สสส
สำนักงานกองทุนสนับสนุนการสร้างเสริมสุขภาพ

การวิเคราะห์ข้อมูลสุขภาพ ในระบบสาธารณสุขไทย โดยใช้โปรแกรม R

การวิเคราะห์ข้อมูลสุขภาพ ในระบบสาธารณสุขไทย โดยใช้โปรแกรม R

การวิเคราะห์ข้อมูลสุขภาพ ในระบบสาธารณสุขไทย โดยใช้โปรแกรม R

ผู้เขียน

ดร.กมลทิพย์ วิจิตรสุนทรกุล	สำนักโรคไม่ติดต่อ กรมควบคุมโรค
พญ.พันธนีย์ ธิติชัย	สำนักระบาดวิทยา กรมควบคุมโรค
นพ.กิตติพันธุ์ ฉลอม	โรงพยาบาลเชียงดาว
ดร.อรพันธ์ อันติมานนท์	ศูนย์พัฒนาวิชาการอาชีวอนามัยและสิ่งแวดล้อม จังหวัดสมุทรปราการ
รศ.ดร.นงเยาว์ เกษตร์ภิบาล	คณะพยาบาลศาสตร์ มหาวิทยาลัยเชียงใหม่
ผศ.สพ.ญ.ดร.กรรณิการ์ ณ ลำปาง	คณะสัตวแพทยศาสตร์ มหาวิทยาลัยเชียงใหม่
ดร.นิรันดร์ อินทรัตน์	คณะแพทยศาสตร์ มหาวิทยาลัยมหาสารคาม
ผศ.ดร.เชษฐา งามจรัส	คณะสาธารณสุขศาสตร์ มหาวิทยาลัยขอนแก่น
รศ.ดร.อภิรดี แซ่ลิ่ม	คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสงขลานครินทร์ ปัตตานี

ISBN	978-616-271-574-7
พิมพ์ครั้งที่ 1	จำนวน 200 เล่ม
ปีที่พิมพ์	2563
ผู้จัดพิมพ์	สถาบันวิจัยและพัฒนาสุขภาพภาคใต้ คณะแพทยศาสตร์ มหาวิทยาลัยสงขลานครินทร์
พิมพ์ที่	ไอคิว มีเดีย, สงขลา (089-4660752)

คำนำ

บุคลากรสาธารณสุขถือเป็นตัวแปรสำคัญต่อการพัฒนาสุขภาพประชาชนเป็นอย่างมาก เนื่องจากบุคลากรดังกล่าวมีฐานข้อมูลสุขภาพของประชาชนในพื้นที่และรวบรวมมาอย่างต่อเนื่องหลายสิบปีอยู่ในมือ แต่ยังขาดทักษะด้านสถิติและการวิเคราะห์เพื่อการแปลผลของข้อมูล

สถาบันวิจัยและพัฒนาสุขภาพภาคใต้ คณะแพทยศาสตร์ มหาวิทยาลัยสงขลานครินทร์จึงได้ดำเนินโครงการการสร้างความรู้ด้านระบาดวิทยาของโรคไม่ติดต่อในประเทศไทย ภายใต้การสนับสนุนงบประมาณจากสำนักงานกองทุนสนับสนุนการสร้างเสริมสุขภาพ (สสส.) เพื่อพัฒนาศักยภาพบุคลากรสาธารณสุขด้านการวิเคราะห์ข้อมูล โดยได้ดำเนินการจัดอบรมเชิงปฏิบัติการวิเคราะห์ข้อมูลชุดสุขภาพ 43 แฟ้ม (ซึ่งปัจจุบันมี 52 แฟ้ม) โดยใช้โปรแกรม R ทั่วทั้งประเทศไทย แบ่งเป็นโครงการย่อยแต่ละภูมิภาค ได้แก่ ภาคเหนือ ภาคตะวันออกเฉียงเหนือ ภาคกลาง และภาคใต้ เพื่อให้ครอบคลุมหน่วยงานหลักได้แก่ โรงพยาบาลประจำจังหวัด สำนักงานสาธารณสุขประจำจังหวัด สำนักงานป้องกันควบคุมโรค ทั้ง 12 เขต โดยเชิญเจ้าหน้าที่ไอทีและเจ้าหน้าที่ด้านโรคไม่ติดต่อ รวมแล้ว 367 คน เข้าร่วมโครงการอบรม ทำให้ผู้เข้ารับการอบรมจำนวนมากมีเครื่องมือและทักษะที่สามารถกลับไปวิเคราะห์ข้อมูลด้านสุขภาพที่ตนเองมีอยู่ แล้วพัฒนาการทำงานของตนเองและหน่วยงานได้

จากความสำเร็จดังกล่าว โครงการฯ จึงได้รวบรวมแนวทางการวิเคราะห์ข้อมูลเพื่อให้สะดวกต่อการใช้งาน จึงเกิดเป็น หนังสือการวิเคราะห์ข้อมูลสุขภาพในระบบสาธารณสุขไทยโดยใช้โปรแกรม R ภายใต้ความร่วมมือของวิทยากรทุกภูมิภาค เพื่อต้องการให้บุคลากรสาธารณสุขและบุคคลทั่วไปใช้เป็นคู่มือประกอบการวิเคราะห์ โดยทางโครงการเลือกใช้โปรแกรม R เนื่องจากเป็นโปรแกรมที่มีประสิทธิภาพในการวิเคราะห์ข้อมูลทางสถิติสูง เป็นโปรแกรมที่ใช้งานฟรี (Free software) กล่าวคือเป็นลักษณะที่เป็น open source หมายถึง โปรแกรมที่สามารถนำไปใช้ได้อย่างเสรี โดยสามารถเขียน package ส่งไปยัง R-CRAN เพื่อเพิ่มฟังก์ชันการทำงาน อีกทั้งยังสามารถแขวนไว้ในเว็บไซต์ของโปรแกรม R เพื่อแบ่งปันให้ผู้อื่นได้ใช้งาน ทำให้โปรแกรม R เป็นโปรแกรมที่มีการพัฒนาอย่างไม่หยุดยั้ง เนื้อหาของหนังสือเล่มนี้มีเนื้อหาที่อธิบายถึงโครงสร้างข้อมูล 43 แฟ้ม การติดตั้งโปรแกรม R ความรู้ด้านสถิติและการวิเคราะห์ข้อมูลตั้งแต่การทำความสะอาดข้อมูลจนถึงการวิเคราะห์เชิงตัวแบบ ดังนั้น ผู้ที่ไม่มีพื้นฐานทางสถิติหรือไม่มีพื้นฐานด้านการเขียนภาษาโปรแกรมก็สามารถอ่านและเข้าใจเนื้อหาในหนังสือเล่มนี้ได้ไม่ยาก

ทั้งนี้ ทางโครงการขอขอบคุณผู้เขียนทุกท่านที่ให้ความร่วมมือ ร่วมแรงร่วมใจในการช่วยกันเขียนหนังสือเล่มนี้เพื่อเป็นประโยชน์ต่อสาธารณชน และขอขอบคุณคณะผู้ทรงคุณวุฒิ ดร.วิทย์ วิชัยดิษฐ นางสาวนันท์นภัส พรุเพชรแก้ว นางจิราวรรณ จายุพันธ์ นักวิจัยสังกัดหน่วยระบาดวิทยาและนางฉันทนา เจริญสิน พยาบาลวิชาชีพโรงพยาบาลสงขลานครินทร์ที่ช่วยกรุณาให้ข้อเสนอแนะและตรวจสอบความถูกต้อง และขอขอบคุณหน่วยระบาดวิทยา คณะแพทยศาสตร์ มหาวิทยาลัยสงขลานครินทร์ และคณะผู้ร่วมโครงการทุกท่านที่คอยสนับสนุน ให้กำลังใจและเป็นแรงผลักดันให้หนังสือเล่มนี้สำเร็จลุล่วงไปด้วยดี

ศาสตราจารย์ ดร.นพ.วีระศักดิ์ จงสู่วิวัฒน์วงศ์
สถาบันวิจัยและพัฒนาสุขภาพภาคใต้
คณะแพทยศาสตร์ มหาวิทยาลัยสงขลานครินทร์

สารบัญ

บทที่ 1

ความเป็นมาของ 9 Voluntary Global Targets

บทที่ 1

ความเป็นมาของ 9 Voluntary Global Targets

ดร.กมลทิพย์ วิจิตรสุนทรกุล
E-mail: kamolthipph123@gmail.com

กว่าหกทศวรรษที่ผ่านมา ปัญหาโรคไม่ติดต่อ (Non-communicable diseases: NCD) ได้กลายเป็นปัญหาสุขภาพสำคัญในประเทศพัฒนาแล้ว และช่วงสามทศวรรษที่ผ่านมา แนวโน้มการป่วยและการเสียชีวิตด้วยโรคไม่ติดต่อเพิ่มขึ้นและขยายวงกว้างออกไปทั่วโลก การเพิ่มขึ้นของโรคไม่ติดต่อในประเทศกำลังพัฒนาเป็นการเพิ่มขึ้นของปัญหาจำเป็นต้องรีบแก้ไขควบคู่กับการพัฒนาประเทศและสังคม ทำให้ปัญหาโรคไม่ติดต่อจึงไม่ใช่การเปลี่ยนแปลงเฉพาะระดับบุคคลเท่านั้น แต่ยังเป็นผลมาจากการเปลี่ยนแปลงทางสังคม วัฒนธรรม และสิ่งแวดล้อมที่ส่งผลต่อการดำเนินชีวิตตั้งแต่วัยเด็ก วัยศึกษา วัยทำงาน และวัยสูงอายุ การค้นหาหนทางแก้ไขหรือวิธีการป้องกันควบคุมโรคไม่ติดต่อที่มีประสิทธิภาพ จึงต้องการความเข้าใจปัญหาทางสังคมอย่างถ่องแท้ ปัญหาโรคไม่ติดต่อตามแนวทางการแก้ไขขององค์การระดับประเทศที่สำคัญมีดังนี้

ตารางที่ 1.1 แนวทางการแก้ไขและควบคุมโรคไม่ติดต่อขององค์กรระดับประเทศ

ช่วงเวลา	แนวทางแก้ไขปัญหาโรคไม่ติดต่อ
เดือน พฤษภาคม พ.ศ. 2554	ปัญหาโรคไม่ติดต่อได้รับการแสดงถึงขนาดของปัญหาและความรุนแรงอย่างเร่งด่วนในการประชุมประจำปีขององค์การอนามัยโลก จึงได้จัดทำปฏิญญากรุงมอสโกว่าด้วย NCDs หรือ Moscow Declaration on NCDs ที่ได้รับการรับรองจากรัฐมนตรีว่าการกระทรวงสาธารณสุขประเทศสมาชิก
เดือน สิงหาคม พ.ศ. 2554	หลังจากนั้น 3 เดือน ปัญหาโรคไม่ติดต่อได้ถูกนำเข้าสู่การประชุมผู้นำประเทศขององค์การสหประชาชาติ ผู้นำประเทศได้ร่วมกันประกาศเจตนารมณ์ทางการเมืองต่อองค์การสหประชาชาติ ด้าน NCDs หรือ UN Political Declaration on NCDs ในการจัดการปัญหาโรคไม่ติดต่อ นับว่าโรคไม่ติดต่อเป็นปัญหาสุขภาพที่มีความสำคัญเป็นลำดับที่สามรองจากโรคเอดส์และโรคมาเลเรีย ที่มีการนำเข้าสู่การประชุมขององค์การสหประชาชาติ
เดือน พฤษภาคม พ.ศ. 2556	ต่อจากนั้นอีกประมาณ 2 ปี ในที่ประชุมประจำปีขององค์การอนามัยโลก ประเทศสมาชิกได้รับทราบและรับรองแผนปฏิบัติการเพื่อป้องกันและควบคุมโรคไม่ติดต่อ พ.ศ. 2556-2563 (Global action plan for prevention and control of non-communicable diseases 2013-2020) ที่ความมุ่งหวังให้ประเทศสมาชิกได้ดำเนินการต่อสู้กับปัญหาโรคไม่ติดต่ออย่างจริงจังและกำหนด 9 Voluntary Global Targets ขึ้นเป็นเป้าหมายวัดผลสำเร็จการดำเนินงาน

แผนปฏิบัติการเพื่อป้องกันและควบคุมโรคไม่ติดต่อ พ.ศ. 2556-2563 (Global action plan for prevention and control of non-communicable diseases 2013-2020) ได้กล่าวถึงแนวทางการดำเนินงานป้องกันควบคุมโรคไม่ติดต่อที่สำคัญไว้ 6 ประการ มีสาระสำคัญดังนี้

1. สร้างความร่วมมือในการดำเนินงานและขับเคลื่อนเชิงนโยบายระหว่างประเทศ ภายใต้กรอบดำเนินงานระหว่างประเทศต่างๆ เพื่อสนับสนุนความพยายามในการแก้ไขปัญหาระหว่างกันในทุกด้าน เช่น เทคโนโลยีทางสุขภาพ การพัฒนาข้อมูล งานวิจัย รวมถึงแนวปฏิบัติทางสุขภาพ

2. สร้างการตอบสนองจากทุกส่วนของสังคม ได้แก่ การศึกษา การเกษตร อุตสาหกรรม การสื่อสาร การค้า สิ่งแวดล้อม การเงิน แรงงาน ฯลฯ และภาคประชาสังคม มีการสร้างความแข็งแกร่งร่วมกันในการพัฒนาองค์ความรู้ วิชาการ ความเชี่ยวชาญในการแก้ไขปัญหา รวมถึงพัฒนาศักยภาพขององค์กร สถาบันและกำลังคน

3. ลดปัจจัยเสี่ยงและปัจจัยทางสังคม มีมาตรการส่งเสริมสุขภาพ มีการใช้กฎระเบียบ กฎหมายควบคุมหรือลดอันตรายต่อสุขภาพจากบุหรี่ เครื่องดื่มแอลกอฮอล์ อาหาร เป็นต้น มีกลไกส่งเสริมการตลาดอาหารสุขภาพ ส่งเสริมการมีกิจกรรมทางกาย ครอบคลุมเป้าหมายทุกกลุ่มวัย

4. สร้างความเข้มแข็งของระบบสาธารณสุขและหลักประกันสุขภาพ ประชาชนทุกคนได้รับบริการสุขภาพพื้นฐานอย่างเท่าเทียม ระบบบริการสุขภาพมีประสิทธิภาพในการค้นหาความเจ็บป่วยได้รวดเร็ว มีแผนการดูแลรักษาผู้ป่วยโรคไม่ติดต่อระยะยาว เพิ่มพูนความรู้และทักษะให้เจ้าหน้าที่อย่างสม่ำเสมอ ให้แรงจูงใจและค่าตอบแทนความก้าวหน้าในวิชาชีพ พัฒนาความเชี่ยวชาญซึ่งเป็นทรัพยากรบุคลากรที่สำคัญ

5. ส่งเสริม สนับสนุนงานวิจัยและนวัตกรรมการป้องกันและควบคุมโรคไม่ติดต่อ เพื่อขยายผลการวิจัยสำหรับการตัดสินใจพัฒนานโยบาย การเพิ่มประสิทธิภาพระบบบริการสุขภาพ

6. การเฝ้าระวังและการติดตามผล มีเป้าหมายสำหรับการติดตามการดำเนินงานระดับประเทศและระดับโลก เปรียบเทียบแนวโน้มระหว่างประเทศ ณ ช่วงเวลาต่างๆ พัฒนาศักยภาพในการเฝ้าระวัง คำนึงถึงการพัฒนาเทคโนโลยีและนวัตกรรม เพื่อเพิ่มประสิทธิภาพในการเก็บข้อมูลและการวิเคราะห์ข้อมูล

นับว่าแผนปฏิบัติการเพื่อป้องกันและควบคุมโรคไม่ติดต่อ พ.ศ. 2556-2563 (Global action plan for prevention and control of non-communicable diseases 2013-2020) เป็นแนวทางดำเนินงานระดับประเทศ ซึ่งประเทศสมาชิกจะนำแนวทางมาปรับการดำเนินงานตามบริบทของประเทศสมาชิกให้ประสบความสำเร็จในการลดภาระโรค ค่าใช้จ่ายด้านสุขภาพ และเพิ่มคุณภาพชีวิต

ดังนั้น การกำหนด 9 Voluntary Global Targets ขึ้นจึงเป็นเป้าหมายจากผลการดำเนินงานระดับประเทศ ดังแสดงในตารางที่ 1.2 ซึ่งประเทศไทยต้องให้ความสำคัญและจัดการดำเนินงานที่สอดคล้องในทุกระดับให้บรรลุความสำเร็จ

ตารางที่ 1.2 กรอบ 9 Voluntary Global Targets

เป้าหมายที่	แผนการดำเนินงาน
เป้าหมายที่ 1	ลดการเสียชีวิตก่อนวัยอันควรของ 4 โรคหลัก คือ โรคหัวใจและหลอดเลือด โรคมะเร็ง โรคเบาหวาน และโรคปอดเรื้อรัง
เป้าหมายที่ 2-7	ลดพฤติกรรมเสี่ยงและปัจจัยเสี่ยง การบริโภคเครื่องดื่มแอลกอฮอล์ การบริโภคยาสูบ การกิจกรรมทางกายที่ไม่เพียงพอ การบริโภคโซเดียม ภาวะความดันโลหิตสูง ภาวะอ้วนและน้ำหนักเกิน โรคเบาหวาน
เป้าหมายที่ 8-9	เพิ่มมาตรฐานระบบบริการสุขภาพให้ได้รับยาที่จำเป็น และมีเทคโนโลยีการรักษาขั้นพื้นฐานสำหรับการบริการผู้ป่วยโรคไม่ติดต่อ

การวัดความสำเร็จแผนงานตาม 9 เป้าหมาย ด้วยการกำกับของ 25 ตัวชี้วัดสุขภาพ แบบเลือกได้ สำหรับการเปรียบเทียบและการติดตามแนวโน้ม โดยค่าพื้นฐาน (Baseline) เป็นค่าเป้าหมายตามเกณฑ์บริบทของประเทศสมาชิกในปี 2556 และค่าเป้าหมายตามเกณฑ์ ปี 2568 เป็นปีสิ้นสุด

ตารางที่ 1.3 **9** เป้าหมายโดยสมัครใจ (9 Voluntary Global Targets and 25 Indicators)

ลำดับ	9 เป้าหมายโดยสมัครใจ	25 ตัวชี้วัดขององค์การอนามัยโลก	การรายงานตัวชี้วัดของประเทศไทย/แหล่งข้อมูลที่ใช้รายงาน
1	A 25% relative reduction in risk of premature mortality from cardiovascular disease, cancer, diabetes, or chronic respiratory diseases. ลดการเสียชีวิตก่อนวัยอันควรจากโรคหัวใจและหลอดเลือด มะเร็ง โรคเบาหวานหรือโรคทางเดินหายใจเรื้อรัง	1. โอกาสของความน่าจะเป็นต่อการเสียชีวิตระหว่าง 30-70 ปี แบบไม่มีเงื่อนไข จากโรคหัวใจและหลอดเลือด โรคมะเร็ง โรคเบาหวาน หรือโรคทางเดินหายใจเรื้อรัง (Unconditional probability of dying between 30-70 years from 4 major NCD diseases) (ทางเลือก) 2. อุบัติการณ์ของโรคมะเร็งแต่ละประเภทต่อประชากร 100,000 คน	1. โอกาสของความน่าจะเป็นต่อการเสียชีวิตระหว่าง 30-70 ปี แบบไม่มีเงื่อนไข จากโรคหัวใจและหลอดเลือด โรคมะเร็ง โรคเบาหวาน หรือโรคทางเดินหายใจเรื้อรัง 2. อัตราตายระหว่างกลุ่มอายุ 30-69 ปี จากโรคหัวใจและหลอดเลือด โรคมะเร็ง โรคเบาหวาน หรือโรคทางเดินหายใจเรื้อรัง แหล่งข้อมูล Thai Burden of Diseases
2	At least 10% relative reduction in harmful use of alcohol as appropriate within the national context. การดื่มเครื่องดื่มแอลกอฮอล์แบบ harmful (ดื่มครั้งละมากกว่า 5 แก้วมาตรฐาน) ลดลงอย่างน้อย 10%	3. ปริมาณการบริโภคแอลกอฮอล์ต่อหัว (รวมแอลกอฮอล์ในระบบภาษีและนอกระบบภาษี) ประชากรอายุตั้งแต่ 15 ปีขึ้นไป (หน่วยลิตรของแอลกอฮอล์บริสุทธิ์) 4. ความชุกของการดื่มแอลกอฮอล์อย่างหนัก 5. อัตราการเสียชีวิตและอัตราการป่วยของวัยรุ่นและวัยผู้ใหญ่ที่เกิดจากการดื่มแอลกอฮอล์	3. ความชุกของการดื่มแอลกอฮอล์อย่างหนัก ใน 12 เดือนที่ผ่านมา แหล่งข้อมูล การสำรวจสุขภาพประชาชนโดยการตรวจร่างกาย 4. ปริมาณการบริโภคแอลกอฮอล์ (บริสุทธิ์) ต่อหัว แหล่งข้อมูล กรมศุลกากร กระทรวงการคลัง

ลำดับ	9 เป้าหมายโดยสมัครใจ	25 ตัวชี้วัดขององค์การอนามัยโลก	การรายงานตัวชี้วัดของประเทศไทย/แหล่งข้อมูลที่ใช้รายงาน
3	A 10% relative reduction in prevalence of insufficiency physical activity. ความชุกของผู้มีกิจกรรมทางกายไม่เพียงพอ ลดลง 10%	6. ความชุกของผู้ที่มีกิจกรรมทางกายไม่เพียงพอของวัยรุ่น (หมายถึง การมีกิจกรรมทางกายระดับปานกลางถึงระดับหนักน้อยกว่า 60 นาทีต่อสัปดาห์) 7.ความชุกของผู้มีกิจกรรมทางกายไม่เพียงพอของประชากรอายุ 18 ปีขึ้นไป (หมายถึง การมีกิจกรรมทางกายระดับปานกลางน้อยกว่า 150 นาทีต่อสัปดาห์ หรือมีระดับที่เทียบเท่ากัน)	5. ความชุกของผู้ที่มีกิจกรรมทางกายไม่เพียงพอในประชากรอายุ 15 ปีขึ้นไป แหล่งข้อมูล การสำรวจสุขภาพประชาชนโดยการตรวจร่างกาย
4	A 30% relative reduction in mean of population intake in salt /sodium. ค่าเฉลี่ยการบริโภคเกลือ/โซเดียมของประชากรลดลง 30%	8. ค่าเฉลี่ยประชากรปรับฐานอายุของการบริโภคโซเดียมคลอไรด์ (จำนวนกรัมต่อวัน) ในประชากรอายุ 18 ปี ขึ้นไป	6. ค่าเฉลี่ยประชากรของการบริโภคโซเดียมคลอไรด์ (จำนวนกรัมต่อวัน) ในประชากรอายุ 18 ปีขึ้นไป แหล่งข้อมูล การสำรวจสุขภาพประชาชนโดยการตรวจร่างกาย
5	A 30% relative reduction in prevalence of current tobacco use in person aged 15 years+ ความชุกการบริโภคยาสูบในปัจจุบันของผู้มีอายุ 15 ปีขึ้นไป ลดลง 30%	9. ความชุกของการสูบบุหรี่ในวัยรุ่น 10. ความชุกของการสูบบุหรี่ในประชากรอายุ 18 ปีขึ้นไป	7. ความชุกการสูบบุหรี่ ในประชากรอายุ 18 ปีขึ้นไป แหล่งข้อมูล การสำรวจของสำนักงานสถิติแห่งชาติ
6	A 25% relative reduction in prevalence of raise blood pressure or contain of raise blood pressure according to national circumstance. ความชุกผู้มีความดันโลหิตสูงลดลง 25%	11. ความชุกปรับฐานอายุของผู้มีภาวะความดันโลหิตสูงในประชากรอายุ 18 ปีขึ้นไป (หมายถึง ความดันซิสโตลิกมากกว่า 140 มิลลิเมตรปรอท และ 3/3 หรือความดันไดแอสโตลิก มากกว่า 90 มิลลิเมตรปรอท)	8. ความชุกภาวะความดันโลหิตสูงในประชากรอายุ 18 ปีขึ้นไป แหล่งข้อมูล การสำรวจสุขภาพประชาชนโดยการตรวจร่างกาย

ลำดับ	9 เป้าหมายโดยสมัครใจ	25 ตัวชี้วัดขององค์การอนามัยโลก	การรายงานตัวชี้วัดของประเทศไทย/แหล่งข้อมูลที่ใช้รายงาน
7	Halt the raise in diabetes and obesity. หยุดการเพิ่มขึ้นของโรคเบาหวานและภาวะอ้วน	12. ความชุกปรับฐานอายุของผู้มีภาวะน้ำตาลในเลือดสูง / โรคเบาหวานในประชากรอายุ 18 ปีขึ้นไป (ตรวจพบ Fasting blood sugar มากกว่าหรือเท่ากับ 7.0 mmol/l หรือ 126 mg/dl)	9. ความชุกโรคเบาหวานในประชากรอายุ 18 ปีขึ้นไป 10. ความชุกภาวะน้ำหนักเกินและภาวะอ้วนในประชากรอายุ 18 ปีขึ้นไป แหล่งข้อมูล การสำรวจสุขภาพประชาชนโดยการตรวจร่างกาย
		13. ความชุกภาวะน้ำหนักเกินและภาวะอ้วนในประชากรวัยรุ่น (ตามเกณฑ์ WHO Growth Reference)	
		14. ความชุกปรับฐานภาวะน้ำหนักเกินและภาวะอ้วนในประชากรอายุ 18 ปีขึ้นไป (ภาวะน้ำหนักเกิน (BMI) เท่ากับหรือมากกว่า 25.00 kg/m^2 ภาวะอ้วน BMI เท่ากับหรือมากกว่า 30.00 kg/m^2 ขึ้นไป)	
		(ทางเลือก) 15. ค่าเฉลี่ยปรับฐานอายุของสัดส่วนปริมาณพลังงานจากการบริโภคไขมันอิ่มตัว (Saturated fatty acids) ในประชากรอายุ 18 ปีขึ้นไป	
		(ทางเลือก) 16. ความชุกปรับฐานอายุผู้รับประทานผักและผลไม้น้อยกว่า 5 หน่วยมาตรฐานต่อวัน (400 กรัมต่อวัน)	
		(ทางเลือก) 17. ความชุกปรับฐานอายุผู้มีระดับคลอเลสเตอรอลในเลือดสูงในประชากรอายุ 18 ปีขึ้นไป (ค่าคอเลสเตอรอลรวม มากกว่า 5 mmol/l หรือ 190 mg/dl)	

ลำดับ	9 เป้าหมายโดยสมัครใจ	25 ตัวชี้วัดขององค์การอนามัยโลก	การรายงานตัวชี้วัดของประเทศไทย/แหล่งข้อมูลที่ใช้รายงาน
8	At least 50% of eligible people receives drug therapy and consulting (including glycemic control) to prevent heart attack and stokes. อย่างน้อย 50% ผู้ป่วยได้รับยาและคำแนะนำการปฏิบัติตน (รวมทั้งได้รับคำแนะนำในการควบคุมระดับน้ำตาลในเลือด) เพื่อป้องกันการเกิดโรคหัวใจและโรคหลอดเลือดสมอง	18. สัดส่วนผู้อายุ 40 ปีขึ้นไปที่มีความเสี่ยงต่อการเกิดโรคหัวใจและหลอดเลือดใน 10 ปีข้างหน้า หรือความเสี่ยงมากกว่า 30 % ได้รับการรักษาด้วยยาควบคุมระดับน้ำตาลในเลือด และคำแนะนำในการปฏิบัติตนเพื่อป้องกันการเกิดโรคหัวใจวายเฉียบพลันและโรคหลอดเลือดสมอง	ยังไม่มีการรายงาน
9	An 80% availability of the affordable basic technologies and essential medicine, including generics, required to treat major non-communicable diseases in both public and private facilitates. 80% ของสถานพยาบาลรัฐและเอกชนมีเทคโนโลยีพื้นฐานและยาสำคัญที่ประชาชนสามารถเข้าถึงได้ซึ่งรวมถึงยาสำคัญในบัญชียาหลักเพื่อรักษาโรคไม่ติดต่อ	19.ความสามารถในการจัดยาที่จำเป็นสำหรับการรักษาโรคไม่ติดต่ออย่างมีคุณภาพ มีความปลอดภัยและเกิดประสิทธิภาพ รวมทั้งมีเทคโนโลยีขั้นพื้นฐานในสถานพยาบาลของภาครัฐและเอกชน (ทางเลือก) 20. การเข้าถึงวิธีการรักษาเพื่อบรรเทาอาการปวดด้วย Morphine equivalent ในกลุ่มยา strong opioid analgesic (ยกเว้น การใช้ยา methadone ของผู้เสียชีวิตด้วยโรคมะเร็ง) (ทางเลือก) 21. การประกาศใช้นโยบายระดับประเทศในการควบคุมปริมาณกรดไขมันอิ่มตัวในอาหารและหลีกเลี่ยงการใช้น้ำมันประเภท partially hydrogenated vegetable oils (PHVO) ในกระบวนการผลิตอาหาร	ยังไม่มีการรายงาน

