการแทนค่าข้อมูล

2) Estimation step เป็นขั้นตอนการวิเคราะห์แต่ละชุดข้อมูลด้วยวิธีการทางสถิติที่เหมาะสมเพื่อประมาณค่าพารามิเตอร์และค่าความแปรปรวนจากข้อมูลแต่ละชุด

3) Pooling step เป็นขั้นตอนการคำนวณรวมค่าพารามิเตอร์และค่าความแปรปรวนจากผลการวิเคราะห์ทั้งภายในชุดข้อมูลและระหว่างชุดข้อมูลทุกชุดข้อมูลเพื่อสรุปผลการศึกษา

Multiple imputation เป็นวิธีการที่มีศักยภาพช่วยให้อาจารย์ได้ผลการศึกษาทางด้านการศึกษาแพทย์และสาธารณสุขที่มีข้อมูลสูญหายมีความถูกต้องมากยิ่งขึ้น เนื่องจากใช้

วิธีการสร้างชุดข้อมูลขึ้นมาหลายชุดและแทนค่าข้อมูลสูญหายในแต่ละชุดข้อมูลด้วยค่าทำนายจากโมเดลทางสถิติที่เหมาะสม วิธีการนี้จะสะท้อนความไม่แน่นอนแบบสุ่มของค่าสังเกตของกลุ่มตัวอย่าง นอกจากนี้ยังคำนึงถึงความสัมพันธ์ของความหลากหลายของกลุ่มตัวอย่างด้วย ความถูกต้องเหมาะสมของโมเดลทางสถิติที่ใช้ทำนายค่าข้อมูลสูญหายควรพิจารณาวิเคราะห์ภายใต้การแจกแจงของค่าสังเกตของแต่ละตัวแปรที่ทราบค่า (Sterne et al., 2009)

### บทสรุปและข้อเสนอแนะเพื่อการนำไปใช้

ข้อมูลสูญหายเป็นปัญหาที่สามารถพบได้ในทุกรูปแบบการวิจัย ส่งผลให้ขนาดตัวอย่างลดลง ลดความสามารถในการเป็นตัวแทนของกลุ่มตัวอย่าง เกิดความเอนเอียงและความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ สำหรับการทดสอบสมมติฐานมีผลทำให้อำนาจการทดสอบทางสถิติลดลง เพิ่มความคลาดเคลื่อนชนิดที่ 1 การที่นักวิจัยเพิกเฉยหรือจัดการไม่เหมาะสมจะส่งผลโดยตรงต่อข้อสรุปของผลการศึกษา โดยที่ระดับความรุนแรงของผลกระทบขึ้นอยู่กับขนาดของข้อมูลสูญหายและประเภทของข้อมูลสูญหาย วิธีการที่ดีที่สุดในการจัดการข้อมูลสูญหาย คือ การป้องกันการเกิดข้อมูลสูญหายด้วยการออกแบบระเบียบวิธีวิจัยและกระบวนการเก็บรวบรวมข้อมูลที่ดี เมื่อมีข้อมูลสูญหายในชุดข้อมูลเกิดขึ้น มีหลายวิธีในการจัดการแก้ไข ซึ่งแต่ละวิธีมีข้อดีและเหมาะสมกับชุดข้อมูลแตกต่างกัน การพิจารณาเลือกใช้วิธีการทางสถิติที่เหมาะสมในการจัดการข้อมูลสูญหายจะช่วยลดปัญหาได้ โดยวิธี Multiple imputation เป็นวิธีการทางสถิติที่มีศักยภาพในการแก้ไขเมื่อเกิดข้อมูลสูญหายขึ้นในชุดข้อมูล เป็นวิธีการที่ไม่ซับซ้อนมากนัก ช่วยให้ผลการศึกษามีความถูกต้องมากยิ่งขึ้นและมีโปรแกรมวิเคราะห์ทางสถิติรองรับจำนวนมากในปัจจุบัน

### เอกสารอ้างอิง

- Bennett, D. A. (2001). How can I deal with missing data in my study?. *Australian and New Zealand journal of public health*, 25(5), 464-469.
- Dong, Y., & Peng, C. Y. (2013). Principled missing data methods for researchers. *Springer Plus*, 2(1), 222. <https://doi.org/10.1186/2193-1801-2-222>
- Dziura, J. D., Post, L. A., Zhao, Q., Fu, Z., & Peduzzi, P. (2013). Strategies for dealing with missing data in clinical trials: From design to analysis. *The Yale Journal of Biology and Medicine*, 86(3), 343-358.
- Fielding, S., Fayers, P. M., Loge, J. H., Jordhoy, M. S., & Kaasa, S. (2006). Methods for handling missing data in palliative care research. *Palliative Medicine*, 20(8), 791-798.
- Haukoos, J. S., & Newgard, C. D. (2007). Advanced statistics: Missing data in clinical research part 1: An introduction and conceptual framework. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*, 14(7), 662-668.
- Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputation be used for handling Missing data in randomised clinical trials—a practical guide with flowcharts. *BMC Medical Research Methodology*, 17(1), 162. <https://doi.org/10.1186/s12874-017-0442-1>
- Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 64(5), 402-406.

- Karahalios, A., Baglietto, L., Carlin, J. B., English, D. R., & Simpson, J. A. (2012). A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. *BMC Medical Research Methodology*, 12, 96. <https://doi.org/10.1186/1471-2288-12-96>
- Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., et al. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology*, 9, 157-166.
- Prasitwattanaseree, P., & Prasitwattanaseree, S. (2009). Missing data and management. *Data Management and Biostatistics Journal*, 4(3), 52-61. (in Thai)
- Sookcharoen, W. (2015). Dealing with missing data. *Romphruek Journal*, 33(2), 11-32. (in Thai)
- Srisuttiyakorn, S. (2014). Missing data analysis. *Journal of Education Studies*, 42(1), 207-223. (in Thai)
- Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., et al. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ*, 338, b2393. <https://doi.org/10.1136/bmj.b2393>
- Suwannasaen, P. (2018). Management approach of missing data: K-Nearest neighbor imputation. *Science and Technology Research Journal Nakhon Ratchasima Rajabhat University*, 4(1), 1-9. (in Thai)
- Wisniewski, S. R., Leon, A. C., Otto, M. W., & Trivedi, M. H. (2006). Prevention of missing data in clinical research studies. *Biological Psychiatry*, 59(11), 997-1000