ลำดับ	9 เป้าหมายโดยสมัครใจ	25 ตัวชี้วัดขององค์การอนามัยโลก	การรายงานตัวชี้วัดของประเทศไทย/แหล่งข้อมูลที่ใช้รายงาน
		(ทางเลือก) 22. ความสามารถในการจัดหาวัคซีนป้องกันมะเร็งปากมดลูก (HPV vaccines) ในราคาที่เหมาะสม ทั้งนี้ อยู่กับการดำเนินการนโยบายของประเทศ	
		(ทางเลือก) 23. การประกาศใช้นโยบายเพื่อลดผลกระทบต่อเด็กจากกลยุทธ์ทางการตลาดด้านอาหารและเครื่องดื่มที่ไม่มีแอลกอฮอล์ ที่มีส่วนผสมของไขมันอิ่มตัว ไขมันทรานส์ สารให้ความหวานแทนน้ำตาล และเกลือในปริมาณสูง	
		(ทางเลือก) 24. ความครอบคลุมการได้รับวัคซีนป้องกันโรคตับอักเสบ ครั้งที่ 3 ในกลุ่มเด็กทารก	
		(ทางเลือก) 25. สัดส่วนประชากรหญิงที่มีอายุ 30-49 ปี ซึ่งได้รับการตรวจคัดกรองโรคมะเร็งปากมดลูกอย่างน้อยหนึ่งครั้งหรือถี่มากกว่านี้ หรือในกลุ่มอายุน้อยกว่าหรือมากกว่านี้ ได้รับการตรวจคัดกรองมะเร็งปากมดลูกขึ้นกับนโยบายประเทศต่างๆ	

บทที่ 2

โครงสร้างชุดข้อมูลด้านการแพทย์ และสุขภาพ 43 แฟ้ม

บทที่ 2

โครงสร้างชุดข้อมูลด้านการแพทย์และสุขภาพ 43 แฟ้ม

พญ.พันธนีย์ ธิติชัย และ นพ.กิตติพันธุ์ ฉลอม
E-mail: snookermail@gmail.com

กระทรวงสาธารณสุขได้กำหนดให้สถานพยาบาลในสังกัดกระทรวงสาธารณสุขส่งข้อมูลรายงานกิจกรรมที่สำคัญให้กับกระทรวง ผ่านสำนักงานสาธารณสุขจังหวัดแต่ละแห่งทุกเดือน โดยมีวัตถุประสงค์ให้เกิดการวิเคราะห์ข้อมูลเพื่อให้ทราบสถานการณ์สุขภาพของประชาชนในแต่ละตำบล อำเภอ จังหวัด และใช้ข้อมูลดังกล่าวในการติดตามผลการดำเนินการรักษาโรคและส่งเสริมสุขภาพ รวมถึงการจัดสรรทรัพยากรให้กับสถานพยาบาลได้อย่างมีประสิทธิภาพ

การส่งออกข้อมูลให้กับสำนักงานสาธารณสุขจังหวัดนั้น กำหนดให้ส่งข้อมูลตามมาตรฐาน เพื่อระบุรายละเอียดข้อมูลเฉพาะที่จำเป็น โดยกระทรวงสาธารณสุขได้กำหนดเป็นชุดแฟ้มมาตรฐาน มาตั้งแต่ ปี พ.ศ.2543 เรียกว่า ชุดข้อมูลแฟ้มมาตรฐาน 12 แฟ้ม ต่อมามีการปรับปรุงชุดแฟ้มมาตรฐานหลายครั้ง จนเป็นชุดแฟ้มมาตรฐาน 43 แฟ้ม ด้านการแพทย์และสุขภาพ เริ่มใช้ในปี 2559 โดยในปัจจุบันมีจำนวนชุดแฟ้มมาตรฐานทั้งสิ้น 52 แฟ้ม (version 2.3 ตุลาคม 2560) แต่ส่วนใหญ่ก็มักจะยังคงเรียกกันว่าข้อมูล 43 แฟ้ม

องค์การอนามัยโลกได้กำหนดความสำคัญในการป้องกันและควบคุม NCDs อย่างเร่งด่วนตาม “4 x 4 x 4 model” อันประกอบไปด้วย โรคหลัก 4 โรค ได้แก่ (1) โรคหัวใจและหลอดเลือด (2) โรคเบาหวาน (3) โรคมะเร็ง (4) โรคทางเดินหายใจเรื้อรัง ซึ่งเกิดจากการเปลี่ยนแปลงทางสรีรวิทยาสำคัญ 4 ปัจจัย ได้แก่ (1) ภาวะไขมันในเลือดสูง (2) ภาวะความดันโลหิตสูง (3) ภาวะน้ำตาลในเลือดสูง (4) ภาวะน้ำหนักเกินและอ้วน โดยการเปลี่ยนแปลงดังกล่าว เกิดจากการมีพฤติกรรมที่ไม่เหมาะสม ซึ่งมีปัจจัยเสี่ยงร่วมสำคัญ 4 ปัจจัย ได้แก่ (1) การบริโภคยาสูบ (2) การดื่มเครื่องดื่มแอลกอฮอล์ (3) การบริโภคอาหารที่ไม่เหมาะสม (4) การมีกิจกรรมทางกายไม่เพียงพอ

จากประเด็นข้างต้นจะพบว่าข้อมูลส่วนใหญ่สามารถวิเคราะห์ได้โดยใช้ข้อมูลจากโครงสร้างมาตรฐานข้อมูลด้านการแพทย์และสุขภาพ กระทรวงสาธารณสุข จะมีเพียงพฤติกรรมเสี่ยงด้านกิจกรรมทางกายและการบริโภคอาหาร ที่ไม่มีข้อมูล ส่วนประเด็นอื่นๆ สามารถใช้ข้อมูลตามแฟ้มที่เกี่ยวข้องเพื่อใช้ในการวิเคราะห์สถานการณ์ NCDs ในแต่ละพื้นที่ได้ ไม่ว่าจะเป็นประเด็นด้านการเปลี่ยนแปลงทางสรีรวิทยา และการวินิจฉัยโรคต่างๆ (ตารางที่ 2.1) ทั้งนี้ ผู้เขียนขออธิบายตัวแปรที่สามารถนำมาใช้ในการวิเคราะห์สถานการณ์ NCDs ในแต่ละแฟ้ม โดยกำกับชื่อตัวแปรในวงเล็บ เพื่อให้สะดวกต่อการนำไปใช้งานในขั้นตอนของการส่งออกและวิเคราะห์ข้อมูล

2.1 แฟ้มข้อมูล PERSON

เป็นการบันทึกข้อมูลทั่วไปของประชาชนในเขตรับผิดชอบ และผู้ที่มาใช้บริการ โดยจะมีการสำรวจปีละ 1 ครั้ง และมีการลงทะเบียนผู้ป่วยรายใหม่ หรือปรับแก้ข้อมูลในกรณีที่ผู้ป่วยมารับบริการในแต่ละครั้ง โดยจะมีการบันทึกข้อมูลทั่วไป ได้แก่ รหัสสถานบริการตามมาตรฐานกองยุทธศาสตร์และแผนงาน (HOSPCODE) เลขที่บัตรประชาชน (CID) ทะเบียนบุคคล (PID) ชื่อ (NAME) นามสกุล (LNAME) เพศ (SEX) วันเกิด (BIRTH) สัญชาติ (NATION) และตัวแปรที่สำคัญในการบอกสถานะบุคคล (TYPEAREA) ซึ่งจะแบ่งเป็น 5 กลุ่ม ได้แก่ 1) มีชื่ออยู่ตามทะเบียนบ้านในเขตรับผิดชอบและอยู่จริง 2) มีชื่ออยู่ตามทะเบียนบ้านในเขตรับผิดชอบแต่ตัวไม่อยู่จริง 3) อาศัยอยู่ในเขตรับผิดชอบ (ตามทะเบียนบ้านในเขตรับผิดชอบ) แต่ทะเบียนบ้านอยู่นอกเขตรับผิดชอบ 4) อาศัยอยู่นอกเขตรับผิดชอบและทะเบียนบ้านไม่อยู่ในเขตรับผิดชอบ เข้ามารับบริการ หรือเคยอยู่ในเขตรับผิดชอบ 5) อาศัยในเขตรับผิดชอบแต่ไม่ได้อยู่ตามทะเบียนบ้านในเขตรับผิดชอบ เช่น คนเร่ร่อน ไม่มีที่พักอาศัย เป็นต้น ซึ่งหากต้องการที่จะวิเคราะห์สถานการณ์เฉพาะของพื้นที่ มักจะเลือกข้อมูลเฉพาะประชากรที่ขึ้นทะเบียนในเขตรับผิดชอบของพื้นที่นั้น โดยเลือกเฉพาะผู้ที่มี TYPEAREA = 1 และ 3 เท่านั้น ข้อมูลในแฟ้ม PERSON สามารถนำมาใช้วิเคราะห์เพื่อดูการกระจายของโรคตามสถานพยาบาล เพศ อายุ ได้

ผู้ป่วยแต่ละรายในสถานบริการนั้นๆ จะถูกกำหนดรหัส PID เพื่อใช้สำหรับเชื่อมโยงตัวบุคคลกับแฟ้มอื่นๆ โดยโปรแกรมจะกำหนดรหัสขึ้นเองสำหรับผู้ป่วยแต่ละราย ข้อมูลในแต่ละแฟ้มของฐานข้อมูลจะเชื่อมต่อกันโดยใช้ HOSPCODE และ PID โดยจะเห็นว่าแฟ้ม PERSON ที่มีตัวแปรสำคัญ คือ เลขบัตรประชาชน และชื่อ ที่สามารถระบุตัวบุคคลเฉพาะราย นั่นคือ หากผู้ป่วย 1 รายไปรับบริการมากกว่า 1 สถานพยาบาล การระบุตัวตนผู้ป่วยเฉพาะรายจะต้องมีการเชื่อมกับแฟ้ม PERSON เพื่อใช้ข้อมูล CID จะไม่สามารถระบุจาก HOSPCODE และ PID เพียงอย่างเดียวได้ อย่างไรก็ตามจะต้องมีการตรวจสอบความถูกต้องของข้อมูลก่อนการวิเคราะห์ เนื่องจากพบว่ามีการบันทึกข้อมูล CID หรือ ชื่อ ผิดพลาดอยู่บ้าง และบางรายพบตัวเลขที่ไม่ถูกต้อง อาทิ “0000000000000”

ข้อควรระวังในการใช้ข้อมูล

การประมวลผลตัวชี้วัดโดย Health Data Center หรือ HDC นั้น ตัวชี้วัดที่แสดงผลใน HDC dashboard จะประมวลผลเฉพาะบุคคลที่ TYPEAREA เป็นค่า 1 หรือ 3 ของพื้นที่นั้นๆ เท่านั้น แต่อย่างไรก็ตาม โดยปกติหน่วยบริการจะให้บริการผู้ป่วยทุกรายที่มาโรงพยาบาลโดยไม่เลือกปฏิบัติ จึงมีผู้ป่วยมารับบริการทั้ง 4 TYPEAREA ดังนั้น การเลือกเฉพาะบาง TYPEAREA มาวิเคราะห์ข้อมูลอาจจะทำให้เกิดความคลาดเคลื่อนได้ และจากการเก็บข้อมูลในพื้นที่พบว่า ข้อมูล TYPEAREA ยังมีความคลาดเคลื่อนอยู่พอสมควร เนื่องจากความเข้าใจผิดจากประชาชนที่ถูกสำรวจ จึงทำให้มีการแจ้ง TYPEAREA คลาดเคลื่อน ทำให้เกิดกรณี เช่น ผู้ป่วยบางรายที่เป็น TYPEAREA 1 ของโรงพยาบาล A แต่ไม่เคยมารักษาที่โรงพยาบาล A เลย และมีผู้ป่วยหลายรายที่เป็น TYPEAREA 2 หรือ 4 มารับบริการที่โรงพยาบาล A

2.2 แฟ้มข้อมูล HOME/ADDRESS

แฟ้ม HOME เป็นแฟ้มข้อมูลที่ตั้งของหลังคาเรือนในเขตรับผิดชอบของสถานบริการ โดยบันทึกข้อมูลจากการสำรวจปีละครั้ง และขึ้นทะเบียนเมื่อมีหลังคาเรือนใหม่ ความสำคัญของแฟ้มนี้คือการใช้ข้อมูลที่อยู่ในการพรรณนาทางระบาดวิทยา ในด้านของสถานที่ โดยใช้ข้อมูลจากตัวแปร หมู่ที่ (VILLAGE) ตำบล (TAMBON) อำเภอ (AMPUR) และจังหวัด (CHANGWAT) โดยแต่ละตัวแปรมีรหัส 2 หลัก ตามกำหนดของกรมการปกครอง กรณีไม่ทราบใช้รหัส 99 ทั้งนี้ แฟ้ม HOME มีข้อมูลหลักของประชาชนผู้ที่อาศัยในเขตรับผิดชอบ ส่วนข้อมูลในแฟ้ม ADDRESS บันทึกเฉพาะข้อมูลที่อยู่ตามทะเบียนบ้านเฉพาะกรณีที่ทะเบียนบ้านไม่ตรงกับที่อยู่จริงในเขตรับผิดชอบเท่านั้น ซึ่งแฟ้ม ADDRESS มีตัวแปร หมู่ที่ (VILLAGE) ตำบล (TAMBON) อำเภอ (AMPUR) และจังหวัด (CHANGWAT) เช่นเดียวกัน ดังนั้น ในทางปฏิบัติจะใช้ข้อมูลที่อยู่จากแฟ้ม HOME ซึ่งเป็นการสำรวจที่อยู่จริงของผู้ป่วยแต่ละรายก่อน หากไม่มีข้อมูลจึงจะใช้ข้อมูลจากแฟ้ม ADDRESS โดยเฉพาะอย่างยิ่งในผู้ป่วยที่เป็น TYPEAREA 1 หรือ 3

2.3 แฟ้มข้อมูล DEATH

เป็นแฟ้มข้อมูลประวัติการเสียชีวิตของประชาชนที่อาศัยในเขตรับผิดชอบ และผู้ป่วยที่มารับบริการ โดยจะมีการสำรวจปีละ 1 ครั้ง และบันทึกกรณีมีผู้เสียชีวิตรายใหม่เพิ่มเติมระหว่างปี โดยจะมีตัวแปรที่สำคัญคือรหัสโรคที่เป็นสาเหตุการตาย โดยบันทึกเป็นรหัส ICD-10-TM 6 หลัก สถานที่ตาย (PDEATH 1=ในสถานพยาบาล 2=นอกสถานพยาบาล) และวันที่ตาย (DDEATH) โดยจะมีการบันทึกรหัสโรคตามลำดับตามหนังสือรับรองการตาย (CDEATH_A, CDEATH_B, CDEATH_C, CDEATH_D, ODISEASE) สาเหตุการตายตามหนังสือรับรองการตาย (CDEATH) ทั้งนี้ พบว่าการบันทึกข้อมูลในแฟ้ม DEATH อาจยังไม่ครบถ้วน ครอบคลุมทั้งหมด เนื่องจากยังมีผู้เสียชีวิตส่วนหนึ่ง โดยเฉพาะผู้เสียชีวิตนอกโรงพยาบาล ที่ไม่ได้มีการลงบันทึกในระบบข้อมูลของสถานพยาบาล (hospital information system) จึงไม่มีข้อมูลส่งออกมายังฐานข้อมูล

ข้อควรระวังในการใช้ข้อมูล

ควรระวังในการใช้ข้อมูลจากแฟ้มนี้ในแง่ของผลการรักษา (treatment outcome/result) โดยไม่ได้คำนึงถึงความไม่ครอบคลุมของการลงบันทึกข้อมูล ซึ่งอาจจะทำให้ได้ผลการรักษาที่ดีเกินจริง เนื่องจากขาดข้อมูลผู้เสียชีวิตบางส่วน นอกจากนั้นยังสามารถตรวจสอบข้อมูลการเสียชีวิตได้จากสถานะของผู้ป่วยในแฟ้มอื่นๆ เช่น PERSON (DISCHARGE), CHRONIC (TYPEDISCH), REFER_RESULT (REFER_RESULT), ADMISSION (DISCHSTATUS)

2.4 แฟ้มข้อมูล SERVICE

แฟ้มนี้ประกอบไปด้วยข้อมูลประวัติการมารับบริการแต่ละสถานบริการ ซึ่งจะมีข้อมูลของผู้ป่วยทุกครั้งที่มารับบริการ ในแฟ้มนี้จะมีตัวแปรที่เกี่ยวข้องกับ NCDs คือ ค่าความดันโลหิต (SBP และ DBP) และวันที่ให้บริการ (DATE_SERV) ข้อมูลดังกล่าวนี้สามารถใช้ในกรณีที่ต้องการใช้ข้อมูลความดันโลหิตของผู้ป่วยที่มารับบริการในแต่ละครั้ง โดยที่ไม่จำเป็นว่าผู้ป่วยได้ขึ้นทะเบียนในแฟ้ม CHRONIC หรือไม่ อาทิ การเชื่อมข้อมูลกับแฟ้ม DIAGNOSIS เพื่อดึงข้อมูลความดันโลหิตในผู้ป่วยที่ได้รับการวินิจฉัยด้วยรหัสโรคต่าง ๆ ที่ไม่ได้มีการลงทะเบียนและส่งออกข้อมูลในแฟ้ม CHRONIC

ข้อควรระวังในการใช้ข้อมูล

บางรายที่เข้ารับบริการ เช่น ทำแผล ฉีดยา หรือญาติรับยาแทน อาจจะไม่มีข้อมูล SBP DBP

2.5 แฟ้มข้อมูล DIAGNOSIS (DIAGNOSIS_OPD และ DIAGNOSIS_IPD)

แฟ้ม DIAGNOSIS_OPD เป็นการบันทึกการวินิจฉัยโรคของผู้ป่วยนอก แต่ละครั้งที่ผู้ป่วยมารับบริการประกอบไปด้วยตัวแปรสำคัญ ได้แก่ วันที่ให้บริการ (DATE_SERV) รหัสโรคที่วินิจฉัยตามรูปแบบ ICD-10-TM (DIAGCODE) โดย 1 record จะมีข้อมูลรหัสโรค 1 โรค หากผู้ป่วยถูกวินิจฉัยมากกว่า 1 โรคต่อ 1 ครั้งบริการจะมีข้อมูลมากกว่า 1 record ส่วนในแฟ้ม DIAGNOSIS_IPD จะเป็นข้อมูลการวินิจฉัยโรคของผู้ป่วยใน โดยมีตัวแปรสำคัญเช่นเดียวกันกับ DIAGNOSIS_OPD เพียงแต่ชื่อตัวแปรวันที่ให้บริการจะเป็น DATETIME_ADMIT

ข้อควรระวังในการใช้ข้อมูล

ควรระวังความถูกต้องของข้อมูล ICD-10 เนื่องจากผู้ให้รหัสในระบบอาจจะไม่ได้เป็นผู้ที่รับผิดชอบหรือมีความเชี่ยวชาญด้านการลงรหัสโรคโดยตรง เช่น แพทย์ พยาบาล ทำให้รหัสโรคมีความคลาดเคลื่อน เช่น บางครั้งผู้ป่วยมีภาวะความดันโลหิตสูงแต่ยังไม่ถึงเกณฑ์การวินิจฉัยโรคความดันโลหิตสูง แต่ก็มีการให้รหัส I10 ซึ่งคือโรค Hypertension เนื่องจากแพทย์ไม่ทราบว่าควรลงรหัสเป็น R03 Raised Blood Pressure มากกว่า ในกรณีนี้ก็จะทำให้ มีความชุกของโรคความดันโลหิตสูงมากกว่าปกติได้ นอกจากนั้นการศึกษาภาวะแทรกซ้อนของโรคความดันโลหิตสูงและเบาหวานก็ต้องระวังการใช้รหัส ICD-10 ซึ่งจะเป็นคนละรหัสกับโรคที่ไม่มีภาวะแทรกซ้อน และควรตรวจสอบความถูกต้องของรหัสก่อนวิเคราะห์ข้อมูลทุกครั้ง

2.6 แฟ้มข้อมูล DRUG (DRUG_OPD และ DRUG_IPD)

ข้อมูลการจ่ายยาสำหรับผู้ป่วยก็จะถูกส่งออกแยกเป็น 2 ส่วนเช่นเดียวกับการวินิจฉัย คือ DRUG_OPD สำหรับผู้ป่วยนอก และ DRUG_IPD สำหรับผู้ป่วยใน ทั้งยาที่จ่ายให้ขณะรับการรักษาในโรงพยาบาลและยาที่จ่ายเพื่อกลับไปใช้ต่อที่บ้าน ตัวแปรสำคัญ คือ วัน (DATE_SERV สำหรับ DRUG_OPD และ DATETIME_ADMIT สำหรับ DRUG_IPD) รหัสยา 24 หลักมาตรฐาน (DIDSTD) ข้อมูลจากแฟ้มนี้สามารถนำมาใช้ในกรณีที่ต้องการวิเคราะห์ความครอบคลุมของการรักษาด้วยยา อาทิ การหาความครอบคลุมของการได้รับยาในกลุ่ม aspirin หรือ statin ในกลุ่มผู้ป่วย NCDs

ข้อควรระวังในการใช้ข้อมูล

ควรรวบรวมรหัสยาให้ครอบคลุมยาที่ต้องการศึกษาทั้งหมด เพื่อให้ได้ข้อมูลครบถ้วนที่สุด ยาบางอย่างที่ไม่ได้ซื้อในโรงพยาบาลที่ไปรับการรักษา เช่น ยาที่มีราคาแพง ยาที่ไม่สามารถเบิกจ่ายได้ด้วยสิทธิการรักษานั้นๆ

อาจจะมีการใช้ยาตัวนั้น แต่อาจจะไม่ได้บันทึกในระบบ เนื่องจากเป็นการซื้อยาจากนอกโรงพยาบาล นอกจากนั้นไม่ควรอนุมานว่ายาที่เห็นเป็นยาทั้งหมดที่ผู้ป่วยใช้เสมอ เนื่องจากผู้ป่วยอาจใช้ยาที่รับจากโรงพยาบาลอื่น หรือรับยาไปแต่ไม่ได้ใช้จริง

2.7 แฟ้มข้อมูล CHRONIC

แฟ้มนี้เป็นแฟ้มสะสมข้อมูลผู้ป่วยโรคเรื้อรัง ของผู้ป่วยที่อาศัยในเขตรับผิดชอบ และ/หรือผู้มารับบริการที่คลินิกโรคเรื้อรังของหน่วยบริการ (NCDs Clinic) โดยทั่วไปจะต้องมีการขึ้นทะเบียนผู้ป่วยในระบบข้อมูลของสถานพยาบาล (Hospital information system: HIS) จึงจะมีการส่งออกข้อมูลมายังแฟ้มนี้ หากผู้ป่วยถูกวินิจฉัยด้วย NCDs แต่ไม่ถูกขึ้นทะเบียนในระบบ HIS ข้อมูลที่ถูกส่งออกมายังฐานข้อมูล จะมีเพียงการวินิจฉัยในแฟ้ม DIAGNOSIS เท่านั้น จะไม่มีข้อมูลผู้ป่วยในแฟ้ม CHRONIC

ตัวแปรสำคัญในแฟ้ม CHRONIC ได้แก่ วันที่ตรวจพบครั้งแรก (DATE_DIAG) รหัสวินิจฉัยโรคเรื้อรัง (CHRONIC) วันที่จำหน่าย (DATE_DISCH) และประเภทของการจำหน่าย (TYPEDISCH) ซึ่งโดยปกติแล้วผู้ป่วยที่ถูกวินิจฉัยด้วยโรคเรื้อรังจะถูกจำหน่ายผู้ป่วยเนื่องจากสาเหตุ “ตาย” (รหัส = 02) เท่านั้น

ทั้งนี้ พบว่าข้อมูลตัวแปร DATE_DIAG มีข้อจำกัดในบางโปรแกรม ซึ่งไม่สามารถลงข้อมูลวันที่ย้อนหลัง ทำให้ข้อมูลที่ส่งออกมายังแฟ้ม CHRONIC เป็นวันที่มารับบริการในวันที่ขึ้นทะเบียน ซึ่งในความเป็นจริงต้องมีการถามประวัติผู้ป่วยย้อนหลัง แล้วบันทึกข้อมูลตั้งแต่วันที่ผู้ป่วยถูกวินิจฉัยครั้งแรก แม้ว่าจะมีการแก้ไขปัญหานี้แล้ว แต่ข้อมูลที่ถูกบันทึกไปแล้วยังอาจจะมีความไม่ถูกต้องอยู่บ้าง ทำให้ศูนย์เทคโนโลยีสารสนเทศและการสื่อสาร (Health Data Center, HDC) มีการสร้างตัวแปรขึ้นมาใน Data Exchange เพื่อให้แต่ละพื้นที่สามารถดูข้อมูลการวินิจฉัยแรกสุดที่มีการบันทึกลงในฐานข้อมูล เพื่อให้ได้ข้อมูลการวินิจฉัยแรกที่แม่นยำยิ่งขึ้น

ข้อควรระวังในการใช้ข้อมูล

ผู้ป่วยที่ได้รับการขึ้นทะเบียนในแฟ้มนี้ เป็นเพียงบางส่วนของจำนวนผู้ป่วยทั้งหมดเท่านั้น จึงควรระวังการใช้ข้อมูลในแง่ของการคิดความชุกของโรค นอกจากนั้นข้อมูลตัวแปร DATE_DIAG ซึ่งจะนำมาใช้ในการหาอุบัติการณ์ของโรคอาจจะมีความคาดเคลื่อนทั้งจากตัวโปรแกรมเองและความเข้าใจของเจ้าหน้าที่ในการลงบันทึกข้อมูล จึงควรระมัดระวังหรือมีการตรวจสอบความถูกต้องของข้อมูลก่อนวิเคราะห์

2.8 แฟ้มข้อมูล CHRONICFU

เป็นแฟ้มข้อมูลการตรวจติดตามผู้ป่วยโรคเบาหวาน และความดันโลหิตสูง โดยจะมีการบันทึกข้อมูลผู้ป่วยทุกครั้งที่มารับบริการผู้ป่วยโรคเรื้อรัง (NCDs Clinic) โดยผู้ป่วยจะต้องถูกบันทึกในแฟ้ม CHRONIC ก่อน และจะต้องมีการส่งตรวจในห้องตรวจของคลินิกโรคเบาหวาน หรือ โรคความดันโลหิตสูง จึงจะถูกส่งออกข้อมูลมายังแฟ้ม CHRONICFU ตัวแปรสำคัญในแฟ้มนี้ ได้แก่ วันที่ตรวจ (DATE_SERV) น้ำหนัก (WEIGHT) ส่วนสูง (HEIGHT) เส้นรอบเอว (WAIST_CM) ความดันโลหิตซิสโตลิก (SBP) ความดันโลหิตไดแอสโตลิก (DBP) การตรวจเท้า (FOOT) การตรวจจอประสาทตา (RETINA) และสถานที่ตรวจติดตาม (CHRONICFUPLACE)

การใช้ประโยชน์จากแฟ้มนี้ จะใช้ในการวิเคราะห์สถานการณ์การควบคุมความดันโลหิตในผู้ป่วยโรคความดันโลหิตสูง การคำนวณภาวะอ้วนโดยใช้ดัชนีมวลกาย หรือเส้นรอบเอว และความครอบคลุมของการตรวจภาวะแทรกซ้อน ได้แก่ การตรวจเท้า และการตรวจจอประสาทตาในผู้ป่วยโรคเบาหวาน

2.9 แฟ้มข้อมูล LABFU

เป็นการบันทึกข้อมูลการตรวจทางห้องปฏิบัติการของผู้ป่วยโรคเรื้อรัง โดยมีกลุ่มเป้าหมายคือ ผู้ป่วยโรคเบาหวาน โรคความดันโลหิตสูง และโรคเรื้อรังอื่นๆ ปัจจัยที่เกี่ยวข้อง ได้แก่ วันที่ตรวจ (DATE_SERV) รหัสการตรวจทางห้องปฏิบัติการ (LABTEST) ผลของการตรวจทางห้องปฏิบัติการ (LABRESULT) ส่วนใหญ่จะใช้ข้อมูลจากแฟ้มนี้ในการวิเคราะห์ผลตรวจทางห้องปฏิบัติการ อาทิ ระดับน้ำตาลในเลือดเพื่อใช้ในการประเมินสถานการณ์ควบคุมโรคเบาหวาน ระดับไขมันในเลือดเพื่อประเมินความเสี่ยงของการเกิดโรคหัวใจและหลอดเลือด การตรวจ albumin ในปัสสาวะหรือการตรวจระดับ creatinine เพื่อประเมินภาวะแทรกซ้อนของ NCDs

ตัวอย่าง รหัสการตรวจทางห้องปฏิบัติการ (LABTEST) ที่มักจะได้ใช้ในการวิเคราะห์ข้อมูล NCDs ได้แก่

- 531002 Glucose, serum/plasma การตรวจหาน้ำตาลกลูโคสในซีรั่ม/พลาสม่า
- 531004 Glucose, vein (NPO) การตรวจน้ำตาลในเลือดจากหลอดเลือดดำ โดยไม่อดอาหาร
- 531101 Glucose, semi-quantitative (test strip), whole blood การตรวจหาน้ำตาลกลูโคสกึ่งเชิงปริมาณ (โดยใช้แถบทดสอบ) ในเลือด
- 531102 Glucose, capillaries การตรวจน้ำตาลในเลือดจากเส้นเลือดฝอย โดยไม่อดอาหาร
- 531601 Glycosylated hemoglobin whole blood (HbA1c) การตรวจหา glycosylated hemoglobin ในเลือด
- 546602 Triglycerides, serum/plasma การตรวจหา triglycerides ในซีรั่ม/พลาสม่า
- 541602 Cholesterol, total, serum/plasma การตรวจหาคอเลสเตอรอลทั้งหมดในซีรั่ม/พลาสม่า
- 541202 HDL Cholesterol, serum/plasma การตรวจหาคอเลสเตอรอลชนิด HDL ในซีรั่ม/พลาสม่า
- 541402 LDL Cholesterol, serum/plasma การตรวจหาคอเลสเตอรอลชนิด LDL ในซีรั่ม/พลาสม่า
- 440203 Albumin, urine การตรวจอัลบูมินในปัสสาวะ/ตรวจโปรตีน macroalbumin ในปัสสาวะ (ใน filed ผลการตรวจใส่ค่า 0=negative, 1=trace, 2=positive)
- 440204 Microalbumin protein การตรวจโปรตีน microalbumin ในปัสสาวะ (ใน filed ผลการตรวจใส่ค่า 0=negative, 1=trace, 2=positive)
- 581902 Creatinine, serum/plasma การตรวจหา creatinine ในซีรั่ม/พลาสม่า
- 581903 Creatinine, urine การตรวจหา creatinine ในปัสสาวะ
- 581904 eGFR (CKD-EPI formula) การตรวจหาค่า eGFR (ใช้สูตร CKD-EPI formula)

ข้อควรระวังในการใช้ข้อมูล

อาจจะมีความคลาดเคลื่อนของรหัสการตรวจทางห้องปฏิบัติการ (LABTEST) กับผลของการตรวจทางห้องปฏิบัติการ (LABRESULT) ควรมีการตรวจสอบข้อมูลก่อนการวิเคราะห์

2.10 แฟ้มข้อมูล NCDSCREEN

แฟ้มข้างต้นที่กล่าวมาจะเป็นข้อมูลด้านการรักษาจากสถานพยาบาลเป็นหลัก แต่ในแฟ้ม NCDSCREEN จะเป็นข้อมูลการให้บริการคัดกรองโรคความดันโลหิตสูงและโรคเบาหวาน โดยสถานพยาบาลจะคัดกรองเชิงรุกในประชาชนกลุ่มเป้าหมายที่มีอายุ 35 ปีขึ้นไป ที่ยังไม่ได้เป็นผู้ป่วยโรคความดันโลหิตสูง หรือโรคเบาหวาน ตามแนวทางของกระทรวงสาธารณสุข ปีละ 1 ครั้ง

ตัวแปรสำคัญ ประกอบด้วย วันที่ตรวจ (DATE_SERV) ประวัติสูบบุหรี่ (SMOKE) ประวัติดื่มเครื่องดื่มแอลกอฮอล์ (ALCOHOL) น้ำหนัก (WEIGHT) ส่วนสูง (HEIGHT) เส้นรอบเอว (WAIST_CM) ความดันโลหิตครั้งที่ 1 (SBP_1 และ DBP_1) ความดันโลหิตครั้งที่ 2 (SBP_2 และ DBP_2) ระดับน้ำตาลในเลือด (BSLEVEL) วิธีการตรวจน้ำตาลในเลือด (BSTEST)

ในการบันทึกข้อมูลประวัติสูบบุหรี่ จะให้รหัสดังต่อไปนี้ คือ 1 = ไม่สูบ 2 = สูบนานๆ ครั้ง 3 = สูบเป็นครั้งคราว 4 = สูบเป็นประจำ และ 9 = ไม่ทราบ ส่วนประวัติดื่มเครื่องดื่มแอลกอฮอล์ ให้รหัส คือ 1 = ไม่ดื่ม (ไม่ดื่มในรอบ 12 เดือนที่ผ่านมา) 2 = ดื่มนานๆ ครั้ง (ดื่ม 1-3 วัน/เดือน หรือดื่ม 1-11 วัน/ปี) 3 = ดื่มเป็นครั้งคราว (ดื่ม 1-4 วัน/สัปดาห์) 4 = ดื่มเป็นประจำ (ดื่ม 5-7 วัน/สัปดาห์) และ 9 = ไม่ทราบ

ส่วนรหัสวิธีการตรวจน้ำตาลในเลือด ได้แก่ 1 = ตรวจน้ำตาลในเลือดจากหลอดเลือดดำ หลังอดอาหาร 2 = ตรวจน้ำตาลในเลือดจากหลอดเลือดดำ โดยไม่อดอาหาร 3 = ตรวจน้ำตาลในเลือดจากเส้นเลือดฝอย หลังอดอาหาร 4 = ตรวจน้ำตาลในเลือดจากเส้นเลือดฝอย โดยไม่อดอาหาร และ 9 = ไม่ตรวจน้ำตาลในเลือด

แฟ้มนี้สามารถใช้ประมาณสถานการณ์ NCDs ในชุมชนได้ เนื่องจากการเป็นการเก็บข้อมูลจากการสำรวจในชุมชน นอกจากนั้น ยังสามารถใช้ข้อมูลในการติดตามผู้ป่วยที่คัดกรองผิดปกติ และใช้ประเมินประสิทธิภาพของการคัดกรองได้ เช่น สัดส่วนของผู้ป่วยที่คัดกรองผิดปกติแล้วถูกส่งต่อไปวินิจฉัยและขึ้นทะเบียนที่โรงพยาบาล เป็นต้น

จะเห็นว่าข้อมูลที่มีการส่งออกจากสถานพยาบาลทั่วประเทศ มีข้อมูลเป็นจำนวนมาก แม้ว่าจะพบปัญหาเรื่องความถูกต้องหรือครบถ้วนของข้อมูลอยู่บ้าง แต่ในภาพรวมก็ถือว่ามีความเพียงพอในการนำมาใช้ประเมินสถานการณ์ NCDs ในพื้นที่ ไม่ว่าจะเป็นด้านปัจจัยเสี่ยงด้านพฤติกรรม ปัจจัยเฉพาะบุคคล ภาวะโรค และการเสียชีวิต ซึ่งสามารถแสดงให้เห็นบริบทเฉพาะของพื้นที่เพื่อนำไปใช้ในการวางมาตรการที่เหมาะสมในการควบคุมป้องกัน NCDs นอกจากนี้ การใช้ข้อมูลดังกล่าวในการศึกษาวิเคราะห์ จะทำให้ผู้ศึกษาเข้าใจรายละเอียดของตัวแปร และอาจพบประเด็นในการพัฒนาเรื่องความครบถ้วน และคุณภาพของข้อมูล ซึ่งสามารถนำไปพัฒนาคุณภาพข้อมูล เพื่อให้มีความครบถ้วน ถูกต้อง และสามารถนำผลการวิเคราะห์มาสะท้อนบริบท และปัญหาของพื้นที่ได้ชัดเจนยิ่งขึ้นต่อไป

ข้อควรระวังในการใช้ข้อมูล

อาจจะมีความคาดเคลื่อนของการบันทึกข้อมูล เนื่องจากแฟ้มนี้เป็นส่วนหนึ่งของตัวชี้วัดความครอบคลุมของการคัดกรอง จึงอาจจะทำให้มีการนำเข้าข้อมูลผิดพลาดได้

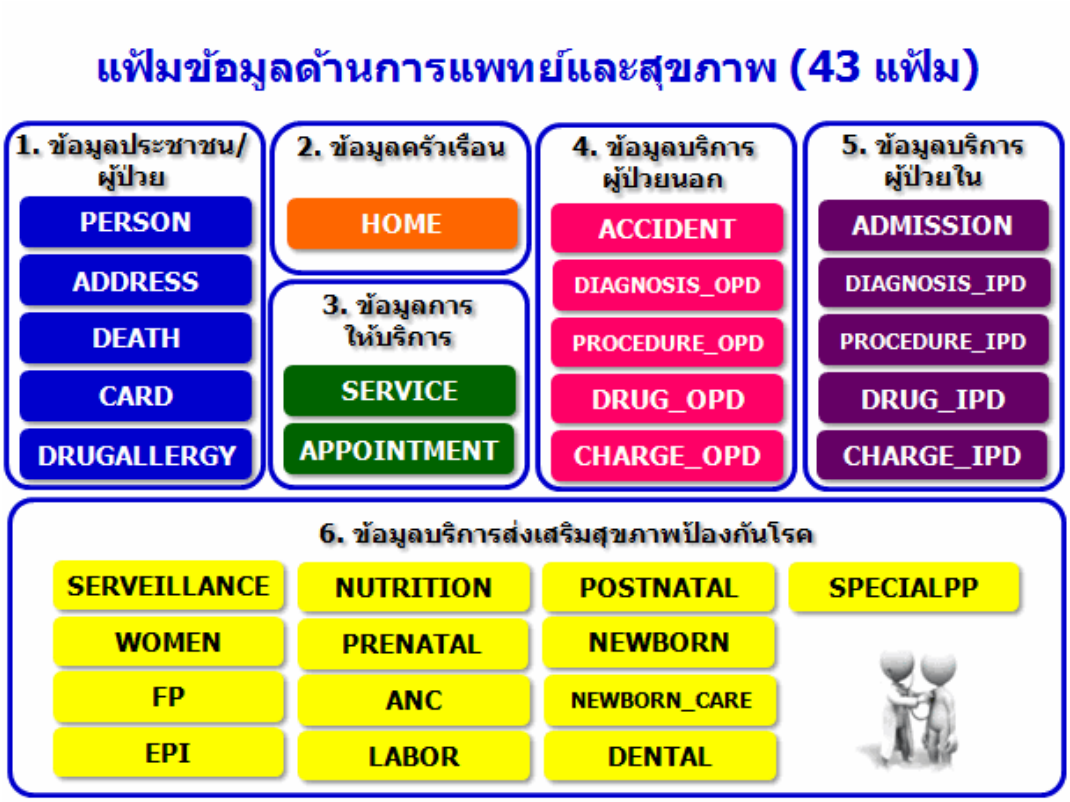
ตารางที่ 2.1 รายชื่อแฟ้มที่เกี่ยวข้องกับการวิเคราะห์ข้อมูล NCDs ในโครงสร้างมาตรฐานข้อมูลด้านการแพทย์และสุขภาพ กระทรวงสาธารณสุข

ประเด็นสุขภาพ	PERSON/ HOME/ADDRESS	DEATH	SERVICE	DIAGNOSIS	DRUG	CHRONIC	CHRONICFU	LABFU	NCDSCREEN
ข้อมูลทั่วไป	√								
พฤติกรรมเสี่ยง									
กิจกรรมทางกาย									
โภชนาการ									√
การสูบบุหรี่									√
การดื่มเครื่องดื่มแอลกอฮอล์									
การเปลี่ยนแปลงทางสรีรวิทยา			√	√	√	√			√
ความดันโลหิตสูง			√	√	√		√		√
น้ำตาลในเลือดสูง			√	√	√		√		√
ไขมันในเลือดสูง			√			√			√
ภาวะอ้วน									
โรคสำคัญ		√	√	√	√				
โรคหลอดเลือดหัวใจ		√	√	√	√				
โรคเบาหวาน		√	√	√					
โรคมะเร็ง		√	√	√	√				
โรคทางเดินหายใจเรื้อรัง									

ตารางที่ 2.2 ตัวแปรที่เกี่ยวข้องกับการวิเคราะห์ข้อมูล NCDs ในโครงสร้างมาตรฐานข้อมูลด้านการแพทย์และสุขภาพ กระทรวงสาธารณสุข

ลำดับ	ชื่อแฟ้ม	ตัวแปรที่เกี่ยวข้อง
1	PERSON	CID, NAME, LNAME, SEX, BIRTH, NATION, TYPEAREA
2	HOME	VILLAGE, TAMBON, AMPUR, CHANGWAT
	ADDRESS	VILLAGE, TAMBON, AMPUR, CHANGWAT
3	DEATH	DDEATH, CDEATH_A, CDEATH_B, CDEATH_C, CDEATH_D, ODISEASE, CDEATH, PDEATH
4	SERVICE	DATE_SERV, SBP, DBP
5	DIAGNOSIS_OPD	DATE_SERV, DIAGCODE
	DIAGNOSIS_IPD	DATETIME_ADMIT, DIAGCODE
6	DRUG_OPD	DATE_SERV, DIDSTD
	DRUG_IPD	DATETIME_ADMIT, DIDSTD
7	CHRONIC	DATE_DIAG, CHRONIC, DATE_DISCH, TYPEDISCH
8	CHRONICFU	DATE_SERV, WEIGHT, HEIGHT, WAIST_CM, SBP, DBP, FOOT, RETINA, CHRONICFUPLACE
9	LABFU	DATE_SERV, LABTEST, LABRESULT
10	NCDSCREEN	DATE_SERV, SMOKE, ALCOHOL, WEIGHT, HEIGHT, WAIST_CM, SBP_1, DBP_1, SBP_2, DBP_2, BSLEVEL, BSTEST

*ตัวแปรหลักที่ใช้ในการเชื่อมข้อมูลระหว่างแฟ้ม ได้แก่ HOSPCODE และ PID



รูปที่ 2.1 โครงสร้างข้อมูล 43 แฟ้ม แฟ้มที่ 1-31

รูปที่ 2.2 โครงสร้างข้อมูล 43 แฟ้ม แฟ้มที่ 32-43

2.11 บรรณานุกรม

1. อรรถเกียรติ กาญจนพิบูลวงศ์, บรรณาธิการ. (2560). รายงานสถานการณ์โรค NCDs ฉบับที่ 2. นนทบุรี : สำนักงานพัฒนานโยบายสุขภาพระหว่างประเทศ.
2. วรรษา เปาอินทร์, มะลิวัลย์ ยืนยงสุวรรณ, บรรณาธิการ. (2559). มาตรฐานการส่งออกข้อมูลตามแฟ้มมาตรฐานกระทรวงสาธารณสุข. กรุงเทพฯ : สำนักกิจการโรงพิมพ์ องค์การสงเคราะห์ทหารผ่านศึกในพระบรมราชูปถัมภ์.

บทที่ 3

การใช้งานโปรแกรม R เบื้องต้น

บทที่ 3

การใช้งานโปรแกรม R เบื้องต้น

ดร.อรพันธ์ อันติมานนท์
E-mail: untimanon99@hotmail.com

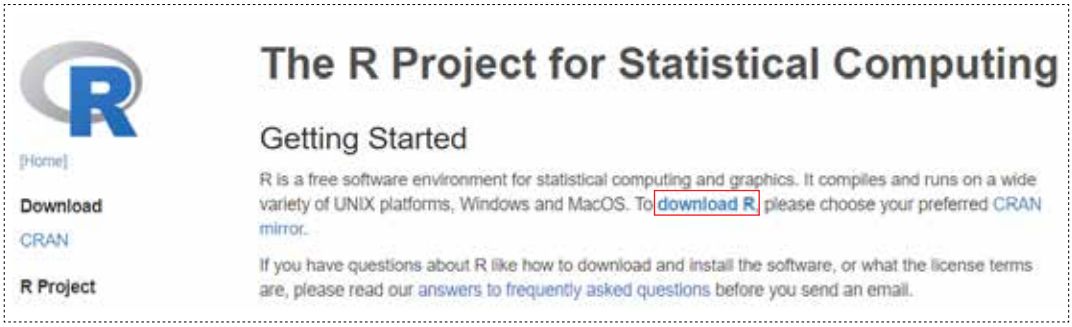
3.1 ความสำคัญของ R

R เป็นภาษาโปรแกรมที่ถูกออกแบบมาเพื่อใช้ในการวิเคราะห์ข้อมูลทางสถิติ และนำเสนอข้อมูลเป็นกราฟ รวมทั้งสามารถนำมาวิเคราะห์ข้อมูลขนาดใหญ่ (Big data) หรืองานที่ซับซ้อนได้ เช่น แผนที่ทางภูมิศาสตร์ที่สำคัญ R เป็นโปรแกรมฟรี ไม่มีค่าใช้จ่าย จึงเป็นโปรแกรมที่ได้รับความนิยมในวงวิชาการมากขึ้นเรื่อยๆ

R เป็นโปรแกรมประเภท Open source ทำให้มีการร่วมพัฒนาต่อยอดให้มีความสามารถที่หลากหลายอย่างต่อเนื่อง สามารถใช้งานได้ทั้งระบบปฏิบัติการ Windows Mac OS และ Linux โดยการทำงานของ R เป็นที่นิยม เนื่องจากมี built-in function ทางด้านสถิติจำนวนมาก รวมถึงมีสถิติที่ใช้ในการวิเคราะห์ข้อมูลขนาดใหญ่ และสามารถแสดงผลการทำงานในรูปแบบกราฟแบบต่างๆ ได้อย่างสวยงาม ซึ่งปัจจุบันโปรแกรม R มี package กว่า 10,000 package จากเหตุผลต่างๆ ที่ได้กล่าวมาข้างต้น จึงทำให้ความนิยมในการใช้โปรแกรม R เพิ่มขึ้นอย่างต่อเนื่อง

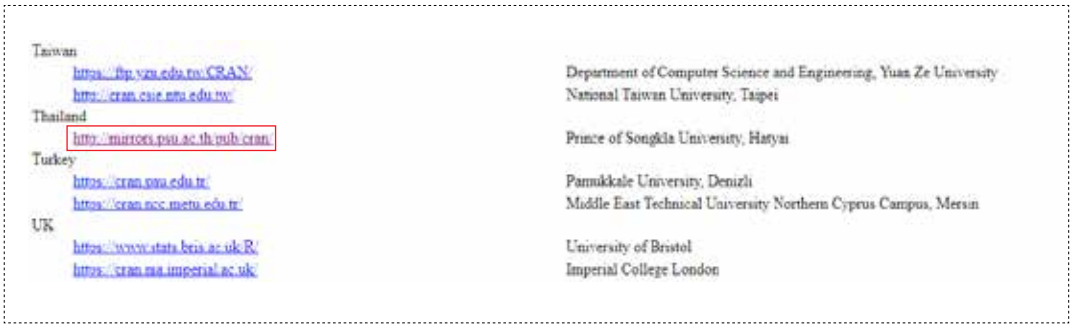
3.2 การติดตั้งโปรแกรม R

3.2.1 เปิดเว็บไซต์ https://www.r-project.org และคลิก download R ตามภาพที่ 3.1



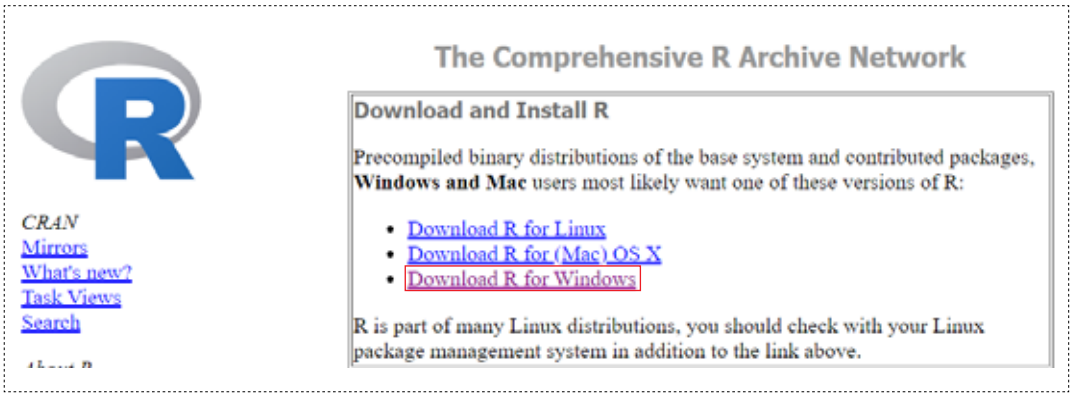
ภาพที่ 3.1 เว็บไซต์สำหรับดาวน์โหลด R

3.2.2 เลือก CRAN Mirrors เพื่อ download เลือก Thailand, Prince of Songkla University, Hatyai ตามภาพที่ 3.2



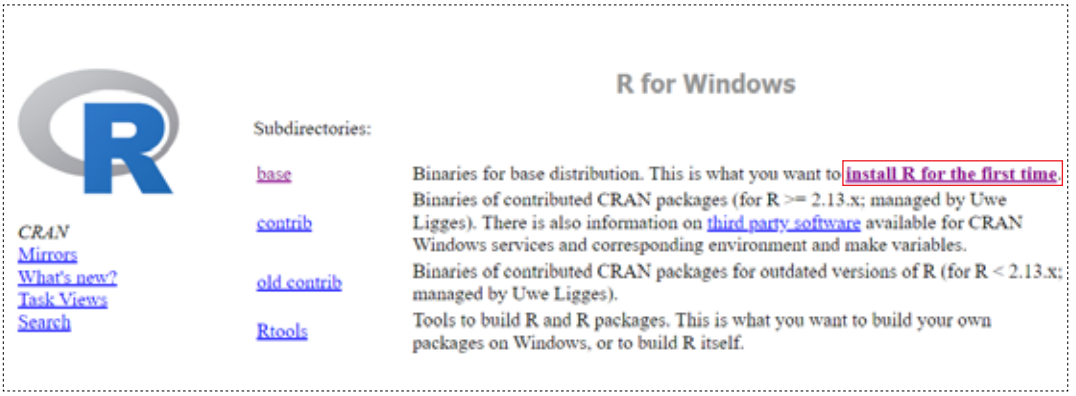
ภาพที่ 3.2 CRAN Mirrors สำหรับดาวน์โหลด R

3.2.3 เลือกระบบปฏิบัติการของคอมพิวเตอร์ ตัวอย่างนี้เลือก Download R for Windows ตามภาพที่ 3.3



ภาพที่ 3.3 ระบบปฏิบัติการที่ต้องการติดตั้ง R

3.2.4 เลือก install R for the first time ตามภาพที่ 3.4



ภาพที่ 3.4 การติดตั้ง R สำหรับ Windows

3.2.5 เลือก Download R 3.6.1 for Windows (เวอร์ชั่น ณ วันที่ 10 เดือน พฤศจิกายน 2562) ตามภาพที่ 3.5

ภาพที่ 3.5 โปรแกรม R สำหรับการดาวน์โหลด

3.2.6 สำหรับระบบปฏิบัติการ windows หลังจากดาวน์โหลดโปรแกรม R ตัวติดตั้งจะอยู่ในโฟลเดอร์ Downloads ไฟล์ที่ดาวน์โหลดมีชื่อว่า R-3.6.1-win32.exe ซึ่งเมื่อ double click ระบบจะแสดงหน้าจอการติดตั้งโปรแกรม ให้กด Next ไปเรื่อยๆ และกดปุ่ม Finish เมื่อการติดตั้งเสร็จสิ้น

3.3 การติดตั้งโปรแกรม R-studio

Rstudio เป็นโปรแกรมหนึ่งที่ช่วยให้ผู้ใช้งานมีความเข้าใจการทำงานของโปรแกรม R ได้ดีขึ้น และทำให้การทำงานในโปรแกรม R มีความสะดวกและง่ายขึ้น การติดตั้งโปรแกรม Rstudio จะต้องติดตั้งโปรแกรม R ก่อนเสมอ หากไม่ได้ติดตั้งก่อน โปรแกรม Rstudio จะใช้งานไม่ได้ โดยการดาวน์โหลดโปรแกรม Rstudio สามารถทำได้ตามขั้นตอนต่อไปนี้

3.3.1 เปิดเว็บไซต์ https://rstudio.com/ และคลิก DOWNLOAD RSTUDIO ตามภาพที่ 3.6

ภาพที่ 3.6 เว็บไซต์สำหรับดาวน์โหลด Rstudio

3.3.2 เลื่อนลงมาที่ Installers for Supported Platforms สำหรับระบบปฏิบัติการ Windows ให้เลือก RStudio 1.2.5019 - Windows 10/8/7 (64-bit) (เวอร์ชั่น ณ วันที่ 11 เดือน พฤศจิกายน 2562) ตามภาพที่ 3.7

Installers for Supported Platforms

Installers	Size	Date	MD5
RStudio 1.2.5019 - Ubuntu 18/Debian 10 (64-bit)	106.04 MB	2019-11-01	a6c9af3d8b1621eb155d23c879c1a75a
RStudio 1.2.5019 - Debian 9 (64-bit)	106.39 MB	2019-11-01	bc7b0b25b41e39fb6f1aefa74163a133
RStudio 1.2.5019 - Fedora 28/Red Hat 8 (64-bit)	120.89 MB	2019-11-01	2291b1befb02622b3aa02c43638ee5c2
RStudio 1.2.5019 - macOS 10.12+ (64-bit)	126.88 MB	2019-11-01	55738355277e8ec660e628acaf2a401b
RStudio 1.2.5019 - SLES/OpenSUSE 12 (64-bit)	99.04 MB	2019-11-01	3bcbf47f40944cc4a5ef4f6fb42319c1
RStudio 1.2.5019 - OpenSUSE 15 (64-bit)	107.09 MB	2019-11-01	29d07b198b7aac92356f8487911efbfa
RStudio 1.2.5019 - Fedora 19/Red Hat 7 (64-bit)	120.26 MB	2019-11-01	dab1cb5f0ed39f5bcf0c795e2938fa94
RStudio 1.2.5019 - Ubuntu 14/Debian 8 (64-bit)	96.93 MB	2019-11-01	f86811fce50b48850fed259d6ce7ef13
RStudio 1.2.5019 - Windows 10/8/7 (64-bit)	149.82 MB	2019-11-01	4d6521a9b89d70c3bf50414c8b6708f2
RStudio 1.2.5019 - Ubuntu 16 (64-bit)	104.91 MB	2019-11-01	67d5a2c255f2bc1a171c7e417853102c

ภาพที่ 3.7 ระบบปฏิบัติการและ Rstudio ที่ให้ดาวน์โหลด

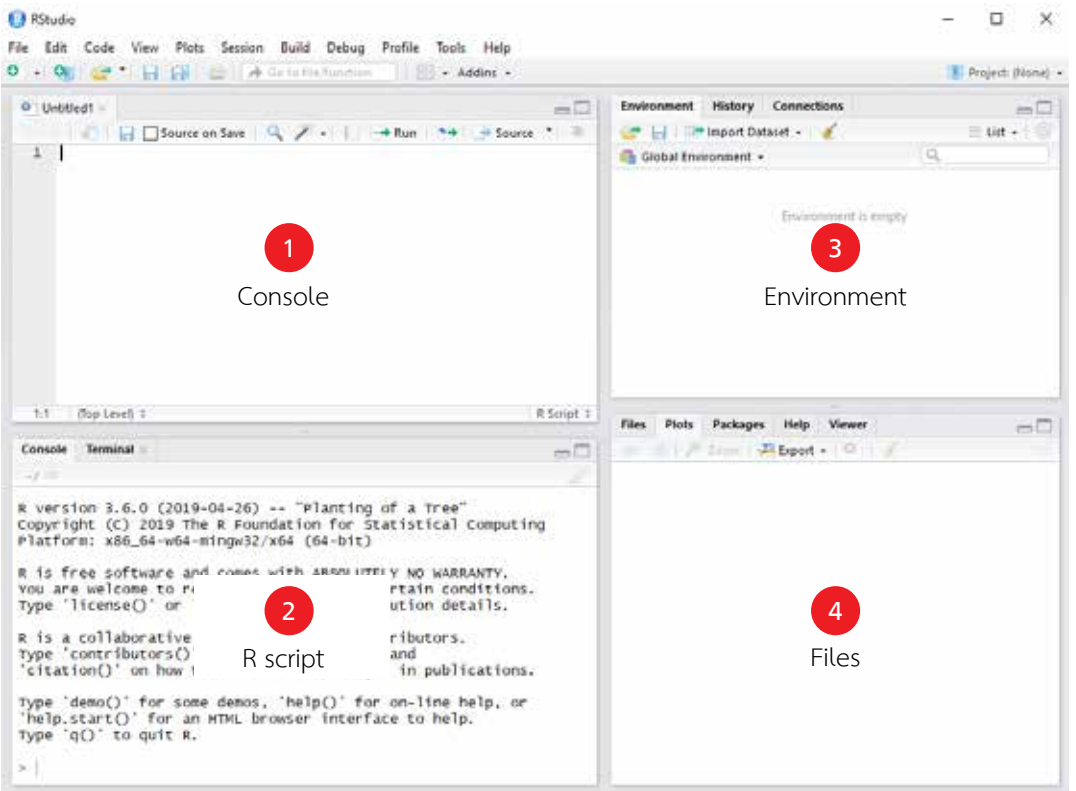
3.3.3 หลังจากดาวน์โหลดโปรแกรม Rstudio ตัวติดตั้งจะอยู่ในโฟลเดอร์ Downloads ไฟล์ที่ดาวน์โหลดมามีชื่อว่า 1.2.5019.exe ซึ่งเมื่อ double click ระบบจะแสดงหน้าจอการติดตั้งโปรแกรม ให้กด next ไปเรื่อยๆ และกดปุ่ม Finish เมื่อการติดตั้งเสร็จสิ้น

3.4 การใช้งานโปรแกรม R- studio

โปรแกรม RStudio แบ่งหน้าต่าง ออกเป็น 4 ส่วน ดังนี้

- หน้าต่าง R Script ถือได้ว่าเป็นหน้าต่างหลักที่เราจะสื่อสารกับโปรแกรม คำสั่งต่างๆ จะถูกสั่งและบันทึกจากหน้าต่างนี้
- หน้าต่าง Console สำหรับการพิมพ์คำสั่งและแสดงผลลัพธ์ เมื่อเขียนคำสั่งในหน้าต่าง R Script แล้ว highlight คำสั่งทั้งหมด จากนั้น กด ctr+enter หรือกด Run ผลการวิเคราะห์จะปรากฏที่หน้าต่าง console นั้น หมายความว่า คำสั่ง ถูกสั่ง แล้ว และโปรแกรมจะทำการประมวลผล
- หน้าต่าง Environment and history แสดงผลของค่าที่กำหนด เช่น จำนวนตัวแปร ชื่อตัวแปร ข้อมูลที่ใช้ เป็นต้น รวมทั้งผู้ใช้สามารถดูประวัติการใช้คำสั่งได้

หน้าต่าง Files เป็นหน้าจอแสดงโฟลเดอร์ ณ ปัจจุบันที่กำลังทำงานอยู่ Plots เป็นพื้นที่แสดงกราฟ, Packages แสดง package ที่มีอยู่ แล้วยังสามารถคลิกเพื่อ install package และ update package ได้ Help แสดงรายการช่วยเหลือที่ต้องการ รายละเอียดตามภาพที่ 3.8



ภาพที่ 3.8 หน้าต่าง Rstudio

3.5 การใช้ R เบื้องต้น

3.5.1 การคำนวณใน R

การคำนวณทางคณิตศาสตร์ใน R สามารถทำได้ง่ายและสะดวกรวดเร็ว เช่น

- สัญลักษณ์พื้นฐานที่ใช้ในการคำนวณ คือ บวก (+) ลบ (-) คูณ (*) หาร (/) และยกกำลัง (^)
- การคำนวณฟังก์ชันทางคณิตศาสตร์ต่างๆ เช่น รากที่สอง sqrt() ค่าลอการิทึม log() เอกซ์โพเนนเชียล exp() เป็นต้น แสดงตัวอย่างการคำนวณด้วยโปรแกรม R

```
8+4
   [1] 12
8-4
   [1] 4
8*4
   [1] 32
8/4
   [1] 2
8^2
   [1] 64
sqrt(9)
   [1] 3
log(1)
   [1] 0
exp(0)
   [1] 1
```

3.5.2 การเก็บค่าการกระทำต่างๆ ใน R

การเก็บค่าการกระทำต่างๆ ใน R มีสัญลักษณ์หลักๆ ดังนี้

<- เป็นการนำค่าหรือผลจากการคำนวณหรือการทำงานของคำสั่งที่ปลายลูกศร (ด้านขวา) ไปใส่ในชื่อวัตถุที่เขียนไว้ที่อยู่ด้านหัวลูกศร (ด้านซ้าย)

-> เป็นการนำค่าหรือผลจากการคำนวณหรือการทำงานของคำสั่งที่ปลายลูกศร (ด้านซ้าย) ไปใส่ในชื่อวัตถุที่เขียนไว้ที่อยู่ด้านหัวลูกศร (ด้านขวา)

= มีความหมายเช่นเดียวกับ <- แต่ไม่นิยมใช้ เพราะจะสับสนกับเครื่องหมาย == ที่เป็นเครื่องหมายเท่ากับทางตรรกศาสตร์

Tip ชื่อวัตถุหรือตัวแปร อักษรตัวเล็กและใหญ่ถือว่าต่างกัน

แสดงตัวอย่างการเก็บค่าการกระทำต่างๆ ใน R โดย a คือ ผลบวกของ 8 และ 4 ส่วน b คือ ผลต่างของ 8 และ 4 และ c คือ ผลหารของ a และ b

```
a <- 8+4
b <- 8-4
c <- a/b
c
   [1] 3
d <- c^2
sqrt(d)
   [1] 3
```

3.5.2 สัญลักษณ์ทางตรรกศาสตร์ใน R

สัญลักษณ์ทางตรรกศาสตร์ใน R แสดงตามตารางที่ 3.1

ตารางที่ 3.1 สัญลักษณ์ทางตรรกศาสตร์ใน R

สัญลักษณ์	ความหมาย
==	การเท่ากัน
!=	การไม่เท่ากัน
<	น้อยกว่า
>	มากกว่า
<=	น้อยกว่าหรือเท่ากับ
>=	มากกว่าหรือเท่ากับ
&	และ
|	หรือ

ตัวอย่างสัญลักษณ์ทางตรรกศาสตร์ใน R ซึ่งเปรียบเทียบค่า a b และ c โดยผลที่ออกมาคือ TRUE และ FALSE

```
a <- 8
b <- 8
c <- 6
a==b
   [1] TRUE
a==c
   [1] FALSE
a != c
   [1] TRUE
```

3.5.4 ชนิดของข้อมูลใน R

ตารางที่ 3.2 ชนิดของข้อมูลใน R

ชนิดของข้อมูล	ตัวอย่าง	ตรวจสอบ
ตรรกะ Logical	TRUE, FALSE	v <- TRUE class(v) ผลลัพธ์ของคำสั่งเป็นดังนี้ [1] “logical”
ตัวเลข (มีจุดทศนิยม) Numeric	12.3, 5, 999	v <- 23.5 class(v) ผลลัพธ์ของคำสั่งเป็นดังนี้ [1] “numeric”
จำนวนเต็ม (ไม่มีจุดทศนิยม) Integer	2L, 34L, 0L	v <- 2L class(v) ผลลัพธ์ของคำสั่งเป็นดังนี้ [1] “integer”
ข้อความ Character	‘A’, “ดี”, “true”, “yellow”	v <- c(“Hello”, “Hi”, “Ni Hao”) class(v) ผลลัพธ์ของคำสั่งเป็นดังนี้ [1] “character”

3.5.5 คำสั่งพื้นฐานใน R

- การกำหนด Working directory
  getwd() ตรวจสอบว่าขณะนี้เราทำงานอยู่ directory ไหน
  setwd() กำหนด directory ที่ใช้ทำงาน เช่น setwd(“d:/DataNCD”)
  dir() เรียกดูไฟล์ทั้งหมดที่จัดเก็บอยู่ใน directory
- ตั้งค่าให้ R อ่านภาษาไทยได้
  Sys.setlocale(locale=“Thai”)
- การศึกษาแต่ละคำสั่งหรือ function จาก help() ซึ่งมีความสำคัญอย่างมาก เช่น ถ้าต้องการอยากรู้ว่า setwd คืออะไร และ ใช้งานอย่างไร ? สามารถพิมพ์ว่า help(setwd) เพื่อดูรายละเอียดได้เลย

3.6 โครงสร้างข้อมูลใน R

โครงสร้างข้อมูลใน R หลักๆ ประกอบด้วย เวกเตอร์ (Vector) เมทริกซ์ (Matrix) อาร์เรย์ (Array) และกรอบข้อมูล (Data frame)

- เวกเตอร์ (Vector) เป็นโครงสร้างข้อมูลพื้นฐานใน R
- เมทริกซ์ (Matrix) เป็นเวกเตอร์หลายๆ เวกเตอร์มาประกอบกัน
- Array คือกลุ่มของข้อมูลที่เรียงลำดับกัน มีจำนวนแน่นอน ซึ่งข้อมูลจะเป็นประเภทเดียวกัน
- กรอบข้อมูล (Data frame) ข้อมูลที่ประกอบด้วยแถวและคอลัมน์

แสดงตัวอย่างโครงสร้างข้อมูลใน R โดยสร้างเวกเตอร์ อายุ (age) และ เพศ (sex) จากนั้นสร้าง เมทริกซ์ person จากเวกเตอร์ดังกล่าว และสร้างกรอบข้อมูล dat จากเมทริกซ์ดังกล่าว

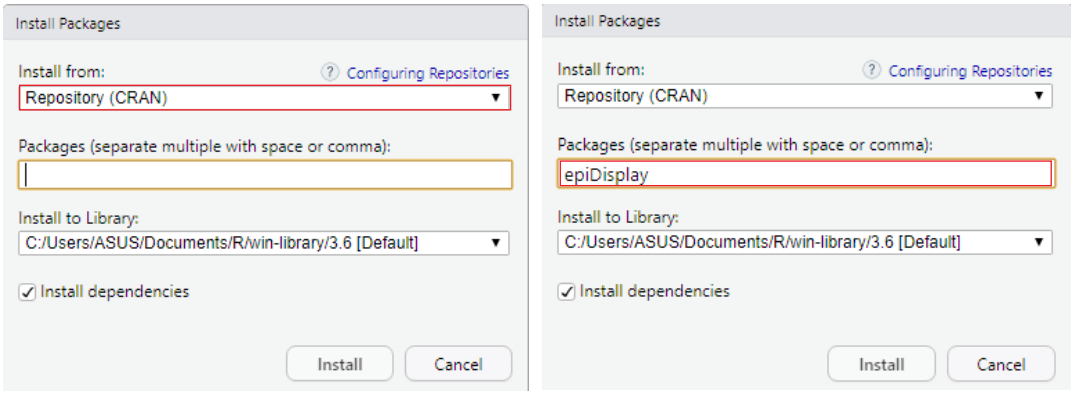
```
#Vector
age <- c(10,20,30,40,50)
age
   [1] 10 20 30 40 50
sex <- c("male", "male", "female", "female", "male")
sex
   [1] "male"   "male"   "female" "female" "male"
#Matrix
person <- cbind(age,sex)
person
      age  sex
   [1,] "10" "male"
   [2,] "20" "male"
   [3,] "30" "female"
   [4,] "40" "female"
   [5,] "50" "male"
#Data frame
dat <- data.frame(person)
dat
   age    sex
1  10   male
2  20   male
3  30 female
4  40 female
5  50   male
```

3.6.1 การใช้ Package ใน R

โปรแกรม R มี Package มากกว่า 10,000 Package ที่สามารถติดตั้งและนำมาใช้งานเพื่อช่วยในการจัดการข้อมูล วิเคราะห์ข้อมูล และแสดงผลในรูปแบบต่างๆ โดยสามารถติดตั้งและทำการ load package และเรียกใช้ library() เมื่อเสร็จแล้ว โดยติดตั้งเพียงครั้งเดียว หลังจากนั้นใช้งานได้ตลอด การติดตั้งสามารถทำได้ 2 วิธี คือ การติดตั้ง package ผ่านอินเตอร์เน็ต และการติดตั้ง package โดยใช้ไฟล์

- **การติดตั้ง package ผ่านอินเตอร์เน็ต**

ไปที่ Tools เลือก Install Packages เลือก Repository (CRAN) และพิมพ์ชื่อ package ในตัวอย่างนี้ติดตั้ง package epiDisplay



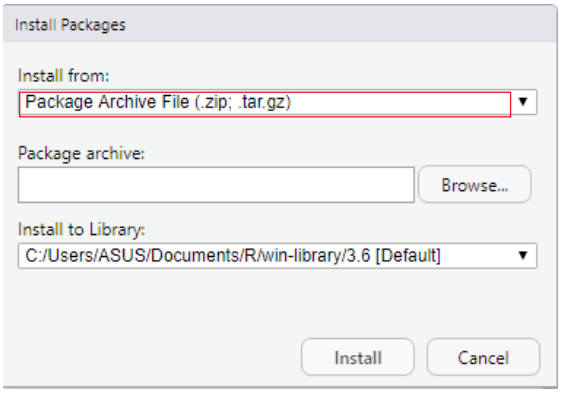
ภาพที่ 3.9 ติดตั้ง package ผ่านอินเตอร์เน็ต

อีกวิธีหนึ่งคือติดตั้งผ่านคำสั่ง R

installed.packages(“epiDisplay”)

- **การติดตั้ง package โดยใช้ไฟล์**

ไปที่ Tools เลือก Install Packages... เลือก Package Archive File (.zip; .tar.gz) และเลือกไฟล์ package ที่จะติดตั้ง



ภาพที่ 3.10 ติดตั้ง package โดยใช้ไฟล์

อีกวิธีหนึ่งคือติดตั้งผ่านคำสั่ง R

installed.packages(“directory file”)

3.6.2 การใช้ package epiDisplay

epiDisplay เป็น package ที่พัฒนาขึ้นโดย ศ.ดร.นพ.วีระศักดิ์ จงสู่วิวัฒน์วงศ์ จากหน่วยระบาดวิทยา คณะแพทยศาสตร์ มหาวิทยาลัยสงขลานครินทร์ โดยเป็น package ที่ใช้สำหรับการคำนวณสถิติด้านระบาดวิทยา หากทำการติดตั้งแล้วสามารถเรียกใช้โดยคำสั่ง library (epiDisplay)

- คำสั่งที่ใช้บ่อยใน epiDisplay

  des() ดูลักษณะของตัวแปรต่างๆ

  summ() คำสั่งสำหรับสรุปค่าทางสถิติเบื้องต้น

  codebook() คำสั่งสำหรับการคำนวณค่าสถิติเชิงพรรณนาที่เหมาะสมกับข้อมูลแต่ละประเภท

3.6.3 คำสั่งอื่นๆ ที่ใช้บ่อยใน R

length() แสดงจำนวนของข้อมูล

str() แสดงโครงสร้างของข้อมูล

class() แสดงชนิดของข้อมูล

names() แสดงชื่อตัวแปรของข้อมูล

cbind() รวมข้อมูลโดยคอลัมน์

rbind() รวมข้อมูลโดยแถว

ls() แสดงวัตถุหรือตัวแปร

rm() ลบวัตถุหรือตัวแปร

factor() ใช้สำหรับเปลี่ยนประเภทตัวแปรเป็นตัวแปรแบบกลุ่ม (categorical data)

cut() เปลี่ยนข้อมูลตัวเลข ให้เป็นกลุ่มช่วง

จากการศึกษาเนื้อหาในบทนี้คาดหวังว่าผู้เรียนสามารถติดตั้งโปรแกรมทั้ง 2 โปรแกรมที่จำเป็นต้องมีไว้ในเครื่องคอมพิวเตอร์เพื่อใช้ในการจัดการและวิเคราะห์ข้อมูล รวมทั้งสามารถเรียนรู้คำสั่งเบื้องต้นทั้งใน R และใน package epiDisplay เพื่อเตรียมความพร้อมเข้าสู่บทอื่นๆ ทั้งนี้ ผู้เรียนสามารถฝึกทำแบบฝึกหัดท้ายบทเพื่อทบทวนการใช้คำสั่งต่างๆ ที่จะใช้ในการจัดการข้อมูลต่อไป

3.7 แบบฝึกหัดท้ายบท

ฝึกปฏิบัติการใช้คำสั่งใน package epiDisplay จากข้อมูล ชื่อ Cars93 ซึ่งอยู่ใน package”MASS” โปรดเรียกข้อมูลดังกล่าวมาสำรวจ โดยเริ่มด้วยคำสั่ง

library(epiDisplay)

data(Cars93, package= “MASS”)

ข้อมูลนี้เป็นข้อมูลเกี่ยวกับรถในสหรัฐอเมริกา โปรดตอบคำถามต่อไปนี้

1. ข้อมูลมีทั้งหมดกี่ตัวแปร
2. ตัวแปรกลุ่มหรือ factor มีกี่ตัว อะไรบ้าง
3. ราคาเฉลี่ย ราคาสูงสุดและต่ำสุด ของรถทั้งหมดเป็นเท่าไร
4. มีรถขับเคลื่อนล้อหน้ากี่คัน คิดเป็นร้อยละเท่าไร (ตัวแปร=DriveTrain)
5. รถที่ผลิตในสหรัฐอเมริกามีกี่คัน (ตัวแปร = Origin)
6. รถทั้งหมดมียี่ห้ออะไรบ้าง (ตัวแปร = Manufacturer)

3.8 บรรณานุกรม

1. กานต์ ยงศิริวิทย์ โปรแกรมภาษา R เบื้องต้น มหาวิทยาลัยรังสิต สืบค้นเมื่อ 20 มีนาคม 2562 ทางhttp://hdc.moph.go.th/download/document/training/ visualization2018/dr_karn/Rprogramming.pdf
2. จิราวรรณ รอนราญ โปรแกรม R เบื้องต้น มหาวิทยาลัยแม่โจ้ สืบค้นเมื่อ 10 สิงหาคม 2562 ทาง https://erp.mju.ac.th/acticleDetail.aspx?qid=622

a. วิโรจน์ อรุณมานะกุล สถิติและการใช้โปรแกรม R จุฬาลงกรณ์มหาวิทยาลัย สืบค้นเมื่อ 15 มีนาคม 2561 ทาง http://pioneer.chula.ac.th/~awirote/courses/res-tech-ling/statistics-and-r.pdf

4. The Comprehensive R Archive Network. Package ‘epiDisplay’ May 10, 2018 สืบค้นเมื่อ 15 พฤษภาคม 2562 ทาง https://cran.r-project.org/
5. อภิรดี แซ่ลิ่ม. (2559). การจัดการข้อมูล กราฟ และการวิเคราะห์ทางสถิติ. สงขลา : ไอคิว มีเดีย.

บทที่ 4
การเลือกใช้สถิติ

บทที่ 4

การเลือกใช้สถิติ

รศ.ดร.นงเยาว์ เกษตร์ภิบาล
E-mail: nongyaok2003@gmail.com

การดำเนินงานทางด้านสาธารณสุขในประเทศไทย มีการเก็บรวบรวมข้อมูลเกี่ยวกับโรคไม่ติดต่อของผู้รับบริการไว้เป็นจำนวนมาก โดยเก็บไว้ในฐานข้อมูล 43 แฟ้ม ซึ่งหากมีการนำข้อมูลดังกล่าวมาทำการวิเคราะห์และสังเคราะห์องค์ความรู้ ก็จะก่อให้เกิดประโยชน์แก่ผู้รับบริการ หน่วยงาน และองค์กรอย่างมหาศาล องค์ความรู้ที่ได้จากการวิเคราะห์ และสังเคราะห์ข้อมูลในหน่วยงานหรือในพื้นที่ของตนเองนั้น จะช่วยให้การแก้ไขปัญหาสาธารณสุขของประชากรในพื้นที่ได้ตรงประเด็น นอกจากนี้องค์ความรู้ดังกล่าวยังสามารถนำมาใช้ในการพัฒนาองค์กร หรือนำไปเผยแพร่ให้กับหน่วยงานอื่นที่เกี่ยวข้องได้

การที่จะได้มาซึ่งองค์ความรู้ที่ถูกต้องนั้น บุคลากรสาธารณสุขจะต้องมีความรู้ในการตรวจสอบคุณภาพของข้อมูล การจัดการข้อมูล การวิเคราะห์ข้อมูล การแปลผลข้อมูล การนำเสนอข้อมูล และการรายงานข้อมูล ซึ่งในบทนี้จะกล่าวถึงเฉพาะการตรวจสอบคุณภาพของข้อมูล การจัดการข้อมูล ความรู้พื้นฐานทางสถิติ การวิเคราะห์ข้อมูลเบื้องต้น และการแปลผลข้อมูลทางระบาดวิทยา

4.1 ความหมายของสถิติ ข้อมูล สารสนเทศ และการประมวลผล

สถิติ (statistics) มีความหมายเป็น 2 นัย คือ สถิติ หมายถึง ตัวเลขที่แสดงข้อเท็จจริงของข้อมูล ซึ่งเป็นตัวเลขที่ได้มาจากการวิเคราะห์ข้อมูลหรือประมวลผล อีกนัยหนึ่ง สถิติ หมายถึง ศาสตร์หรือวิชาที่ว่าด้วยหลักการและระเบียบวิธีทางสถิติ ซึ่งประกอบด้วยกระบวนการ 4 ขั้นตอน คือ การเก็บรวบรวมข้อมูล การวิเคราะห์ข้อมูล การแปลความหมายข้อมูล และการนำเสนอข้อมูล เพื่อประโยชน์ในการช่วยตัดสินใจและกำหนดนโยบายต่างๆ ให้เป็นไปอย่างมีประสิทธิภาพ รวมถึงการประเมินผลโครงการ อันจะนำไปสู่การพัฒนาต่อไป

ข้อมูล (data) หมายถึง ข้อเท็จจริงหรือเหตุการณ์เกี่ยวกับสิ่งต่างๆ เช่น คน สถานที่ สิ่งของต่างๆ ซึ่งมีการเก็บรวบรวมเอาไว้ ยังไม่มีการประมวล แต่สามารถนำไปประมวลผลได้ด้วยการคำนวณด้วยมือหรือโปรแกรมคอมพิวเตอร์ ทั้งนี้สามารถเรียกเอามาใช้ประโยชน์ได้ในภายหลัง โดยข้อมูลอาจเป็นตัวเลข สัญลักษณ์ ตัวอักษร เสียง ภาพ ภาพเคลื่อนไหว เป็นต้น

ข้อมูลด้านสุขภาพ (health data) หมายถึง ข้อเท็จจริงที่เกี่ยวข้องกับสุขภาพของผู้รับบริการ ซึ่งอาจเป็นข้อมูลของบุคคล ครอบครัว หรือชุมชน ที่บุคลากรสาธารณสุขจะนำไปใช้ประโยชน์ในการรักษาพยาบาล

สารสนเทศ (information) หมายถึง ข้อมูลที่มีสาระอยู่ในตัว สามารถสื่อความหมายให้ผู้ที่ต้องการใช้ข้อมูลนั้นเกิดความเข้าใจ โดยเกิดจากการนำข้อมูล ผ่านระบบการประมวลผล คำนวณ วิเคราะห์ และแปลความหมายเป็นข้อความ อย่างเป็นระบบตามหลักวิชาการ ที่สามารถนำไปใช้ประโยชน์ได้

การประมวลผล (data processing) หมายถึง การนำข้อมูลดิบ (raw data) มาดำเนินการบางประการ เช่น จัดหมวดหมู่ จำแนก คัดแยก คำนวณ บันทึก เปรียบเทียบ เพื่อให้ข้อมูลอยู่ในรูปที่กะทัดรัด มีความหมายและสะดวกต่อการนำไปใช้ประโยชน์ตามที่ต้องการ

4.2 คำศัพท์ที่เกี่ยวข้องกับสถิติ

ประชากร (population) หมายถึง กลุ่มของสิ่งต่างๆ ทั้งหมดที่ผู้ศึกษาหรือผู้วิจัยสนใจ อาจจะเป็นสิ่งมีชีวิต หรือไม่มีชีวิตก็ได้ ประชากรในทางสถิติอาจหมายถึง บุคคล กลุ่มบุคคล องค์กรต่างๆ สัตว์ หรือสิ่งของ

กลุ่มตัวอย่าง (sample) หมายถึง ส่วนหนึ่งของประชากรที่ผู้ศึกษาหรือผู้วิจัยสนใจ กลุ่มตัวอย่างถูกเลือกมาจากประชากรด้วยวิธีการใดวิธีการหนึ่งเพื่อเป็นตัวแทนในการศึกษาและเก็บข้อมูล กลุ่มตัวอย่างที่ดีคือกลุ่มตัวอย่างที่มีลักษณะต่างๆ ที่สำคัญครบถ้วนเหมือนกับกลุ่มประชากร เป็นตัวแทนที่ดีของกลุ่มประชากรได้ ในกรณีที่กลุ่มประชากรที่จะศึกษานั้นเป็นกลุ่มขนาดใหญ่ เกินความสามารถหรือความจำเป็นที่ต้องการ จะใช้กลุ่มตัวอย่างในการวิจัยเพื่อประหยัดในด้านงบประมาณและเวลา

ค่าพารามิเตอร์ (parameter) หมายถึง ค่าต่างๆ ที่คำนวณมาได้จากกลุ่มประชากร

ค่าสถิติ (statistics) หมายถึง ค่าต่างๆ ที่คำนวณมาได้จากกลุ่มตัวอย่าง ค่าที่ได้จะเปลี่ยนแปลงได้ตามกลุ่มตัวอย่างที่เลือกสุ่มมา

ตัวแปร (variable) ในทางสถิติ หมายถึง ลักษณะบางอย่างที่ผู้ศึกษาหรือผู้วิจัยสนใจศึกษา เช่น เพศ อายุ ประสบการณ์การทำงาน เป็นต้น ค่าของตัวแปร อาจอยู่ในรูปข้อความ หรือตัวเลขก็ได้

ค่าที่เป็นไปได้ (possible value) หมายถึง ค่าของตัวแปรที่อาจจะเกิดขึ้นได้จริง

ค่าจากการสังเกต (observed value) หมายถึง ค่าที่เก็บรวบรวมมาได้จริง

4.3 ประเภทของข้อมูล

การมีความรู้ความเข้าใจเกี่ยวกับประเภทของข้อมูลจะช่วยให้สามารถเลือกสถิติที่ใช้ในการวิเคราะห์ข้อมูลได้อย่างถูกต้องเหมาะสม การแบ่งประเภทของข้อมูลสามารถแบ่งได้หลายแบบ ในที่นี้จะกล่าวถึงการแบ่งประเภทข้อมูลตามลักษณะของข้อมูล ตามระดับของข้อมูล และตามแหล่งที่มาของข้อมูล ดังนี้

4.3.1 แบ่งตามลักษณะของข้อมูล ได้ดังนี้

1. **ข้อมูลเชิงปริมาณ (quantitative data)** เป็นข้อมูลที่แสดงความแตกต่างในเรื่องปริมาณหรือขนาด สามารถวัดค่าได้ว่ามีค่ามากหรือน้อย เช่น อายุ ส่วนสูง น้ำหนัก แบ่งเป็น 2 ประเภท

- **ข้อมูลแบบไม่ต่อเนื่อง (discrete data)** หมายถึง ข้อมูลที่มีค่าเป็นเลขจำนวนเต็มที่มีความหมาย เช่น จำนวนบุตร จำนวนการตั้งครรภ์ จำนวนผู้ป่วย จำนวนผู้ติดเชื้อ เป็นต้น
- **ข้อมูลแบบต่อเนื่อง (continuous data)** หมายถึง ข้อมูลที่อยู่ในรูปตัวเลขที่มีค่าได้ทุกค่าในช่วงที่กำหนด และมีความหมาย เช่น รายได้ อายุ น้ำหนัก ส่วนสูง ชีพจร การหายใจ ความดันโลหิต ระดับน้ำตาลในเลือด และระดับไขมันในเลือด เป็นต้น

2. **ข้อมูลเชิงคุณภาพ (qualitative data)** เป็นข้อมูลที่แสดงลักษณะที่แตกต่างกัน แต่ไม่ได้อยู่ในรูปของตัวเลข ไม่สามารถบอกได้ว่ามีค่ามากหรือน้อย แต่สามารถบอกได้ว่าดีหรือไม่ดี บอกลักษณะความเป็นกลุ่มของข้อมูล เช่น เพศ แบ่งเป็น เพศชาย และเพศหญิง การเจ็บป่วย แบ่งเป็น ป่วย และไม่ป่วย การติดเชื้อ แบ่งเป็น ติดเชื้อ และไม่ติดเชื้อ เป็นต้น

4.3.2 แบ่งตามระดับของข้อมูล ได้ดังนี้

1. **มาตรานามบัญญัติ (nominal scales)** เป็นการวัดที่มีระดับต่ำสุด วัดโดยใช้ตัวเลขหรือสัญลักษณ์อื่นๆ บอกถึงการจำแนกหรือแบ่งกลุ่ม สิ่งของ บุคคล หรือคุณลักษณะต่างๆ การวัดแบบนี้เป็นการวัดเพื่อให้เห็นความแตกต่างเท่านั้น แต่นำมาเปรียบเทียบความมากน้อยไม่ได้ เช่น เพศ การเจ็บป่วย สถานภาพสมรส อาชีพ เลขที่โรงพยาบาล หมายเลขโทรศัพท์ และบ้านเลขที่ เป็นต้น

2. **มาตราเรียงลำดับ (ordinal scales)** เป็นการวัดที่บอกความแตกต่างระหว่างกลุ่มหรือประเภทได้ เมื่อนำมาเปรียบเทียบกัน สามารถบอกความมากน้อยกว่ากันได้ แต่ไม่สามารถบอกได้ว่ามีปริมาณแตกต่างกันอยู่เท่าใด เช่น ระดับการศึกษา (ประถม มัธยม ประกาศนียบัตร ปริญญาตรี ปริญญาโท ปริญญาเอก) ความรุนแรงของโรค (มาก ปานกลาง ต่ำ) ความเก่ง ความสวย ความชอบ (มาก ปานกลาง น้อย) และความพึงพอใจ (มากที่สุด มาก ปานกลาง น้อย น้อยที่สุด) เป็นต้น

3. **มาตราอันตรภาค (interval scales)** การวัดในระดับนี้จะบอกระยะห่างหรือความแตกต่างระหว่างสิ่งที่วัดได้ ว่ามากหรือน้อยกว่ากันเท่าใด แต่ไม่มีค่าศูนย์แท้ที่เกิดขึ้นจากการวัดได้ เช่น อุณหภูมิ และคะแนน เป็นต้น กล่าวคือ คนที่ได้คะแนนศูนย์คะแนน ไม่ได้หมายความว่า ไม่มีความรู้

4. **มาตราอัตราส่วน (ratio scales)** เป็นการวัดที่มีความละเอียดและมีความสมบูรณ์มากที่สุด สามารถบอกระยะห่างหรือความแตกต่างระหว่างสิ่งที่วัดได้ ว่ามากหรือน้อยกว่ากันเท่าใด และมีค่าศูนย์แท้ (absolute zero) ที่เกิดขึ้นจากการวัดได้ เช่น รายได้ อายุ น้ำหนัก ส่วนสูง ชีพจร การหายใจ ความดันโลหิต ระดับน้ำตาลในเลือด และระดับไขมันในเลือด เป็นต้น

4.3.3 ข้อมูล แบ่งตามแหล่งที่มาของข้อมูล ได้ดังนี้

1. **ข้อมูลปฐมภูมิ (primary data)** หมายถึง ข้อมูลที่ได้จากการเก็บรวบรวมโดยตรงจากผู้รับบริการด้วยวิธีการต่างๆ เช่น การสัมภาษณ์ การสังเกต การตรวจร่างกาย การทดลอง การตอบแบบสอบถาม และการ

ประเมินด้วยเครื่องมือ หรือแบบวัดชนิดต่างๆ ข้อมูลชนิดนี้สามารถควบคุมคุณภาพของข้อมูลได้ง่ายกว่าและมีความเป็นปัจจุบันมากกว่า

2. ข้อมูลทุติยภูมิ (secondary data) หมายถึง ข้อมูลที่ผู้เก็บข้อมูลรวบรวมจากผู้อื่นหรือแหล่งอื่นที่มีผู้รวบรวมไว้แล้ว ไม่ได้เก็บข้อมูลจากผู้รับบริการโดยตรง เช่น รายงานผู้ป่วย บันทึกการรักษาพยาบาล และรายงานผลการตรวจต่างๆ ของผู้ป่วย เช่น ผลตรวจเลือด ปัสสาวะ อุจจาระ เสมหะ ผลการตรวจทางรังสี เช่น การ X-ray การตรวจมวลกระดูก การตรวจด้วยคอมพิวเตอร์ และการตรวจด้วยคลื่นแม่เหล็กไฟฟ้า เป็นต้น

4.4 การตรวจสอบคุณภาพข้อมูลและการจัดการข้อมูล

การตรวจสอบคุณภาพข้อมูล (data validation) และการจัดการข้อมูล (data management) เป็นสิ่งที่มีความสำคัญอย่างยิ่ง ต้องดำเนินการอย่างละเอียดรอบคอบก่อนนำข้อมูลไปวิเคราะห์ เพราะถ้าไม่มีการตรวจสอบคุณภาพของข้อมูลและการจัดการข้อมูลที่ไม่มีประสิทธิภาพ จะเปรียบเสมือน “การนำขยะเข้าไปทำการวิเคราะห์ สิ่งที่ได้ออกมาคือขยะ (garbage in garbage out)” ดังนั้น ผู้วิเคราะห์ข้อมูลต้องพึงระลึกไว้เสมอว่า ก่อนวิเคราะห์ข้อมูลต้องทำการตรวจสอบคุณภาพข้อมูลและจัดการข้อมูลให้ดีก่อน

การตรวจสอบคุณภาพข้อมูล หลังจากรวบรวมข้อมูลแล้ว ต้องมีการตรวจสอบความถูกต้องของข้อมูล ก่อนนำชุดข้อมูลไปวิเคราะห์ และประมวลผล เพื่อนำไปใช้ในการอ้างอิงหรือแปลผลได้อย่างถูกต้อง การตรวจสอบความถูกต้องของข้อมูลทำได้หลายวิธี เช่น ให้ผู้ป่วย ครอบครัว หรือผู้ดูแลผู้ป่วย เป็นผู้ตรวจสอบข้อมูลว่าสิ่งที่บันทึกกับข้อมูลที่ตอบแบบสอบถามหรือสัมภาษณ์ถูกต้องหรือไม่ หรือเปรียบเทียบข้อมูลที่ได้จากการสัมภาษณ์กับการตรวจร่างกาย การสังเกต หรือกับรายงานบันทึก เป็นต้น สำหรับกรณีที่ลงบันทึกข้อมูลไว้ในระบบฐานข้อมูลเรียบร้อยแล้ว สามารถทำการตรวจสอบได้โดยใช้คำสั่งในโปรแกรมคอมพิวเตอร์ ซึ่งจะกล่าวถึงรายละเอียดในบทต่อไป

กรณีที่ตรวจสอบพบว่าข้อมูลไม่ถูกต้องหรือมีความผิดพลาดในการบันทึกข้อมูล ต้องมีการทำความสะอาดข้อมูล (data cleaning) ก่อนการวิเคราะห์ทุกครั้ง

4.5 ประเภทของสถิติ

สถิติ แบ่งออกเป็น 2 ประเภท ได้แก่

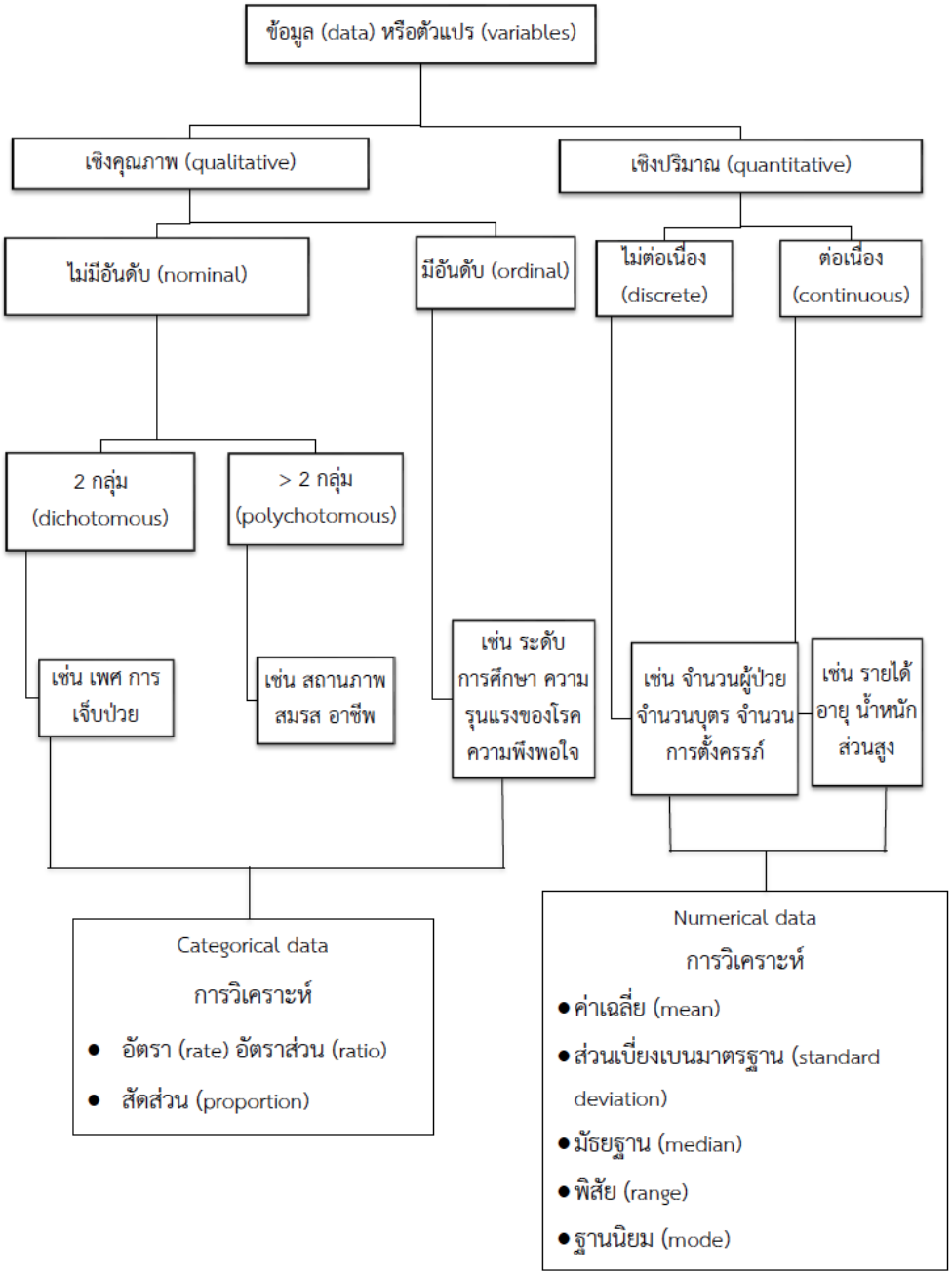
1. สถิติเชิงพรรณนา (descriptive statistics) เป็นวิธีการทางสถิติที่ใช้พรรณนาลักษณะของข้อมูลกลุ่มตัวอย่างที่ทำการศึกษา โดยผลการศึกษาใช้อธิบายเฉพาะกลุ่มที่ศึกษาเท่านั้น ไม่สามารถนำไปอ้างอิงกลุ่มอื่นๆ ที่ไม่ได้ทำการศึกษา

สถิติเชิงพรรณนา เช่น ความถี่ (frequency) ร้อยละ (percentage) อัตรา (rate) อัตราส่วน (ratio) สัดส่วน (proportion) ค่าเฉลี่ย (mean) ส่วนเบี่ยงเบนมาตรฐาน (standard deviation) ค่ามัธยฐาน (median) ค่าพิสัย (range) และค่าฐานนิยม (mode) เป็นต้น

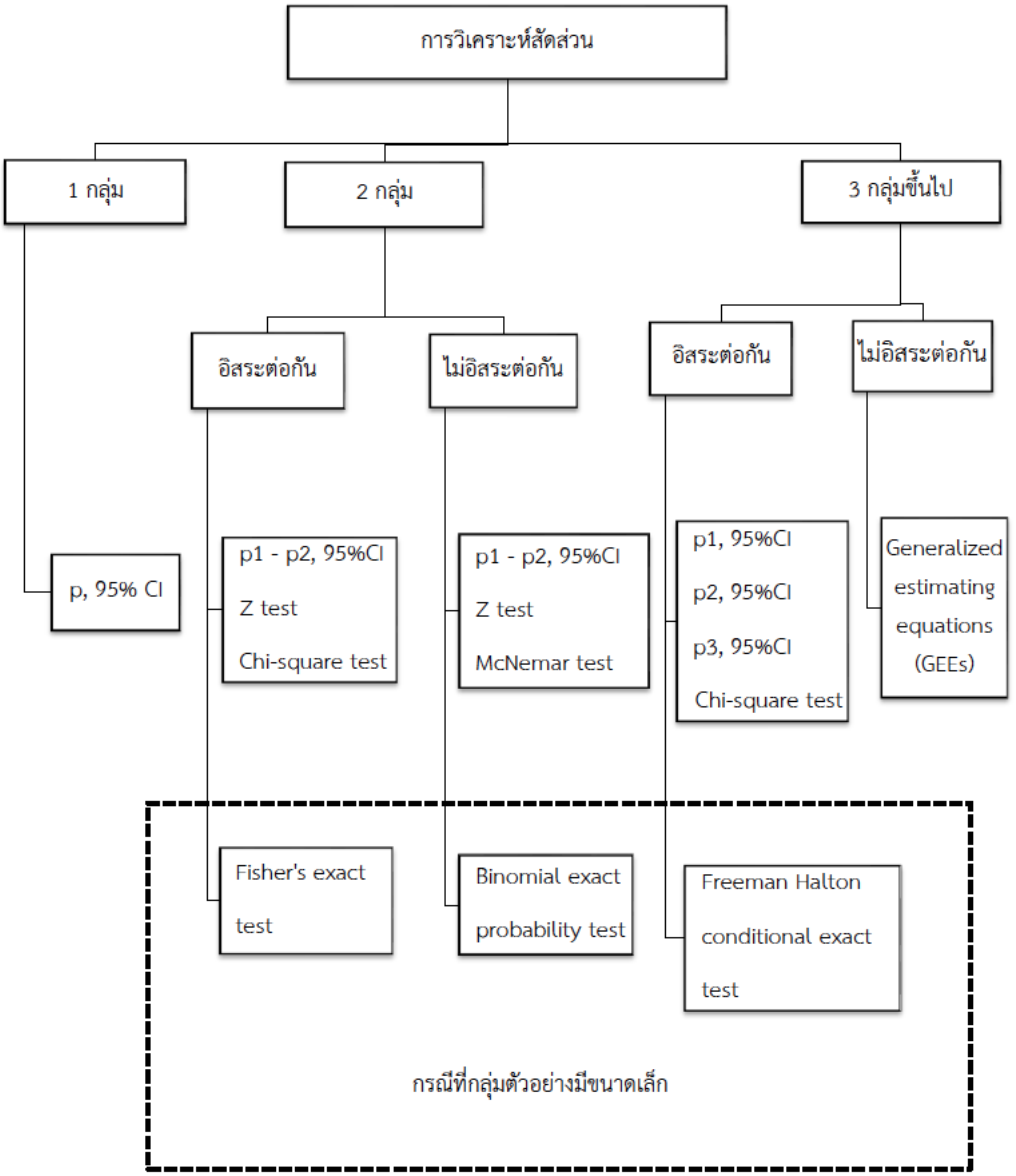
2. สถิติอ้างอิง (inferential statistics) หรือ เรียกอีกอย่างหนึ่งว่า สถิติเชิงอนุมาน เป็นสถิติที่มุ่งศึกษาและอธิบายลักษณะต่างๆ ของกลุ่มประชากรเป้าหมาย โดยการเก็บรวบรวมข้อมูลจากกลุ่มย่อยที่เรียกว่า กลุ่มตัวอย่าง (sample) ซึ่งเป็นส่วนหนึ่งของกลุ่มประชากร แล้วทำการวิเคราะห์ข้อมูลจากกลุ่มตัวอย่าง หลังจากนั้นนำผลการศึกษาไปสรุปอ้างอิงถึงกลุ่มใหญ่ที่เรียกว่า กลุ่มประชากร (population) ซึ่งเป็นกลุ่มเป้าหมายที่ต้องการศึกษา ดังนั้น ค่าที่ได้จะเป็นค่าประมาณการ (estimate) จากกลุ่มตัวอย่าง เพื่ออธิบายลักษณะประชากร จึงอาจมีความผิดพลาดเกิดขึ้นได้ เช่น ความผิดพลาดจากการสุ่มตัวอย่าง (sampling error) ความแปรปรวนของ

ค่าประมาณการ (standard error) ถ้าความแปรปรวนสูง แสดงว่าการประมาณการไม่เที่ยง การลดความแปรปรวนทำได้โดยการเพิ่มขนาดของกลุ่มตัวอย่าง ในทางระบาดวิทยาความผิดพลาดแบ่งเป็น 3 ประเภทคือ ความผิดพลาดในการเลือกกลุ่มตัวอย่าง (selection bias) ความผิดพลาดจากการเก็บรวบรวมข้อมูล (information bias) และความผิดพลาดจากตัวแปรรบกวน (confounding) เพื่อให้การอ้างอิงไปถึงกลุ่มประชากรมีความถูกต้องมากที่สุด ต้องพยายามควบคุมความผิดพลาดต่างๆ ที่อาจเกิดขึ้นให้ได้มากที่สุด ซึ่งสถิติอ้างอิงที่ใช้ในการวิเคราะห์ข้อมูลแสดงดังภาพที่ 4.1 และ 4.2 และตารางที่ 4.1

สถิติอ้างอิงที่ใช้ทดสอบสมมติฐาน เมื่อใดที่ผู้วิจัยต้องการทดสอบสมติฐานเพื่อนำค่าสถิติที่ได้จากกลุ่มตัวอย่างไปอ้างอิงถึงค่าของประชากร (parameter) จะใช้สถิติที่เรียกว่า สถิติพาราเมทริกซ์ (parametric statistic) อย่างไรก็ตาม หากจะใช้สถิติแบบพาราเมทริกซ์ได้ ข้อมูลจะต้องเป็นไปตามข้อตกลงเสียก่อน ซึ่งเมื่อใดที่พบว่าข้อมูลไม่เป็นไปตามข้อตกลง ก็จะไม่สามารถใช้สถิติแบบพาราเมทริกซ์ได้ ดังนั้น การทดสอบสมมติฐานแบบนอนพาราเมทริกซ์จึงเป็นวิธีที่แก้ปัญหาดังกล่าว โดยสถิตินอนพาราเมทริกซ์สามารถใช้ได้ทั้งข้อมูลที่มีการแจกแจงแบบปกติหรือไม่ปกติก็ได้ และใช้ได้กับข้อมูลที่อยู่ในมาตรวัดตั้งแต่นามบัญญัติ (Norminal) ขึ้นไป แต่อำนาจในการวิเคราะห์และแปลผลจะลดลง ซึ่งโดยทั่วไปแล้วถ้าข้อมูลเป็นไปตามข้อตกลงของสถิติพาราเมทริกซ์แล้ว ควรใช้สถิติพาราเมทริกซ์มากกว่าสถิตินอนพาราเมทริกซ์



ภาพที่ 4.1 ชนิดของข้อมูลและสถิติเบื้องต้นที่ใช้ในการวิเคราะห์ข้อมูล



ภาพที่ 4.2 สถิติอ้างอิงที่ใช้ในการวิเคราะห์สัดส่วน

ตารางที่ 4.1 เปรียบเทียบการทดสอบแบบพาราเมทริกซ์และแบบนอนพาราเมทริกซ์

จุดประสงค์ของการทดสอบ	การแจกแจงของข้อมูล	
	ข้อมูลเป็นเส้นโค้งปกติ (normal distribution)	ข้อมูลไม่เป็นเส้นโค้งปกติ (non-normal distribution)
	การทดสอบโดยใช้สถิติพาราเมทริกซ์ (parametric statistics)	การทดสอบโดยใช้สถิตินอนพาราเมทริกซ์ (nonparametric statistics)
1. ทดสอบค่าเฉลี่ยของประชากร 1 กลุ่ม(one sample)	One sample Z-test or one-sample t-test	Runs test
2. ทดสอบค่าเฉลี่ยของประชากร 2 กลุ่มเมื่อกลุ่มตัวอย่างเป็นอิสระต่อกัน (independent samples)	Independent sample t-test	Mann–Whitney U test (Wilcoxon rank-sum test)
3. ทดสอบค่าเฉลี่ยของประชากร 2 กลุ่มเมื่อกลุ่มตัวอย่างไม่เป็นอิสระต่อกัน (non-independent samples)	Paired t-test	Wilcoxon Signed rank test
4. ทดสอบค่าเฉลี่ยของประชากรมากกว่า 2 กลุ่ม (k กลุ่ม) เมื่อกลุ่มตัวอย่างเป็นอิสระต่อกัน (independent samples)	One-way ANOVA	Kruskal-Wallis test
5. ทดสอบค่าเฉลี่ยของประชากรมากกว่า 2 กลุ่ม (k กลุ่ม) เมื่อกลุ่มตัวอย่างไม่เป็นอิสระต่อกัน (non-independent samples)	Repeated measures ANOVA	Friedman test
6. การทดสอบความสัมพันธ์ (correlation coefficient)	Pearson product moment Partial correlation Eta (norminal-interval)	Spearman rank Chi-square test for independent Cram'er & Phi Contingency Lambda * Gamma * Somer' d * Kendall rank, Kendall partial rank * Kendall coefficient of concordance *

*สถิติที่ใช้น้อยทางด้านการแพทย์

4.6 การวิเคราะห์ข้อมูลทางระบาดวิทยาพื้นฐาน

1. ความถี่ (frequency) เป็นวิธีการที่ง่ายที่สุดในการวัดปริมาณ หมายถึง จำนวนของเหตุการณ์ ที่เกิดขึ้นจริง ในกลุ่มประชากรที่ศึกษา หรือมีลักษณะบางสิ่งบางอย่างร่วมกัน ณ พื้นที่ที่กำหนดและในระยะเวลาที่ศึกษา

2. อัตรา (rate) เป็นการวัดโอกาสที่เป็นไปได้ของการเกิด เป็นการเปรียบเทียบจำนวนความถี่ของการเกิดโรคหรือเหตุการณ์ที่สนใจหรือลักษณะบางอย่างในกลุ่มประชากรที่ศึกษาในช่วงเวลาที่กำหนด เป็นการวัดการเกิดโรคขั้นพื้นฐาน ประกอบด้วย

$$\text{อัตรา} = \frac{\text{จำนวนผู้ป่วยหรือจำนวนเหตุการณ์ที่เกิดขึ้นในช่วงเวลาที่กำหนด} \times k}{\text{จำนวนผู้ที่เสี่ยงต่อเหตุการณ์ดังกล่าวในช่วงเวลาที่กำหนด}}$$

เมื่อ k คือ ค่าคงที่ ซึ่งมีค่าเท่ากับ 100 หรือ 1,000 หรือ 100,000 แล้วแต่ความเหมาะสม (10^n)

$$\text{ตัวอย่าง อัตราผู้ป่วยเบาหวานรายใหม่} = \frac{\text{จำนวนผู้ป่วยเบาหวานรายใหม่} \times 100{,}000}{\text{จำนวนประชากรกลุ่มเสี่ยง}}$$

3. อัตราส่วน (ratio) เป็นค่าเปรียบเทียบระหว่างตัวเลข 2 จำนวน หรือ เหตุการณ์ 2 เหตุการณ์ โดยที่เลขเศษไม่ได้เป็นส่วนหนึ่งของเลขตัวส่วน

ตัวอย่าง โรงพยาบาลแห่งหนึ่งมีผู้ป่วยทั้งสิ้น 7,530 คน เป็นเพศชาย 4,110 คน เพศหญิง 3,420 คน

$$\text{อัตราส่วนของผู้ป่วยเพศชายต่อเพศหญิง} = \frac{4{,}110}{3{,}420} = 1.20$$

นั่นคือ อัตราส่วนของมีผู้ป่วยเพศชายต่อผู้ป่วยเพศหญิง 1 คน คือ 1.20 คน

4. สัดส่วน (proportion) เป็นความสัมพันธ์ของจำนวนย่อยกับจำนวนรวมทั้งหมด หรือเป็นการวัดการกระจายของเหตุการณ์ย่อยจากเหตุการณ์ทั้งหมด โดยให้ถือว่าจำนวนรวมทั้งหมดเป็น 1 ส่วน (ถ้าคูณด้วย 100 หน่วยจะเป็นร้อยละหรือเปอร์เซ็นต์)

ตัวอย่าง ในการสำรวจประชากรในหมู่บ้านแห่งหนึ่ง มีประชากรทั้งสิ้น จำนวน 800 คน เป็นหญิง 300 คน และเป็นชาย 500 คน

$$\text{สัดส่วนของผู้หญิงในหมู่บ้านนี้คือ} = \frac{300}{800} = 0.37$$

$$\text{สัดส่วนของผู้ชายในหมู่บ้านนี้คือ} = \frac{500}{800} = 0.63$$

5. อัตราร้อยละหรือเปอร์เซ็นต์ (percentage) เป็นการวัดร้อยละของการกระจายของเหตุการณ์ย่อยจากเหตุการณ์ทั้งหมด ให้ถือจำนวนรวมทั้งหมดเป็น 100 ส่วน

ตัวอย่าง โรงพยาบาลแห่งหนึ่ง มีผู้ป่วย จำนวน 750 ราย จำแนกตามโรคที่เป็น ดังนี้ เบาหวาน 180 ราย ความดันโลหิตสูง 154 ราย ระบบประสาท 145 ราย ระบบทางเดินอาหาร 112 ราย ที่เหลือเป็นโรคอื่นๆ 159 ราย คำนวณร้อยละหรือเปอร์เซ็นต์ ได้ดังนี้

โรค	คำนวณร้อยละ
โรคเบาหวาน	$\frac{180}{750} \times 100 = 24.00$
โรคความดันโลหิตสูง	$\frac{154}{750} \times 100 = 20.53$
ระบบประสาท	$\frac{145}{750} \times 100 = 19.33$
ระบบทางเดินอาหาร	$\frac{112}{750} \times 100 = 14.94$
โรคอื่นๆ	$\frac{159}{750} \times 100 = 21.20$
รวม	100.0

6. อัตราความชุก (prevalence rate) จำนวนการเกิดโรคหรือจำนวนคนที่เป็นโรคในช่วงระยะเวลาหนึ่งหารด้วยประชากรเสี่ยงในช่วงระยะเวลานั้นๆ

$$\text{อัตราความชุก (prevalence rate)} = \frac{\text{จำนวนผู้ป่วยเก่า + ใหม่ที่เกิดขึ้นในช่วงระยะเวลาที่ศึกษา} \times k}{\text{จำนวนประชากรที่เสี่ยงต่อการเกิดโรค}}$$

ตัวอย่าง ในการสำรวจที่เมือง Framingham รัฐ Massachusetts ผู้วิจัยได้ตรวจสายตาที่มีอายุระหว่าง 52 ถึง 85 ปี พบว่ามีคนเป็น cataracts 310 คน จากที่ตรวจ 2,477 คน ดังนั้น ความชุกของของคนเป็นโรคสามารถคำนวณได้ ดังนี้

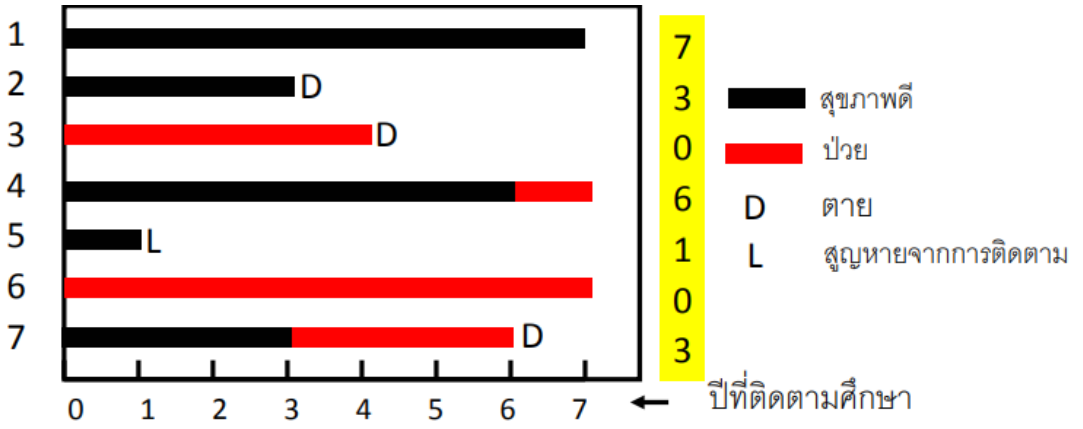
$$\text{อัตราความชุก} = \frac{310 \times 100}{2477}$$

$$= 12.5$$

7. อัตราอุบัติการณ์ (incidence density) เป็นตัวช่วยบอกความรวดเร็วในการเกิดโรคของผู้ป่วยรายใหม่ในประชากรที่ศึกษา จำนวนผู้ป่วยใหม่ที่เกิดขึ้น เทียบกับผลรวมของระยะเวลาที่ไม่ป่วย คำนวณเป็น person-time

$$\text{อัตราอุบัติการณ์} = \frac{\text{จำนวนผู้ป่วยใหม่ที่เกิดขึ้นในช่วงระยะเวลาที่ศึกษาหรือจำนวนครั้งที่ป่วยใหม่} \times k}{\text{ผลรวมของระยะเวลาของแต่ละบุคคลที่เสี่ยงต่อการเกิดโรค}}$$

ตัวอย่าง ช่วงเวลา 7 ปี ที่ติดตามผู้ป่วยของโรงพยาบาลแห่งหนึ่ง



แปลผล : เวลาที่ติดตาม = 7 + 3 + 0 + 6 + 1 + 0 + 3 = 20 คน-ปี
เมื่อติดตามกลุ่มเสี่ยงไป 20 คน/ปี จะพบผู้ป่วย 2 ราย คือ คนที่ 4 และ คนที่ 7
คำถามคือ จะพบผู้ป่วย 1 ราย จะใช้เวลานานเท่าไหร่ นั่นคือ
= 1/Incidence rate คือ 1 x 20 /2 = 10 ปี

8. อัตราป่วยเฉียบพลัน (attack rate) เป็นอัตราอุบัติการณ์ (Incidence rate) ซึ่งมักใช้กับโรคติดเชื้อเฉียบพลัน หรือกรณีมีการระบาดของโรค โดยปกตินิยมใช้หน่วยเป็นร้อยละ

$$\text{อัตราป่วยเฉียบพลัน} = \frac{\text{จำนวนผู้ป่วยใหม่ที่เกิดขึ้นในช่วงระยะเวลาที่ศึกษา} \times 100}{\text{จำนวนประชากรที่เสี่ยง ณ จุดเริ่มต้นของการศึกษา}}$$

9. อัตราตาย (mortality rate) เป็นจำนวนคนตายในช่วงระยะเวลาที่ศึกษา เทียบกับจำนวนประชากรเมื่อเริ่มต้นศึกษา

$$\text{อัตราตาย} = \frac{\text{จำนวนคนตายในช่วงระยะเวลาที่ศึกษา} \times k}{\text{จำนวนประชากรเมื่อเริ่มต้นศึกษา}}$$

10. อัตราตายอย่างหยาบ (crude death rate) เป็นจำนวนคนตายทั้งหมดด้วยทุกสาเหตุในช่วงระยะเวลาที่ศึกษา เทียบกับจำนวนประชากรทั้งหมดในช่วงเวลาเดียวกันหรือประชากรกลางปี

$$\text{อัตราตายอย่างหยาบ} = \frac{\text{จำนวนคนตายทั้งหมดด้วยทุกสาเหตุในช่วงระยะเวลาที่ศึกษา} \times k}{\text{จำนวนประชากรทั้งหมดในช่วงเวลาเดียวกันหรือประชากรกลางปี}}$$

11. อัตราตายเฉพาะ (specific death rate) เป็นจำนวนคนตายเฉพาะอย่าง เช่น เพศ อายุ และสาเหตุในช่วงระยะเวลาที่ศึกษา เทียบกับจำนวนประชากรทั้งหมดในช่วงเวลาเดียวกันหรือประชากรกลางปี

$$\text{อัตราตายเฉพาะ} = \frac{\text{จำนวนคนตายเฉพาะอย่าง (เพศ อายุ และสาเหตุ) ในช่วงระยะเวลาที่ศึกษา} \times k}{\text{จำนวนประชากรทั้งหมดในช่วงเวลาเดียวกันหรือประชากรกลางปี}}$$

12. อัตราป่วยตาย (case-fatality rate) เป็นจำนวนผู้ป่วยตายจากโรคใดโรคหนึ่งในช่วงระยะเวลาหนึ่งเทียบกับจำนวนผู้ป่วยที่ถูกวินิจฉัยว่าเป็นโรคนั้นในระยะเวลาเดียวกัน

$$\text{อัตราป่วยตาย} = \frac{\text{จำนวนผู้ป่วยตายจากโรคใดโรคหนึ่งในช่วงระยะเวลาหนึ่ง} \times k}{\text{จำนวนผู้ป่วยที่ถูกวินิจฉัยว่าเป็นโรคนั้นในระยะเวลาเดียวกัน}}$$

การวิเคราะห์ข้อมูลทางระบาดวิทยาพื้นฐาน มีประโยชน์หลายประการ ตัวอย่างเช่น

1. อัตราความชุก ใช้บอกขนาดของปัญหาทางสาธารณสุขที่มีอยู่ในขณะนั้นหรือช่วงเวลานั้น สามารถใช้เป็นแนวทางในการวางแผนและการให้บริการทางสาธารณสุขแก่ประชาชน มีประโยชน์อย่างมากสำหรับโรคเรื้อรัง

2. อัตราอุบัติการณ์และอัตราป่วยเฉียบพลัน ทำให้ทราบโอกาสหรือความเสี่ยงของการเกิดโรค ทราบความรุนแรงของการระบาด ใช้เป็นข้อมูลสำหรับการศึกษาหาสาเหตุและปัจจัยเสี่ยงของโรค เป็นเครื่องบ่งชี้ถึงความเร่งด่วนในการแก้ไขปัญหาและการควบคุมโรค นอกจากนี้ยังสามารถใช้ในการประเมินผลการดำเนินงานการป้องกันและควบคุมโรคได้อีกด้วย

3. อัตราตาย ใช้บอกสภาวะอนามัยของประชาชนในพื้นที่ การได้รับบริการทางการแพทย์ ประสิทธิภาพการรักษาพยาบาล และประสิทธิภาพการป้องกันและควบคุมโรค

สรุป การที่บุคลากรสาธารณสุขมีความรู้ความเข้าใจที่ถูกต้องเกี่ยวกับประเภทของข้อมูล สถิติที่ใช้ในการวิเคราะห์ข้อมูล การวิเคราะห์ข้อมูลทางระบาดวิทยาเบื้องต้น และการแปลผลข้อมูลทางระบาดวิทยา จะช่วยให้บุคลากรสาธารณสุขนำเสนอข้อมูลที่มีความถูกต้องและน่าเชื่อถือไปยังผู้เกี่ยวข้อง ซึ่งข้อมูลดังกล่าวสามารถนำไปใช้ในการป้องกันและควบคุมโรคไม่ติดต่อ รวมถึงนำไปใช้ในการพัฒนางานด้านโรคไม่ติดต่อในอนาคต

4.7 บรรณานุกรม

1. นงเยาว์ เกษตร์ภิบาล. (2553). การนำเสนอและการแปลผลข้อมูลทางระบาดวิทยาการติดเชื้อ. ในกลยุทธ์สู่การพัฒนาการป้องกันและควบคุมการติดเชื้อ. เชียงใหม่: มิ่งเมือง.
2. บัณฑิต ถิ่นคำรพ. (2543). คู่มือปฏิบัติการชีวสถิติ: สำหรับศึกษาชีวสถิติด้วยตนเอง. ขอนแก่น. โรงพิมพ์คลังนานาวิทยา.
3. บุญธรรม ก่อปรีดาบริสุทธิ์. (2549). สถิติวิเคราะห์เพื่อการวิจัย (พิมพ์ครั้งที่ 4). กรุงเทพฯ: จามจุรีโปรดักท์.
4. บุญพิชชา จิตต์ภักดี, พัชราภรณ์ อารีย์, และประทุม สร้อยวงค์. (2561). เอกสารประกอบการสอนกระบวนวิชา 562708: สถิติสำหรับงานวิจัยทางการพยาบาล. เชียงใหม่: คณะพยาบาลศาสตร์ มหาวิทยาลัยเชียงใหม่.
5. วิสาข์ เกษประทุม. (2544). ความน่าจะเป็น และสถิติเบื้องต้น. กรุงเทพฯ: สำนักพิมพ์พัฒนาศึกษา.
6. ศิริชัย พงษ์วิชัย. (2547). การวิเคราะห์ข้อมูลทางสถิติด้วยคอมพิวเตอร์ (พิมพ์ครั้งที่ 14). กรุงเทพฯ : สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.
7. สีลม แจ่มอุลิตรัตน์. (2540). ระบาดวิทยาพื้นฐาน. สงขลา: โซนพรินท์.
8. Polit, D. F. (2010). Data analysis & statistics for nursing research (2nd Ed.). Connecticut: Appleton & Lange.

บทที่ 5

การนำเข้าและการจัดการข้อมูลโดยใช้โปรแกรม R

บทที่ 5

การนำเข้าและการจัดการข้อมูลโดยใช้โปรแกรม R

ผศ.สพญ.ดร.กรรณิการ์ ณ ลำปาง
E-mail: kna_lampang@hotmail.com

5.1 คำสั่งพื้นฐานก่อนการนำเข้าข้อมูลเข้าสู่โปรแกรม R

คำสั่ง rm()

เป็นคำสั่งสำหรับการลบชุดข้อมูลที่ไม่ต้องการใช้ออกจาก R console เช่น ก่อนการเริ่มต้นวิเคราะห์ข้อมูลโดยใช้ชุดข้อมูลชุดใหม่ การใช้คำสั่ง rm(list=ls()) เป็นการลบชุดข้อมูลที่อาจจะมีค้างอยู่ใน R console จากการทำงานก่อนหน้าออกไปทั้งหมด เพื่อไม่ให้เกิดความสับสนของการเรียกใช้ชุดข้อมูล ยกตัวอย่าง เช่น การสร้างชุดข้อมูลขึ้นมาใหม่ คือ ps1 และ ps2 จากชุดข้อมูลเริ่มต้น คือ ps เมื่อสร้างชุดข้อมูลเพิ่มขึ้นแล้ว และต้องการลบชุดข้อมูลเดิมคือ ps ด้วยการใช้คำสั่ง rm(ps) ชุดข้อมูลที่เหลือคือ ps1 และ ps2 และใช้สำหรับการวิเคราะห์ข้อมูลต่อไป

คำสั่ง getwd()

ใช้สำหรับตรวจสอบพื้นที่ทำงาน (working directory) ที่กำลังทำงานในขณะนั้น

คำสั่ง setwd()

ใช้สำหรับระบุ working directory ที่ต้องการใช้ในการทำงานในแต่ละครั้ง เช่น setwd (“E:/ncd”) เป็นการระบุการทำงาน ที่ drive: E ใน folder ที่ชื่อว่า ncd

คำสั่ง dir()

ใช้สำหรับการตรวจสอบข้อมูลใน working directory ที่ระบุไว้แล้ว

คำสั่ง library()

ใช้สำหรับการเรียกใช้แพ็คเกจ (package) ในการวิเคราะห์ข้อมูลที่เฉพาะ นอกเหนือไปจากคำสั่งที่มีอยู่แล้วในแพ็คเกจพื้นฐาน ทั้งนี้ต้องมีการดำเนินการติดตั้งแพ็คเกจ (install) ที่จะใช้ก่อนเสมอ (อ่านวิธีการติดตั้งแพ็คเกจได้ในบทก่อนหน้า)

คำสั่ง Sys.setlocale()

เป็นการกำหนดให้โปรแกรม R สามารถอ่านค่าข้อมูลที่เป็นภาษาอื่นๆ ได้ ตัวอักษรภาษาไทยได้ หากชุดข้อมูลที่จะใช้มีตัวอักษรภาษาไทย

ตัวอย่างการแสดงผลจากคำสั่งเบื้องต้นก่อนเริ่มนำเข้าข้อมูล

```
rm(list=ls())
getwd()
   [1] "E:/ncdlowernorth"
setwd("E:/ncdlowernorth")
dir()
   [1] "2.Advance"                    "chronic.txt"
   [3] "chronicfu.txt"                "diagnosis_ipd.txt"
   [5] "diagnosis_opd.txt"            "explore data -lwnorth.R"
   [7] "explore data sample.R"        "explore_data_sample.R"
   [9] "explore_data_sample.spin.R"   "explore_data_sample.spin.Rmd"
   [11] "handout"                      "labfu.txt"

library(epiDisplay)
   Loading required package: foreign
   Loading required package: survival
   Loading required package: MASS
   Loading required package: nnet
library(data.table)
Sys.setlocale(locale="Thai")
   [1] "LC_COLLATE=Thai_Thailand.874;LC_CTYPE=Thai_Thailand.874;LC_MONE
TARY=Thai_Thailand.874;LC_NUMERIC=C;LC_TIME=Thai_Thailand.874"
```

rm(list=ls()) คำสั่งนี้จะไม่มีการแสดงผลใดๆ

getwd () คือ โปรแกรมได้ระบุ working directory ที่ทำงานอยู่ในขณะนี้คือ drive: E ใน folder มีชื่อว่า ncdlowernorth

setwd ("E:/ncdlowernorth") เป็นการระบุให้ทำงานใน drive: E ใน folder ที่ชื่อว่า ncdlowernorth dir() คือการแสดงไฟล์ทั้งหมดที่มีอยู่ใน working directory ที่ได้ระบุไว้ ซึ่งมีจำนวนไฟล์ทั้งหมด 12 ไฟล์

library(epiDisplay) คือเรียกใช้ package epiDisplay การที่โปรแกรม R เรียกใช้ package นี้ จะมี package ที่ทำงานร่วมกันโหลดมาด้วยคือ foreign, survival, MASS และ nnet

libraty(data.table) จะไม่เกิดผลลัพธ์ที่แสดงให้เห็น เนื่องจากแพ็คเกจ data.table ไม่ได้เรียกแพ็คเกจอื่นมาทำงานร่วมด้วย

Sys.setlocale (locale= “Thai”) เป็นการกำหนดให้โปรแกรม R สามารถอ่านค่าข้อมูลที่เป็นภาษาไทยได้ หากชุดข้อมูลที่จะใช้มีตัวอักษรภาษาไทย

หมายเหตุ

epiDisplay เป็นแพ็คเกจที่ผู้เขียนสร้างขึ้นมาสำหรับการทำงานในทางระบาดวิทยา ส่วน data.table เป็นแพ็คเกจที่ช่วยทำให้การนำเข้าข้อมูลที่มีขนาดใหญ่ทำได้รวดเร็วมากขึ้น สามารถอ่านเพิ่มเติมได้โดยเลือก epiDisplay หรือ data.table ในหน้าต่าง package

5.2 การนำข้อมูลเข้าสู่โปรแกรม R

การนำข้อมูลเข้าสู่โปรแกรม R สำหรับการวิเคราะห์ข้อมูล ต้องทราบก่อนว่าชุดข้อมูลที่จะนำเข้าเป็นไฟล์ลักษณะใด เช่น เป็นไฟล์ที่บันทึกแบบข้อความ (text) ซึ่งจะมีนามสกุลเป็น .txt หรือ ไฟล์ที่บันทึกในรูปแบบข้อความและคั่นแต่ละคอลัมน์ด้วยเครื่องหมายจุลภาค (,) โดยมีนามสกุลเป็น .csv นอกจากนี้ ควรต้องทราบขนาดของไฟล์ข้อมูลด้วย คำสั่งโดยทั่วไปในการอ่านไฟล์ คือ read.table() หรือ read.csv() (สำหรับการอ่านข้อมูลชนิดอื่นหรือวิธีอื่น ไม่ได้ระบุในบทความนี้)

คำสั่ง read.table() และ read.csv()

read.table() ใช้สำหรับอ่านไฟล์ข้อมูลที่จัดเก็บในรูปแบบ .txt

read.csv() ใช้สำหรับอ่านไฟล์ข้อมูลที่จัดเก็บในรูปแบบ .csv ตัวอย่างรูปแบบคำสั่ง

```
ps <- read.table(“psncd.txt”,head=T, sep=”\t”)
ps <- read.csv(“psncd.csv”,head=T)
```

head = T เป็นคำสั่งเพิ่มเติมที่ระบุว่าชุดข้อมูลนั้นมีแถวแรกเป็นชื่อตัวแปร ถ้าไม่ระบุไว้โปรแกรมจะอ่านแถวแรกเป็นข้อมูลแถวที่ 1

จากตัวอย่างคำสั่ง เป็นการนำไฟล์ที่ชื่อว่า psncd ซึ่งมีนามสกุลเป็น .txt หรือ .csv แล้วแต่กรณี เข้าสู่โปรแกรม R เมื่อนำไฟล์เข้ามาแล้วให้เก็บข้อมูลไว้ในชื่อ ps (หรืออาจจะเป็นชื่ออื่นใดก็ได้ แต่ในที่นี้ระบุเป็น ps) นอกจากนี้ คำสั่ง sep=”\t” เป็นการระบุว่าชุดข้อมูล psncd.txt มีการคั่นแต่ละคอลัมน์ด้วย tab แต่หากชุดข้อมูลใดคั่นด้วยเครื่องหมายจุลภาค (,) ก็ต้องระบุเป็น sep=” , “ โดยที่ไฟล์ psncd มีแถวแรกเป็นชื่อตัวแปร

คำสั่ง fread()

เป็นคำสั่งที่อยู่ในแพ็คเกจ data.table ใช้สำหรับการอ่านไฟล์เหมือนคำสั่ง read() แต่สามารถอ่านไฟล์ได้รวดเร็วกว่า จึงเหมาะสมกับการใช้เมื่อชุดข้อมูลที่จะนำมาวิเคราะห์ข้อมูลมีขนาดใหญ่ (ไฟล์มีขนาด 1 MB)

```
ps<-fread(“psncd.txt”,verbose=T,sep=”\t “, nrows=20000)
```

โดยที่

verbose=T เป็นการบอกให้ R แสดงผลรายละเอียดที่อ่านจากแฟ้มข้อมูล

nrows=20000 เป็นการระบุจำนวนแถวที่ต้องการ ในครั้งนี้ใช้ 20000 แถว

ในกรณีที่มีค่าของข้อมูลบางตัวขาดหายไป ต้องเพิ่ม fill=T ในคำสั่งด้วย จึงจะสามารถนำเข้าข้อมูลได้

5.3 การตรวจสอบข้อมูล

คำสั่ง head() และ tail()

head() เป็นคำสั่งสำหรับแสดงข้อมูล 6 แถวแรกของชุดข้อมูลที่ระบุชื่อในวงเล็บ

tail() เป็นคำสั่งที่สำหรับแสดงข้อมูล 6 แถวสุดท้ายของชุดข้อมูลระบุชื่อในวงเล็บ

คำสั่ง names()

names() เป็นคำสั่งที่สำหรับแสดงชื่อตัวแปรของชุดข้อมูลที่ระบุชื่อในวงเล็บ และใช้เพื่อการเปลี่ยนชื่อตัวแปร

คำสั่ง str() และ des()

str() เป็นคำสั่งเพื่อให้แสดงรายละเอียดของชุดข้อมูลที่ระบุชื่อในวงเล็บ โดยระบุจำนวนข้อมูล จำนวนตัวแปร และแสดงรายละเอียดของตัวแปรทุกตัว ประกอบด้วยชื่อตัวแปร ชนิดของตัวแปร และแสดงค่าของตัวแปรประมาณ 10 แถวแรก

des() เป็นคำสั่งใน package epiDisplay เพื่อแสดงรายละเอียดของชุดข้อมูลที่ระบุชื่อในวงเล็บ โดยระบุจำนวนข้อมูล ลำดับของตัวแปร ชื่อตัวแปร ชนิดของตัวแปร และคำอธิบายของแต่ละตัวแปร การแสดงผลที่ได้จากการทำงานด้วยชุดคำสั่งการนำเข้าข้อมูล การตรวจสอบข้อมูล

หนังสือเล่มนี้ใช้ข้อมูลจากฐานข้อมูลสุขภาพ 43 แฟ้ม โดยใช้ 2 แฟ้มข้อมูลหลักคือ แฟ้ม person และ ncdscreen เนื่องจากเป็นชุดข้อมูลที่มีขนาดใหญ่จึงทำการคัดเลือกข้อมูลบุคคลที่มีอายุ 35 ปีขึ้นไป และทำการสุ่มโดยวิธีสุ่มอย่างง่าย (simple random sampling) จำนวน 200,000 แถว ในแต่ละชุดข้อมูล เพื่อให้ได้ชุดข้อมูลขนาดเล็ก ง่ายต่อการวิเคราะห์ โดยแฟ้ม person เป็นชนิด csv เก็บในชื่อ “person_35_random.csv” ส่วนชุดข้อมูล ncdscreen ที่ได้จากการสุ่มจำนวน 200,000 แถว ถูกเก็บในชื่อ “ncdscreen_35_random.csv” ซึ่งในบทนี้จะแสดงตัวอย่างการวิเคราะห์ ตั้งแต่การนำเข้าข้อมูล การตรวจสอบ การจัดการและการรวมชุดข้อมูล (Data merging) ซึ่งภายหลังการรวมชุดข้อมูลได้ทำการเก็บข้อมูลชุดใหม่ไว้ในชื่อ “person.csv” จะใช้ในบทถัดๆ ไป นอกจากนี้ ข้อมูลชื่อ “blevel.csv” เป็นการดึงข้อมูลบางตัวแปร ที่จะใช้มารวมไว้ด้วยกัน โดยจะใช้ข้อมูลชุดนี้ในบทที่ 8 ทั้งนี้ **ท่านสามารถดาวน์โหลดข้อมูลทุกชุดเพื่อฝึกปฏิบัติและทำแบบฝึกหัดท้ายบท รวมทั้งไฟล์หนังสือเล่มนี้ในรูปแบบ pdf ได้ที่ http://medipe2.psu.ac.th/2019/content?id=30**

แสดงตัวอย่างการนำเข้าข้อมูลโดยใช้ข้อมูล “person_35_random.csv” ดังนี้

```
ps<-read.csv(“person_35_random.csv”,head=T)

head(ps)
    X hospcode                              cid   pid  hid prename name
 1  2    5811 ea169af4ab68d985805f03baa8a6d994 10007 1798     003  \\N
 2  4    5811 73472b017dad8054b11dc4fb91f0d396 10010 2890     003  \\N
 3  8    5811 9e22d2eb57c968a3297c10748a76afad 10021 1660     005  \\N
 4 10    5811 c048668ac116cf23e729d1ce298c191a 10023 3615     004  \\N
 5 18    5811 55f0eb855b2356c68a8e291a8c9e2543 10034 1899     003  \\N
 6 26    5811 5b48a5a5fb288194b34a688cb3317fe5  1006 2847     004  \\N
```

```
  lname    hn sex      birth mstatus occold occnew race nation religion
1   \\N 10007   1 1962-11-29       1      2   9622   99     99        1
2   \\N 10010   1 1928-07-01       1    901   9999    0      0        1
3   \\N 10021   2 1967-08-24       2    901   9999    0      0        1
4   \\N 10023   2 1965-02-05       1      2   9622   99     99        1
5   \\N 10034   1 1979-02-14       1      2   9622   99     99        1
6   \\N  1006   2 1959-05-30       2      2   9622   99     99        1
  edu fstatus father mother couple vstatus     movein discharge ddischarge
1   4       2    \\N    \\N    \\N      NA 1962-11-29         9 0000-00-00
2   9       2    \\N    \\N    \\N      NA 1928-07-01         9 0000-00-00
3   9       2    \\N    \\N    \\N      NA 1967-08-24         9 0000-00-00
4   3       2    \\N    \\N    \\N      NA 1965-02-05         9 0000-00-00
5   5       2    \\N    \\N    \\N      NA 1979-02-14         9 0000-00-00
6   3       2    \\N    \\N    \\N      NA 1959-05-30         9 0000-00-00
  abogroup rhgroup labor passport typearea              update
1       NA      NA    NA       NA        1 2014-03-03 01:57:09
2       NA      NA    23       NA        1 2010-11-02 13:23:47
3       NA      NA    23       NA        1 2010-11-02 13:23:49
4       NA      NA    NA       NA        1 2014-10-24 10:34:15
5       NA      NA    NA       NA        1 2014-04-01 05:00:13
6       NA      NA    NA       NA        1 2013-12-12 08:53:18
names(ps)
  [1] "X"          "hospcode"    "cid"        "pid"        "hid"
  [6] "prename"    "name"        "lname"      "hn"         "sex"
 [11] "birth"      "mstatus"     "occold"     "occnew"     "race"
 [16] "nation"     "religion"    "edu"        "fstatus"    "father"
 [21] "mother"     "couple"      "vstatus"    "movein"     "discharge"
 [26] "ddischarge" "abogroup"    "rhgroup"    "labor"      "passport"
 [31] "typearea"   "update"
names(ps)<-tolower(names(ps))
str(ps)
 'data.frame':    200000 obs. of  33 variables:
  $ x         : int  2 4 8 10 18 26 30 32 33 38 ...
  $ hospcode   : int  5811 5811 5811 5811 5811 5811 5811 5811 5811 5811 ...
  $ cid       : Factor w/ 188626 levels "00003dadcb3298cee17d21a699dafafa",..:
172669 84701 116291 141931 63244 67081 58567 67159 132190 185773 ...
  $ pid       : int  10007 10010 10021 10023 10034 1006 10073 10080 10084 1010 ...
  $ hid       : int  1798 2890 1660 3615 1899 2847 2421 2745 757 1990 ...
  $ prename   : Factor w/ 77 levels "--","003","004",..: 2 2 4 3 2 3 2 2 4 4 ...
  $ name      : Factor w/ 1 level "\\N": 1 1 1 1 1 1 1 1 1 1 ...
  $ lname     : Factor w/ 1 level "\\N": 1 1 1 1 1 1 1 1 1 1 ...
  $ hn        : int  10007 10010 10021 10023 10034 1006 10073 10080 10084 1010 ...
  $ sex       : int  1 1 2 2 1 2 1 1 2 2 ...
  $ birth     : Factor w/ 19309 levels "1916-01-01","1916-01-02",..: 12703 1471
```

```
14432 13502 18624 11424 11315 14287 10502 15250 ...
  $ mstatus   : int  1 1 2 1 1 2 2 1 2 2 ...
  $ occold    : int  2 901 901 2 2 2 3 901 2 3 ...
  $ occnew    : int  9622 9999 9999 9622 9622 9622 5221 9999 9622 5221 ...
  $ race      : int  99 0 0 99 99 99 99 0 99 99 ...
  $ nation    : int  99 0 0 99 99 99 99 0 99 99 ...
  $ religion  : int  1 1 1 1 1 1 1 1 1 1 ...
  $ edu       : int  4 9 9 3 5 3 3 9 3 3 ...
  $ fstatus   : int  2 2 2 2 2 2 2 2 2 2 ...
  $ father    : Factor w/ 1 level "\\N": 1 1 1 1 1 1 1 1 1 1 ...
  $ mother    : Factor w/ 1 level "\\N": 1 1 1 1 1 1 1 1 1 1 ...
  $ couple    : Factor w/ 1 level "\\N": 1 1 1 1 1 1 1 1 1 1 ...
  $ vstatus   : int  NA NA NA NA NA NA NA NA NA NA ...
  $ movein    : Factor w/ 23097 levels "0000-00-00","0001-01-01",..: 11158 1074
12886 11957 17071 9879 9770 12741 8959 13701 ...
  $ discharge : int  9 9 9 9 9 9 9 9 9 9 ...
  $ ddischarge: Factor w/ 1586 levels "0000-00-00","1708-05-22",..: 1 1 1 1 1 1 1
1 1 1 ...
  $ abogroup  : int  NA NA NA NA NA NA NA NA NA 4 ...
  $ rhgroup   : int  NA NA NA NA NA NA NA NA NA NA ...
  $ labor     : int  NA 23 23 NA NA NA NA 23 NA NA ...
  $ passport  : logi  NA NA NA NA NA NA ...
  $ typearea  : int  1 1 1 1 1 1 1 1 1 1 ...
  $ update    : Factor w/ 115493 levels "1899-12-30 00:00:00",..: 71119 1595 1597
89013 72966 67343 74106 1595 67522 14066 ...

des (ps)
  No. of observations =  200000
    Variable      Class           Description
 1  x             integer
 2  hospcode       integer
 3  cid           factor
 4  pid           integer
 5  hid           integer
 6  prename       factor
 7  name          factor
 8  lname         factor
 9  hn            integer
 10 sex           integer
 11 birth         factor
 12 mstatus       integer
 13 occold        integer
 14 occnew        integer
 15 race          integer
```

```
16 nation        integer
17 religion      integer
18 edu           integer
19 fstatus       integer
20 father        factor
21 mother        factor
22 couple        factor
23 vstatus       integer
24 movein        factor
25 discharge     integer
26 ddischarge    factor
27 abogroup      integer
28 rhgroup       integer
29 labor         integer
30 passport      logical
31 typearea      integer
32 update        factor
```

จากการแสดงผลข้างต้น ที่นำเข้าข้อมูล “person_35_random.csv” แล้วให้ชื่อชุดข้อมูลเป็น ps เมื่อตรวจสอบข้อมูลเบื้องต้น พบว่ามีจำนวนทั้งหมด 200,000 แถว มีจำนวนตัวแปรทั้งหมด 32 ตัวแปร

5.4 การจัดการข้อมูลแฟ้ม person_35_random

การสร้างตัวแปรใหม่ที่ชื่อว่า hpid จากตัวแปร hospcode และ pid ด้วยคำสั่ง paste() และตรวจสอบข้อมูลที่ซ้ำด้วยคำสั่ง duplicated()

คำสั่ง paste()

เป็นการนำเอาค่าสังเกตของตัวแปร hospcode และ pid มาเรียงต่อกัน เป็นค่าสังเกตใหม่ และให้ชื่อตัวแปรใหม่เป็น hpid ทั้งนี้ การสร้างตัวแปรใหม่ จากตัวแปร hospcode และ pid เพื่อต้องการตรวจสอบการซ้ำซ้อนของบุคคลเดียวกันภายในสถานพยาบาลเดียวกัน การระบุ sep=”” เป็นการระบุเครื่องหมายที่ใช้คั่นระหว่างค่าสังเกตของตัวแปร 2 ตัวที่นำมาต่อกัน โดยในที่นี่ไม่มีการระบุเครื่องหมายใดๆ คั่นระหว่าง 2 ตัวแปร

```
ps$hpid<-paste(ps$hoscode,ps$pid,sep="")
```

คำสั่ง duplicated()

เป็นคำสั่งที่ใช้ตรวจสอบการซ้ำกันของค่าสังเกตภายใต้ตัวแปรที่ได้ระบุไว้ ดังตัวอย่างคำสั่ง

```
dup <- duplicated (ps$hpid)
```

เป็นการตรวจสอบการซ้ำซ้อนของค่าสังเกตในตัวแปร hpid ที่อยู่ในข้อมูล ps เมื่อตรวจสอบแล้วให้เก็บผลการตรวจสอบความซ้ำกันไว้ในวัตถุที่ชื่อว่า dup การดูผลการวิเคราะห์ความซ้ำซ้อนในครั้งนี้ใช้คำสั่ง table(dup) ซึ่งผลการวิเคราะห์แสดงในรูปตารางของ FALSE และ TRUE โดยที่ TRUE แสดงให้เห็นว่าค่าสังเกตหรือข้อมูลในตัวแปรมีความซ้ำซ้อนกัน

เนื่องจากข้อมูลที่นำมาใช้วิเคราะห์ในครั้งนี้ เป็นแฟ้มข้อมูล psncd ในระดับจังหวัด จึงมีการตรวจสอบความซ้ำกันของค่าสังเกตในตัวแปร cid เมื่อพบความซ้ำซ้อนกัน จะทำการตัดค่าสังเกตที่มีความซ้ำกันออก และ

สร้างชุดข้อมูลใหม่ในชื่อว่า ps2 ซึ่ง ชุดข้อมูล ps2 จะมีจำนวนข้อมูลเท่ากับจำนวนข้อมูลในแฟ้มข้อมูล ps ลบด้วยจำนวนข้อมูลที่มีความซ้ำกัน นอกจากนี้ผู้วิเคราะห์ข้อมูลต้องการวิเคราะห์ข้อมูลเฉพาะผู้ป่วยที่มีสัญชาติหรือเชื้อชาติไทย จึงกำหนดให้บุคคลที่มีตัวแปร race หรือ nation มีค่าเท่ากับ 99 เท่านั้น ให้คงอยู่ในชุดข้อมูล ps

การแสดงผลที่ได้จากการทำงานด้วยชุดคำสั่งการนำเข้าข้อมูล การตรวจสอบความซ้ำกันของค่าสังเกต

```
ps$hpid<-paste(ps$hoscode,ps$pid,sep="")
head(ps$hpid)
 [1] "581110007" "581110010" "581110021" "581110023" "581110034" "58111006"
# check duplication
dup<- duplicated(ps$hpid)
table(dup) # check repeated; False = correct
 dup
  FALSE   TRUE
 198890   1110
dup1<- duplicated(ps$cid)
table(dup1)
 dup1
  FALSE   TRUE
 188626  11374
# delete duplication
ps2<-ps[dup1==FALSE,]
dup2<- duplicated(ps2$hpid)
table(dup2)
 dup2
  FALSE   TRUE
 188618      8
ps3<-ps2[duplicated(ps2$hpid)==FALSE,]
dup3<- duplicated(ps3$hpid)
table(dup3) # check repeated; False = correct
 dup3
  FALSE
 188618
# Select only Thai people  (code99=thai)
str(ps3$race)
  int [1:188618] 99 0 0 99 99 99 99 0 99 99 ...
ps3<-ps3[ps3$race==99|ps3$nation==99,]

dup4<- duplicated(ps3$hpid)
table(dup4)
 dup4
  FALSE   TRUE
 171456    125
```

```
dup5<- duplicated(ps3$cid)
table(dup5)
 dup5
  FALSE   TRUE
 171456    125
ps4<-ps3[duplicated(ps3$hpid)==FALSE,]
table(duplicated(ps4$cid))

  FALSE
 171456
table(duplicated(ps4$hpid))

  FALSE
 171456
#remove ps, dup, dup1
rm(ps2, ps3, dup, dup1, dup2, dup3, dup4, dup5)
ps2<-ps4
rm(ps4)
```

การตรวจสอบความถูกต้องและการจัดการข้อมูลของตัวแปร sex

เนื่องจากตัวแปร sex จะมีค่าสังเกตได้เป็นเพศชายและเพศหญิงเท่านั้น จึงต้องมีการตรวจสอบหากพบการระบุเพศที่ไม่ถูกต้องจะให้ค่าเป็นไม่สามารถระบุได้ หรือ NA โดยการทำงานนี้ใช้คำสั่ง class() tab1() และ levels()

คำสั่ง class()

เป็นคำสั่งสำหรับตรวจสอบประเภทของตัวแปรที่ระบุ เช่น class(ps2$sex) เป็นการตรวจสอบประเภทของตัวแปร sex ซึ่งพบว่าตัวแปร sex เป็นตัวแปรประเภทจำนวนเต็ม (integer)

คำสั่ง tab1()

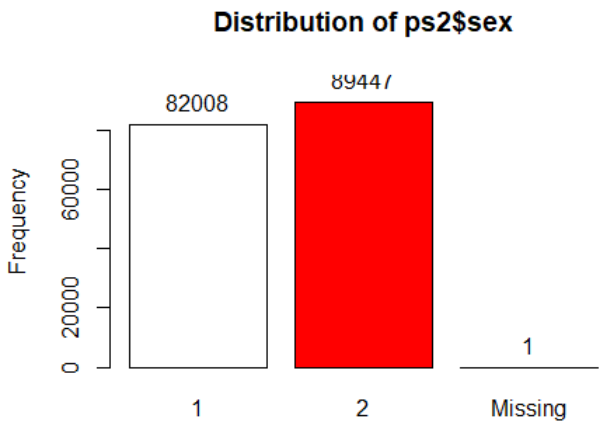
เป็นคำสั่งสำหรับการสร้างตารางทางเดียวของตัวแปรที่ระบุ เช่น tab1(ps2$sex) เป็นการสร้างตารางทางเดียวในตัวแปร sex ซึ่งควรจะมีแค่ 1 (เพศชาย) และ 2 (เพศหญิง) หากพบว่าข้อมูลในตัวแปร sex มีค่าสังเกตอื่นใด ซึ่งไม่ถูกต้อง ผู้วิเคราะห์ข้อมูลจึงสร้างตัวแปรใหม่ที่มีชื่อว่า gender โดยใช้ค่าสังเกตของตัวแปร sex โดยที่ให้ค่าสังเกตที่เป็นตัวเลข 1 และ 2 ยังเป็นเช่นเดิม แต่จะให้ค่าของตัวเลขที่ไม่ใช่ 1 และ 2 เป็น NA

คำสั่ง levels()

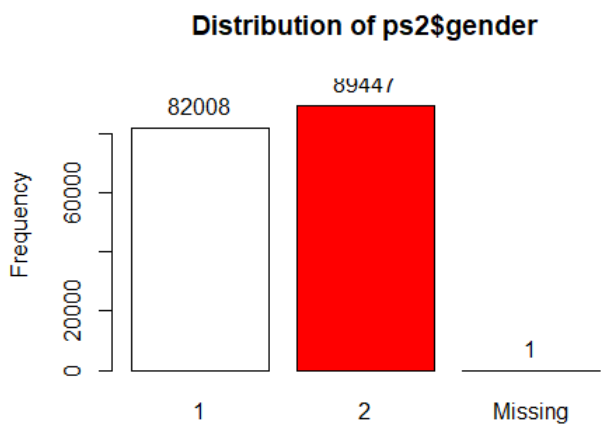
เป็นคำสั่งที่ใช้สำหรับให้ค่าของตัวเลขที่เป็นค่าสังเกตของตัวแปร ทั้งนี้ ตัวแปรที่จะสามารถใช้กับคำสั่ง levels() ได้ ต้องเป็นตัวแปรที่มีลักษณะเป็น factor เท่านั้น เช่น เมื่อสร้างตัวแปร gender แล้วจะระบุค่าตัวเลข 1 และ 2 เป็น male และ female ตามลำดับ

การแสดงผลที่ได้จากการตรวจสอบความถูกต้องและการจัดการข้อมูลของตัวแปร sex

```
# Sex management
class (ps2$sex)
 [1] "integer"
tab1(ps2$sex)
```




```
 ps2$sex :
         Frequency   %(NA+)   %(NA-)
 1           82008     47.8     47.8
 2           89447     52.2     52.2
 <NA>            1       0.0      0.0
   Total    171456    100.0    100.0
ps2$gender<-ifelse(!(ps2$sex%in%c(1,2)),NA,ps2$sex)
tab1(ps2$gender)
```



```
 ps2$gender :
         Frequency   %(NA+)   %(NA-)
 1           82008     47.8     47.8
 2           89447     52.2     52.2
 <NA>            1      0.0      0.0
   Total    171456    100.0    100.0
# definition in gender
ps2$gender<-factor(ps2$gender)
levels(ps2$gender)<-c("male","female")
tab1(ps2$gender, graph=F)
 ps2$gender :
         Frequency   %(NA+)   %(NA-)
 male        82008     47.8     47.8
 female      89447     52.2     52.2
 NA's            1      0.0      0.0
   Total    171456    100.0    100.0
```

การสร้างตัวแปรอายุ (age) จากตัวแปร birth

เนื่องจากชุดข้อมูลจากแฟ้ม person_35_random ไม่มีตัวแปรอายุ จึงต้องสร้างตัวแปรอายุขึ้นมา เพื่อใช้ในการวิเคราะห์ข้อมูลต่อไป ตัวแปรอายุที่สร้างขึ้นใหม่ สร้างมาจากตัวแปร birth ซึ่งเป็นตัวแปร วัน-เดือน-ปีที่เกิด เริ่มต้นการสร้างตัวแปรอายุ โดยการตรวจสอบประเภทของตัวแปร birth ก่อนด้วยคำสั่ง class() และดูรูปแบบของค่าสังเกตด้วยคำสั่ง head() ค่าสังเกตของตัวแปร birth อาจจะเป็นได้หลายรูปแบบ เช่น year-month-day (1974-12-13) หรือ yearmonthday (19741213) หรือ year/month/ day (1974/12/13)

เมื่อตรวจสอบตัวแปร birth แล้ว จะทำการระบุให้ตัวแปร birth เป็นตัวแปรประเภทวันที่ (Date) โดยใช้คำสั่ง as.Date() และเพื่อให้คงตัวแปร birth ไว้ จึงต้องสร้างตัวแปรใหม่ชื่อว่า bdate นอกจากนี้สร้างตัวแปรวันที่ที่เป็นปัจจุบัน คือ ตัวแปร today หรืออาจจะเป็นวันที่ของกลางปีของข้อมูลจากแฟ้มข้อมูล person_35_random เพื่อนำตัวแปรวันเดือนปีเกิดที่สร้างขึ้นใหม่ (bdate) มาลบกับตัวแปร today ก็จะได้ตัวแปรอายุ (age) ที่สร้างขึ้นมาใหม่

คำสั่ง as.Date()

เป็นคำสั่งเพื่อระบุให้ตัวแปรเป็นประเภทวันที่ ทั้งนี้ ต้องระบุรูปแบบของค่าสังเกต ภายใต้ตัวแปรให้ถูกต้อง เช่น หากรูปแบบของค่าสังเกตเป็น 19311105 ต้องระบุเป็น %Y%m%d หรือหากรูปแบบของค่าสังเกตเป็น 1931-11-05 ให้ระบุเป็น %Y-%m-%d

ข้อพึงระวังในการใช้คำสั่ง as.Date() คือ หากตัวแปรที่ต้องการระบุเป็นตัวแปรประเภทอื่น ที่ไม่ใช่ตัวแปรประเภท factor หรือ character เช่น เป็นประเภทจำนวนเต็ม (integer) ต้องมีการเปลี่ยนประเภทของตัวแปรจาก integer ให้เป็น character ก่อนด้วยคำสั่ง as.character()

การแสดงผลที่ได้จากการสร้างตัวแปรอายุ (age)

```
# birth management and generate age
class(ps2$birth)
 [1] "factor"
head(ps2$birth)  #  birth-date style => year-month-day
 [1] 1962-11-29 1965-02-05 1979-02-14 1959-05-30 1959-02-10 1956-11-19
 19309 Levels: 1916-01-01 1916-01-02 1916-01-04 1916-01-08 ... 1980-12-30
# generate  age  variable
ps2$bdate<-as.Date(ps2$birth,format="%Y-%m-%d")
ps2$today<-as.Date("2019-01-01", format("%Y-%m-%d"))
ps2$age<-(ps2$today-ps2$bdate)/365.25

# if birth is not character-> as.character
# ps2$bdate<-as.Date(as.character(ps$birth),format="%Y%m%d")

head(ps2$age)
 Time differences in days
 [1] 56.09035 53.90281 39.87953 59.59206 59.89049 62.11636
tail(ps2$age)
 Time differences in days
 [1] 43.49076 64.83504 45.99863 44.63244 61.17454 95.00068
```

การตรวจสอบความถูกต้องของตัวแปรอายุ (age)

ตัวแปรอายุ (age) ที่สร้างขึ้นใหม่นั้น เมื่อทำการตรวจสอบประเภทของตัวแปรด้วยคำสั่ง class() พบว่ามีประเภทของตัวแปรเป็น "difftime" ซึ่งเป็นตัวแปรที่เป็นความแตกต่างของเวลา จึงต้องเปลี่ยนประเภทของตัวแปรให้เป็นตัวเลข (numeric) ด้วยคำสั่ง as.numeric() เพื่อที่จะสามารถใช้ในการสำรวจและอธิบายตัวแปรอายุ

การอธิบายตัวแปรอายุ ด้วยการหาค่าต่ำสุด สูงสุด ค่าเฉลี่ย เป็นวิธีการหนึ่งในการตรวจสอบความถูกต้องของข้อมูลในตัวแปรอายุ โดยการใช้คำสั่ง summ() หากพบว่าอายุของผู้ป่วยในชุดข้อมูล person_35_random นี้มีค่าอายุที่ไม่สามารถเป็นได้ เช่นอายุเกิน 100 ปี หรือ น้อยกว่า 0 ก็จำเป็นต้องมีการจัดการให้ค่าเหล่านั้นเป็น NA ด้วยคำสั่ง ifelse() และสร้างตัวแปรใหม่ชื่อตัวแปร agenew ที่มีค่าอายุอยู่ระหว่าง 0 ถึง 100 ปี

คำสั่ง as.numeric()

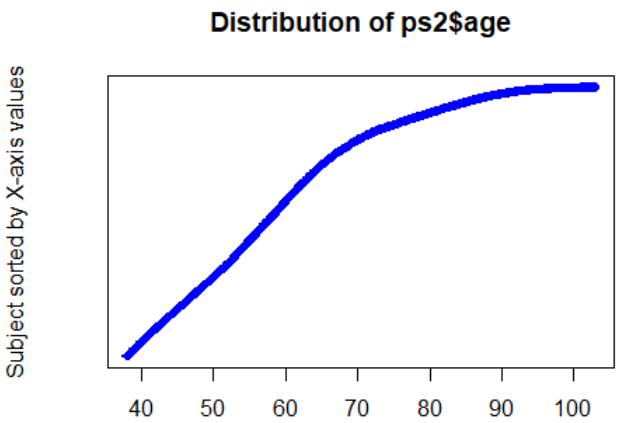
เป็นคำสั่งสำหรับระบุให้ตัวแปรเป็นประเภทตัวเลขหรือ numeric

คำสั่ง summ()

เป็นคำสั่งสำหรับอธิบายค่าสถิติต่างๆ เช่น ค่าต่ำสุด สูงสุด ค่าเฉลี่ย ค่าเบี่ยงเบนมาตรฐาน ของตัวแปรที่ต้องการ สำหรับตรวจสอบข้อมูลที่เป็นตัวเลข พร้อมกันนี้จะมีการสร้างกราฟแสดงการกระจายตัวของข้อมูล หากไม่ต้องการให้สร้างกราฟ ต้องระบุ graph=F ลงไป

การแสดงผลที่ได้จากการตรวจสอบความถูกต้องของตัวแปรอายุ

```
class(ps2$age)
 [1] "difftime"
ps2$age<-as.numeric(ps2$age)
head(ps2$age)
 [1] 56.09035 53.90281 39.87953 59.59206 59.89049 62.11636
tail(ps2$age)
 [1] 43.49076 64.83504 45.99863 44.63244 61.17454 95.00068
summ(ps2$age,graph=F)
  obs.   mean   median  s.d.   min.   max.
  171455 58.988 57.492  13.738 38.004 103.001
summ(ps2$age)
```




```
  obs.   mean   median  s.d.   min.   max.
  171455 58.988 57.492  13.738 38.004 103.001
# สร้างตัวแปร agenew --- แทนที่ตัวแปร age ที่<0 & >100 ด้วย NA
ps2$agenew<-ifelse(ps2$age<=0,NA,ps2$age)
ps2$agenew<-ifelse(ps2$agenew>110,NA,ps2$agenew)

summ(ps2$agenew, graph = F)
  obs.   mean   median  s.d.   min.   max.
  171455 58.988 57.492  13.738 38.004 103.001
```

การสร้างกราฟพีระมิดระหว่างตัวแปรเพศและอายุ

กราฟพีระมิดเป็นกราฟที่แสดงให้เห็นลักษณะเพศและอายุของผู้ป่วย โดยใช้คำสั่ง pyramid() หากต้องการปิดกราฟ จะใช้คำสั่ง dev.off()

คำสั่ง pyramid()

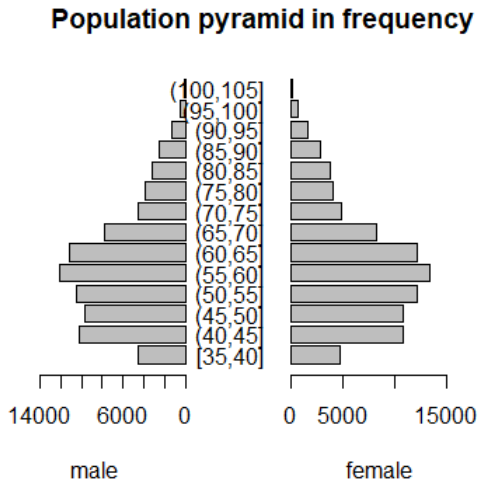
เป็นคำสั่งสำหรับการสร้างกราฟพีระมิด หากต้องการระบุความกว้างของชั้นจะระบุ binwidth ลงในคำสั่ง เช่น สร้างกราฟพีระมิดระหว่างตัวแปรอายุ (agenew) และตัวแปรเพศ (gender) โดยต้องการระบุความกว้างของช่วงอายุให้อยู่ในช่วง 10 ปี ก็จะระบุ binwidth=10 ในคำสั่งนี้สามารถระบุสีให้แต่ละค่าของคอลัมน์ ได้ เช่น ระบุสีของเพศ (col.gender=) นอกจากสามารถระบุชื่อกราฟโดยใช้ main=” ” และหากต้องการระบุค่าร้อยละในทุกชั้นก็เพิ่ม percent= each

คำสั่ง dev.off()

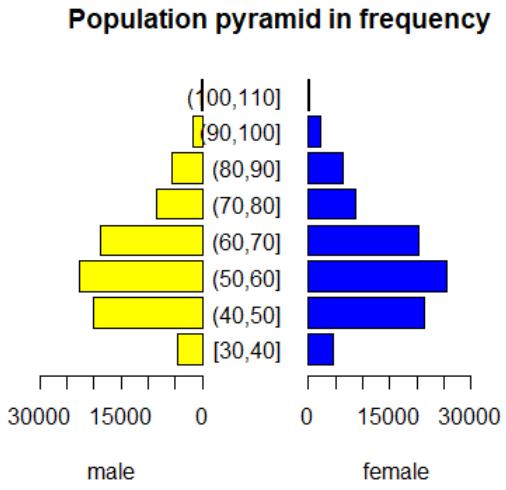
เป็นคำสั่งสำหรับให้ปิดกราฟที่แสดงในขณะนั้น

การแสดงผลที่ได้จากการสร้างกราฟพีระมิด

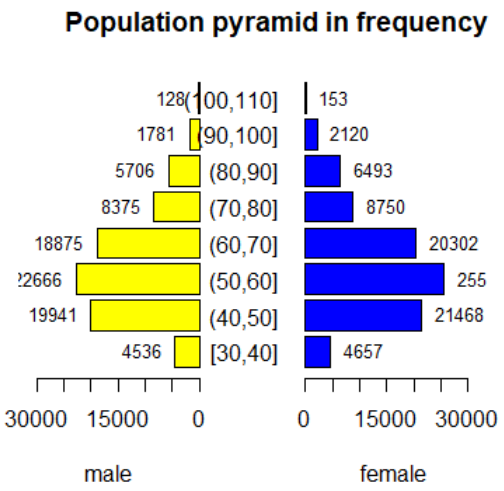
```
# pyramid
pyramid(ps2$agenew,ps2$gender)
```



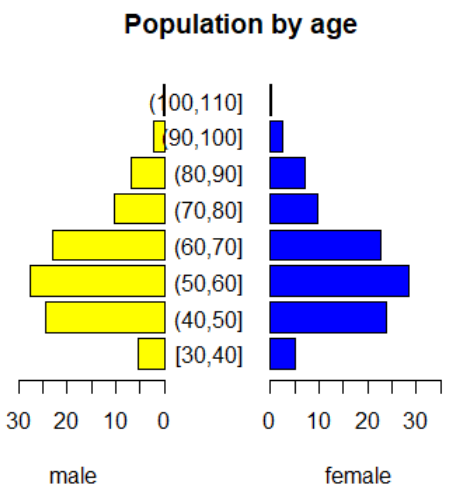

```
pyramid(ps2$agenew,ps2$gender,binwidth = 10,
        col.gender = c(“yellow”,”blue”))
```



```
pyramid(ps2$agenew,ps2$gender,binwidth = 10,
        bar.label=T,col.gender=c("yellow","blue"))
```




```
pyramid(ps2$agenew,ps2$gender,binwidth = 10,
        percent="each",printTable=T, main="Population by age",
        col.gender=c("yellow","blue"))
```




```
 Tabulation of age by sex (percentage of each gender).
          male female
[30,40]    5.5    5.2
(40,50]   24.3   24.0
(50,60]   27.6   28.5
(60,70]   23.0   22.7
(70,80]   10.2    9.8
(80,90]    7.0    7.3
(90,100]   2.2    2.4
(100,110]  0.2    0.2
dev.off() # close graph
null device
```

การจัดแบ่งกลุ่มอายุ

หากต้องการแบ่งกลุ่มอายุจากตัวแปรอายุเดิม ให้เป็นตัวแปรใหม่โดยให้อยู่ในรูปแบบของช่วงอายุก็ สามารถทำได้ โดยใช้คำสั่ง cut() ทั้งนี้ เมื่อแบ่งกลุ่มเป็นช่วงอายุจะสร้างตัวแปรใหม่ ชื่อ age.gr เมื่อสร้างตัวใหม่ขึ้นมาแล้ว ทำการตรวจสอบตัวแปร age.gr โดยการสร้างตาราง ด้วยคำสั่ง tab1(), table() หรือ tabpct() ทั้งนี้ แล้วแต่ความต้องการผลลัพธ์จากการวิเคราะห์ข้อมูล

คำสั่ง cut()

เป็นคำสั่งสำหรับใช้แบ่งกลุ่มค่าสังเกตของตัวแปร cut() โดยข้อมูลที่ได้จะเป็นกลุ่ม เช่น แบ่งช่วงอายุออกเป็น 6 ช่วงอายุ คือ 0-20, 20-30, 30-40, 40-50, 50-60 และ 60-100 การกำหนดช่วงจะระบุด้วย breaks= c() การอ่านค่าความหมายของเครื่องหมายวงเล็บ จากตัวอย่าง (0, 20] มีความหมายว่า อายุตั้งแต่มากกว่า 0 จนถึง 20 ปี คือไม่รวม 0 แต่นับรวม อายุ 20 ปี ตัวแปรใหม่ที่สร้างจากคำสั่งนี้ จะเป็นตัวแปรประเภทกลุ่ม (factor)

คำสั่ง table () และ tabpct()

เป็นคำสั่งสำหรับสร้างตาราง 2 ทาง ระหว่าง 2 ตัวแปร โดยตัวแปรแรกเป็นตัวแปรตามแถว (row) ตัวแปรตัวที่ 2 เป็นตัวแปรตามคอลัมน์ (column) เช่น table (gender, age.gr) หรือ tabpct (gender, age.gr) เป็นการสร้างตาราง 2 ทางระหว่างตัวแปรเพศ และอายุ โดยเพศจะอยู่ในแนวแถว และอายุจะอยู่ในแนวคอลัมน์

ผลลัพธ์ที่ได้จากการใช้คำสั่ง table () และ tabpct() จะต่างกัน โดยคำสั่ง table() จะได้ผลลัพธ์คือตาราง 2 ทาง ที่แสดงความถี่ในแต่ละช่อง (เซลล์: cell) แต่คำสั่ง tabpct() จะให้ผลลัพธ์ที่มากกว่า โดยเพิ่มการแสดงแสดงค่าร้อยละตามแถว และตามคอลัมน์

การแสดงผลที่ได้จากการแบ่งกลุ่มอายุ

```
#การสร้างตัวแปร age.gr จากตัวแปร agenew ด้วยคำสั่ง cut
ps2$age.gr<-cut(ps2$agenew, breaks = c(35,40,50,60,110))

tab1(ps2$age.gr)
 ps2$age.gr :
          Frequency   %(NA+)   %(NA-)
 (35,40]       9193      5.4      5.4
 (40,50]      41409     24.2     24.2
 (50,60]      48170     28.1     28.1
 (60,110]     72683     42.4     42.4
 NA's             1      0.0      0.0
   Total     171456    100.0    100.0

 สร้างตาราง age group &gender ด้วยคำสั่ง tabpct
tabpct(ps2$gender, ps2$age.gr)

 Original table
           ps2$age.gr
 ps2$gender  (35,40]  (40,50]  (50,60]  (60,110]   Total
```

```
    male        4536    19941    22666    34865   82008
    female      4657    21468    25504    37818   89447
    Total       9193    41409    48170    72683  171455

Row percent
          ps2$age.gr
ps2$gender  (35,40]  (40,50]  (50,60]  (60,110]  Total
    male       4536    19941    22666     34865  82008
              (5.5)   (24.3)   (27.6)    (42.5)  (100)
    female     4657    21468    25504     37818  89447
              (5.2)     (24)   (28.5)    (42.3)  (100)

Column percent
          ps2$age.gr
ps2$gender  (35,40]       %  (40,50]       %  (50,60]       %  (60,110]
    male       4536  (49.3)    19941  (48.2)    22666  (47.1)     34865
    female     4657  (50.7)    21468  (51.8)    25504  (52.9)     37818
    Total      9193   (100)    41409   (100)    48170   (100)     72683
          ps2$age.gr
ps2$gender      %
    male     (48)
    female   (52)
    Total   (100)
```

5.5 การนำเข้าและการจัดการข้อมูลจากแฟ้ม ncdscreen_35_random

การนำข้อมูลจากชุดข้อมูล ncdscreen เข้าสู่โปรแกรม R และการตรวจสอบข้อมูลที่นำเข้าสามารถทำได้โดยใช้คำสั่งเดียวกันกับการนำเข้าและการตรวจสอบข้อมูลจากไฟล์ข้อมูล person_35_random และในชุดข้อมูล ncdscreen_35_random ก็จะมีการสร้างตัวแปร hpid จากตัวแปร hospcode และ pid เช่นเดียวกัน ทั้งนี้จะอธิบายวัตถุประสงค์ในการทำงานของคำสั่งภายใต้หัวข้อการแสดงผลลัพธ์ที่จะแสดงผลในต่อไปนี้

การแสดงผลที่ได้จากการนำเข้าและการจัดการข้อมูลจากแฟ้ม ncdscreen_35_random

```
#import ncdscreen data only 200000 record
#ncd screen data
ncd<-read.csv("ncdscreen_35_random.csv", head=T, sep=",")

# หากไฟล์ข้อมูล ncdscreen ไม่ได้ระบุชื่อตัวแปร ต้องใช้คำสั่ง names () ในการให้ชื่อตัวแปร
#ในครั้งนี้เนื่องจาก ชุดข้อมูล ncdscreen ได้กำหนดชื่อตัวแปรแล้ว จึงใส่เครื่องหมาย
# ข้างหน้าคำสั่งทุกบรรทัด เพื่อไม่ให้มีการทำงานในคำสั่งนี้
#names(ncd)[1:21] <- c("hospcode","pid","seq","dateserv",
#                       "servplace","smoke","alcohol",
#                       "dmfamily","htfamily","weight","height",
#                      "waist","sbp1","dbp1","sbp2","dbp2","bslevel",
```

```
#                    "bstest","screenplace","provider","update")

des(ncd)

  No. of observations =  38503
    Variable      Class           Description
 1  X            integer
 2  hospcode      integer
 3  pid          integer
 4  seq          integer
 5  dateserv     factor
 6  servplace    integer
 7  smoke        integer
 8  alcohol      integer
 9  dmfamily     integer
 10 htfamily     integer
 11 weight       integer
 12 height       integer
 13 waist        integer
 14 sbp1         integer
 15 dbp1         integer
 16 sbp2         integer
 17 dbp2         integer
 18 bslevel      integer
 19 bstest       integer
 20 screenplace  integer
 21 provider     factor
 22 update       factor
 23 hcpid        numeric
names(ncd) <- tolower(names(ncd))

# สร้างตัวแปร hpid จาก hospcode & pid
ncd$hpid <- paste0(ncd$hospcode,ncd$pid)
head(ncd$hpid)
 [1] "581110007" "581110116" "581110403" "581110787" "581110872" "581111601"
# ตรวจสอบจำนวนครั้งของการเข้ารับการรักษาของผู้ป่วย
 #table(ncd$hpid)
#บรรทัดนี้เป็นการตรวจสอบจำนวนครั้งของการเข้ารับการรักษาต่อผู้ป่วย 1 คน
#เนื่องจากผลลัพธ์ที่ได้จากการใช้คำสั่งบรรทัดนี้ จะมีจำนวนหลักหมื่น จึงใส่เครื่องหมาย
# เพื่อให้คำสั่งไม่ทำงาน
table(table(ncd$hpid))
 #การสร้างตารางเพื่อตรวจสอบจำนวนผู้ป่วยที่เข้ามารับบริการตามจำนวนครั้ง
```

```
1      2      3      4
 33044  2549   111      7
# การเรียกดูข้อมูลตัวแปร dateserv โดยดูเฉพาะ 10 แถวแรก
ncd$dateserv[1:10]
  [1] 2015-03-31 2015-03-25 2015-03-31 2015-03-27 2015-01-15 2015-03-25
  [7] 2015-03-24 2015-03-25 2015-03-10 2015-03-24
 337 Levels: 2015-01-01 2015-01-02 2015-01-03 2015-01-04 ... 2015-12-28
# เปลี่ยนรูปแบบของตัวแปร dateserv จากรูปแบบ character ให้เป็นรูปแบบวันที่ (date) โดยให้ชื่อ dserv
ncd$dserv <- as.Date(ncd$dateserv,"%Y-%m-%d")

# สร้างตัวแปรใหม่ "yr" จากตัวแปร dserv โดยตัดเฉพาะปี พ.ศ.
ncd$yr <- year(ncd$dserv)
head(ncd$yr)
 [1] 2015 2015 2015 2015 2015 2015
# สร้างตัวแปร dup จากตัวแปร  hpid yr smoke และ alcohol สำหรับตรวจการซ้ำซ้อนของข้อมูล
ncd$dup <- paste(ncd$hpid,ncd$yr,ncd$smoke,ncd$alcohol)
head(ncd$dup)
 [1] "581110007 2015 NA NA" "581110116 2015 NA NA" "581110403 2015 NA NA"
 [4] "581110787 2015 NA NA" "581110872 2015 NA NA" "581111601 2015 NA NA"
tail(ncd$dup)
 [1] "65795025 2015 1 1" "65795090 2015 1 1" "65795135 2015 1 1"
 [4] "65795158 2015 1 1" "65795163 2015 1 1" "65795355 2015 1 1"
# ตรวจสอบความซ้ำกันของข้อมูลจากตัวแปร
table(duplicated(ncd$dup))

 FALSE  TRUE
 35752  2751
# ตัดข้อมูลที่มีความซ้ำกัน แล้วเก็บข้อมูลที่เหลืออยู่ในไฟล์ข้อมูลที่ชื่อว่า ncd1
ncd1 <- ncd[!duplicated(ncd$dup),]
```

5.6 การรวมชุดข้อมูล person_35_random กับ ncdscreen_35_random

การรวมชุดข้อมูลจากแฟ้ม person_35_random และ ncdscreen_35_random เพื่อให้เป็นชุดข้อมูลเดียวกัน ด้วยคำสั่ง merge() โดยการใช้ตัวแปร hpid ซึ่งได้สร้างขึ้นไว้แล้วในชุดข้อมูล person_35_random และ ncdscreen_35_random เมื่อรวมข้อมูลทั้ง 2 ชุด เข้าด้วยกันแล้ว สามารถบันทึกข้อมูลชุดใหม่โดยใช้คำสั่ง write.csv() ซึ่งจะได้ไฟล์ข้อมูลชุดใหม่ที่มีนามสกุลเป็น csv และเก็บไว้ในโฟลเดอร์หรือ directory เดียวกับไฟล์ข้อมูล

เมื่อรวมข้อมูลเข้าด้วยกันแล้ว ทำการตรวจสอบความซ้ำกันของข้อมูล จากตัวแปร hpid หรือ cid หากพบว่ามีความซ้ำกันของข้อมูล ก็ตัดข้อมูลที่ซ้ำกันและเก็บข้อมูลชุดใหม่ในชื่อ psncd

คำสั่ง merge()

เป็นคำสั่งสำหรับการรวมชุดข้อมูลเข้าด้วยกัน ทั้งนี้ ต้องมีการระบุตัวแปรที่จะใช้ในการเชื่อมข้อมูล ตัวอย่างการใช้คำสั่ง

```
psncd <- merge (ps2, ncd1, by.x="hpid",by.y="hpid") หรือ
psncd <- merge (ps2,ncd, by=c("hpid"))
```

คำสั่ง write.csv()

เป็นคำสั่งสำหรับการบันทึกข้อมูลในโปรแกรม R โดยไฟล์ที่บันทึกเป็นลักษณะของ csv

การแสดงผลจากการรวมไฟล์ person_35_random และ ncdscreen_35_random

```
# Merging person & screen

psncd <- merge(ncd1,ps2,by.x="hpid",by.y="hpid")
des(psncd)

  No. of observations =  34710
    Variable       Class           Description
 1  hpid           character
 2  x.x            integer
 3  hospcode.x      integer
 4  pid.x          integer
 5  seq            integer
 6  dateserv       factor
 7  servplace      integer
 8  smoke          integer
 9  alcohol        integer
 10 dmfamily       integer
 11 htfamily       integer
 12 weight         integer
 13 height         integer
 14 waist          integer
 15 sbp1           integer
 16 dbp1           integer
 17 sbp2           integer
 18 dbp2           integer
 19 bslevel        integer
 20 bstest         integer
 21 screenplace    integer
 22 provider       factor
 23 update.x       factor
 24 hcpid          numeric
 25 dserv          Date
 26 yr             integer
 27 dup            character
 28 x.y            integer
```

```
 29 hospcode.y     integer
 30 cid           factor
 31 pid.y         integer
 32 hid           integer
 33 prename       factor
 34 name          factor
 35 lname         factor
 36 hn            integer
 37 sex           integer
 38 birth         factor
 39 mstatus       integer
 40 occold        integer
 41 occnew        integer
 42 race          integer
 43 nation        integer
 44 religion      integer
 45 edu           integer
 46 fstatus       integer
 47 father        factor
 48 mother        factor
 49 couple        factor
 50 vstatus       integer
 51 movein        factor
 52 discharge     integer
 53 ddischarge    factor
 54 abogroup      integer
 55 rhgroup       integer
 56 labor         integer
 57 passport      logical
 58 typearea      integer
 59 update.y      factor
 60 gender        factor
 61 bdate         Date
 62 today         Date
 63 age           numeric
 64 agenew        numeric
 65 age.gr        factor
#check data duplication

table(duplicated(psncd$dup))

 FALSE
 34710
```

```
table(duplicated(psncd$hpid))

 FALSE  TRUE
 34671    39
table(duplicated(psncd$cid))

 FALSE  TRUE
 34671    39
#ลบข้อมูลที่ซ้ำกัน แล้วเก็บข้อมูลในชื่อว่า psncd 1
psncd1<- psncd[duplicated(psncd$hpid)==FALSE,]
table(duplicated(psncd1$hpid))

 FALSE
 34671
# บันทึกข้อมูล psncd1 ในรูปแบบของ csv
write.csv(psncd1, file="psncd.csv")
dir()
```

5.7 แบบฝึกหัดท้ายบท

ให้ผู้อ่านนำข้อมูลที่ตนเองมีอยู่หรืออาจเป็นแฟ้มข้อมูล person ที่ตนเองดูแลอยู่ มาฝึกใช้ในการนำเข้าและจัดการข้อมูล ดังนี้

1. แปลงข้อมูลที่มีอยู่ให้เป็นไฟล์ .csv หรือ .txt เพื่อใช้ในการนำเข้าโปรแกรม R
2. นำเข้าชุดข้อมูลและกำหนดให้สามารถอ่านข้อมูลที่เป็นภาษาไทยได้
3. ทำการสำรวจข้อมูลว่ามีจำนวนกี่แถว กี่คอลัมน์ กี่ตัวแปร และมีตัวแปรอะไรบ้าง
4. ทำการตรวจสอบความซ้ำซ้อนของ id และทำการตัดข้อมูลที่ซ้ำกันทิ้ง แบ่งกลุ่มข้อมูลอายุ และให้เหตุผลว่ายึดหลักใดในความแบ่งช่วงความถี่ หากไม่มีตัวแปรอายุให้ทำการสร้างตัวแปรอายุโดยใช้ตัวแปรวันเดือนปีเกิดตามตัวอย่างที่แสดงไว้ข้างต้น

บทที่ 6

สถิติเชิงพรรณนา (Descriptive Statistics)

บทที่ 6

สถิติเชิงพรรณนา (Descriptive Statistics)

ดร.นิรันดร์ อินทรัตน์
E-mail: nirun.i@msu.ac.th

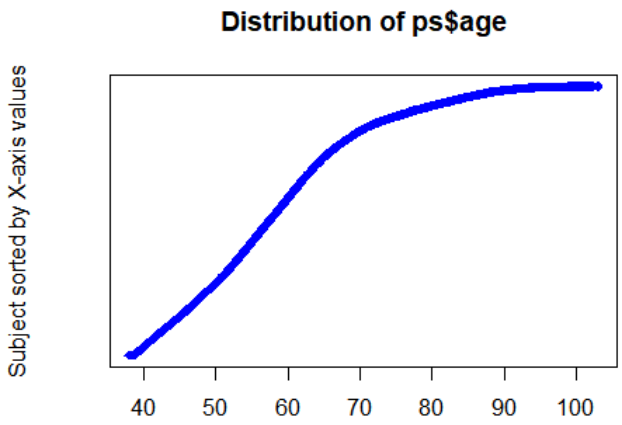
สถิติเชิงพรรณนา คือ สถิติที่ใช้เพื่ออธิบาย บรรยาย หรือสรุป ลักษณะของกลุ่มข้อมูลที่เป็นตัวเลข ที่เก็บรวบรวมมาซึ่งไม่สามารถอ้างอิงลักษณะประชากรได้ การพรรณนาข้อมูลจะต้องระบุตัวแปรให้ได้ก่อนว่ามีลักษณะใด โดยถ้าเป็นตัวแปรกลุ่มจะใช้ค่าความถี่และร้อยละในการพรรณนาข้อมูล ในกรณีที่ข้อมูลมีลักษณะต่อเนื่อง ถ้ามีการแจกแจงปกติจะนิยมรายงานค่าเฉลี่ย (ส่วนเบี่ยงเบนมาตรฐาน) ในกรณีที่ข้อมูลแจกแจงไม่ปกติ จะนิยมรายงานเป็นค่ากลาง (ค่าต่ำสุด ค่าสูงสุด) หรือ ค่ากลาง (ค่าควอไทล์ที่ 1 ค่าควอไทล์ที่ 3) เป็นต้น

ในบทนี้จะแสดงตัวอย่างการพรรณนาข้อมูล โดยใช้ชุดข้อมูลชื่อ psncd ที่ได้จากการรวมกันระหว่างข้อมูล person_35_random และ ncdscreen_35_random เป็นรูปแบบ csv มีจำนวน 34,671 แถว ดังนี้

6.1 การพรรณนาข้อมูลตัวแปรต่อเนื่อง

การพรรณนาข้อมูลใน 43 แฟ้ม ในตัวอย่างนี้จะทำการพรรณนาข้อมูลตัวแปรอายุ

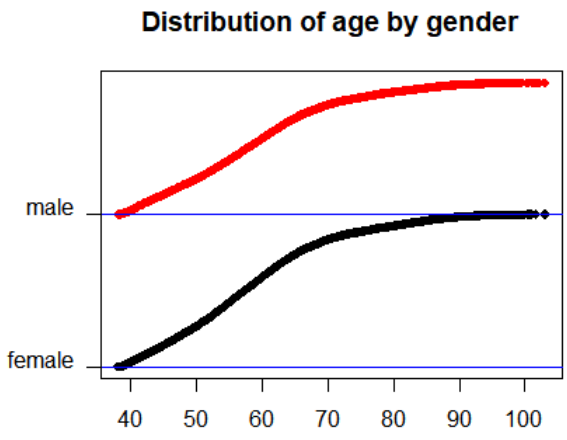
```
setwd("D:/data")
ps <- read.csv("psncd.csv",head=T)
summ(ps$age)
  obs.  mean   median  s.d.   min.   max.
  34671 58.674 57.492  12.453 38.031 103.001
```



จากข้อมูลอายุพบว่ามีค่าเฉลี่ยอยู่ที่ 58.67 ปี และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 12.45 และจากกราฟแสดงให้เห็นว่าข้อมูลอายุมีการเพิ่มขึ้น ทั้งนี้ สามารถแสดงผลข้อมูลอายุ โดยแยกตามเพศได้เช่นกัน

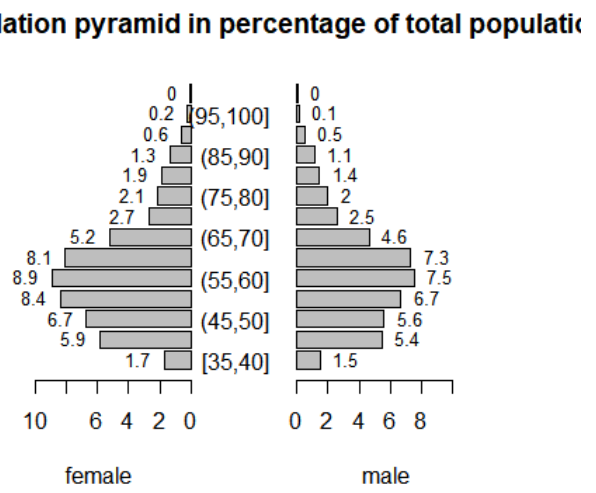
```
with(ps, summ(age,by=gender))
 For gender = female
  obs.  mean   median  s.d.   min.   max.
  18611 58.679 57.358  12.458 38.031 103.001

 For gender = male
  obs.  mean   median  s.d.   min.   max.
  16060 58.668 57.596  12.447 38.07  103.001
```



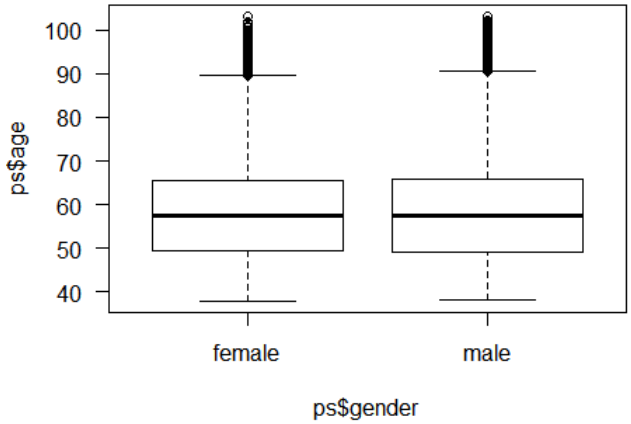
การวิเคราะห์ผลที่ได้ พบว่าเมื่อแบ่งอายุตามเพศแล้วพบว่า เพศหญิงมีค่าเฉลี่ยอยู่ที่ 58.68 ปี และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 12.46 ส่วนเพศชายมีค่าเฉลี่ยอยู่ที่ 58.67 ปี และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 12.45 และกราฟได้แสดงการเพิ่มขึ้นของอายุโดยแบ่งเป็นเพศหญิงและชาย นอกจากนั้น สามารถพรรณนาข้อมูลโดยใช้กราฟพีระมิดได้ ดังนี้

```
pyramid(ps$age,ps$gender,bar.label = TRUE,percent="total",printTable = T )
  Tabulation of age by sex (percentage of total population).
          female male
 [35,40]     1.7  1.5
 (40,45]     5.9  5.4
 (45,50]     6.7  5.6
 (50,55]     8.4  6.7
 (55,60]     8.9  7.5
 (60,65]     8.1  7.3
 (65,70]     5.2  4.6
 (70,75]     2.7  2.5
 (75,80]     2.1  2.0
 (80,85]     1.9  1.4
 (85,90]     1.3  1.1
 (90,95]     0.6  0.5
 (95,100]    0.2  0.1
     (100,105]   0.0  0.0
```



จากการแสดงกราฟพีระมิด พบว่า ประชากรเพศหญิงและชายมีการกระจายที่ใกล้เคียงกัน คืออยู่ในช่วง 46-70 ปี และมีลักษณะเบ้ขวา การแสดงกราฟข้อมูลเมื่อต้องการเปรียบเทียบค่าเฉลี่ยสองกลุ่มโดยการใช้ Box-and-Whisker Plots หรือแผนภาพกล่อง ดังนี้

```
boxplot(ps$age~ps$gender,las=1)
```



คำสั่ง las = 1 คือ การตั้งค่าให้แสดงค่าความถี่เป็นแนวตั้ง โดยกราฟที่แสดงคือ ข้อมูลอายุแบ่งเป็นเพศหญิงและเพศชาย จะพบว่า ค่ากลางอายุของเพศหญิงและชายคือ 58 ปี ซึ่งข้อมูลอายุส่วนใหญ่จะอยู่ในช่วง 50 – 65 ปี แต่มีบางส่วนที่มีอายุมากกว่า 90 ปี

6.2 การพรรณนาข้อมูลตัวแปรกลุ่ม

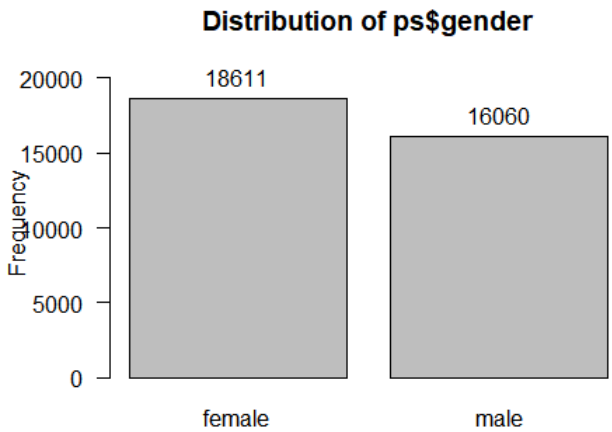
ในกรณีที่ข้อมูลมีลักษณะเป็นตัวแปรกลุ่มเราสามารถพรรณนาข้อมูลโดยการใช้ค่าความถี่ และร้อยละในการรายงานผล

6.2.1 การวิเคราะห์ข้อมูลแบบตารางทางเดียว

เช่น ต้องการคำนวณค่าความถี่และร้อยละของตัวแปรเพศ

```
tab1(ps$gender, col = "gray", las=1)
 ps$gender :
        Frequency Percent Cum. percent
 female      18611    53.7        53.7
 male        16060    46.3       100.0
   Total     34671   100.0       100.0
```

จากผลลัพธ์ได้แสดงจำนวนเพศหญิง เพศชาย เปอร์เซ็นต์แต่ละเพศ รวมถึงเปอร์เซ็นต์สะสม การวิเคราะห์ข้อมูลพบว่า เป็นเพศหญิงจำนวน 18,611 คน คิดเป็นร้อยละ 53.7 ในขณะที่เป็นเพศชายจำนวน 16,060 คน คิดเป็นร้อยละ 46.3 ทั้งนี้ คำสั่งในโปรแกรม R (epiDisplay package) ยังแสดงกราฟแท่งแสดงความถี่มาให้ด้วย



6.2.2 การวิเคราะห์ข้อมูลแบบตารางสองทาง

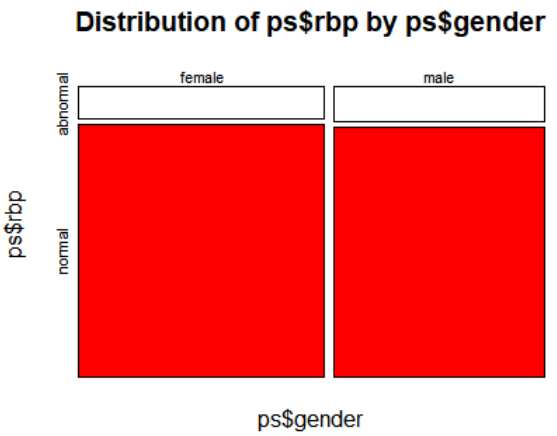
ในกรณีที่เราต้องการคำนวณความถี่และร้อยละ ที่เป็นตัวแปรกลุ่ม 2 ตัวแปร เราจะใช้การวิเคราะห์แบบตารางสองทางโดยใช้คำสั่ง tabpct เช่น ต้องการคำนวณความถี่และร้อยละ ระหว่างเพศและการเป็นโรคความดันโลหิตสูง (ตัวแปร rbp) สามารถทำได้ ดังนี้

```
tabpct(ps$gender, ps$rbp, decimal = 1, percent = "col")

 Column percent
          ps$rbp
 ps$gender  abnormal       %  normal       %
    female      2108  (52.1)   16380  (53.9)
    male        1935  (47.9)   14028  (46.1)
    Total       4043   (100)   30408   (100)
```

จากผลการวิเคราะห์แสดงสัดส่วนของการเป็นโรคและไม่เป็นโรคความดันโลหิตสูงในเพศหญิงและชาย โดยการคำนวณเปอร์เซ็นต์ในแนวคอลัมน์ (percent = "col") และต้องการทศนิยม 1 ตำแหน่ง (decimal =1)

ผลการวิเคราะห์พบว่า ในกลุ่มเพศหญิงมีความถี่ของการเป็นโรคความดันโลหิตสูงเท่ากับ 2,108 ราย คิดเป็นร้อยละ 52.1 ในขณะที่เพศชายจำนวน 1,935 ราย คิดเป็นร้อยละ 47.9 นอกจากนั้นแล้วคำสั่ง tabpct ยังแสดงข้อมูลในรูปแบบกราฟให้มาด้วย



จากกราฟแสดงสัดส่วนของการเป็นโรคความดันโลหิตสูง โดยพื้นที่สีแดงหมายถึง สัดส่วนของการไม่เป็นโรค และพื้นที่สีขาวแสดงถึงการไม่เป็นโรคความดันโลหิตสูง ซึ่งสามารถเข้าใจอย่างหยาบๆ ได้ว่า เพศหญิงและเพศชายพบสัดส่วนการไม่เป็นโรคความดันโลหิตสูงมากกว่าการเป็นโรค

6.3 แบบฝึกหัดท้ายบท

1. ให้สร้างแผนภาพกล่องสำหรับตัวแปรดัชนีมวลกาย (bmi) และเพศ (gender) พร้อมทั้งอธิบายความหมายจากผลลัพธ์
2. ให้วิเคราะห์ตารางสองทาง โดยใช้ตัวแปรกลุ่ม 2 ตัวแปรที่สนใจ พร้อมทั้งอธิบายความหมายจากผลลัพธ์

บทที่ 7

สถิติเชิงอนุมาน (Inferential Statistic)

บทที่ 7

สถิติเชิงอนุมาน (Inferential Statistic)

ผศ.ดร.เชษฐา งามจรัส
E-mail: nchett@kku.ac.th

สถิติเชิงอนุมาน (Inferential statistics) เป็นวิธีการทางสถิติที่ใช้ผลจากการวิเคราะห์ข้อมูลจากกลุ่มตัวอย่างไปสรุปผลหรือไปอธิบายข้อมูลในประชากรที่นักวิจัยต้องการศึกษา สถิติเชิงอนุมานจะประกอบด้วยสถิติที่ใช้ประมาณค่าตัวแปรผลลัพธ์ (outcome) และสถิติที่ใช้ในการทดสอบสมมติฐาน ซึ่งในบทนี้จะอธิบายเฉพาะการทดสอบทางสถิติที่พบบ่อยในการวิเคราะห์ข้อมูลโรคไม่ติดต่อเรื้อรัง

7.1 การทดสอบสมมุติฐานทางสถิติ

การทดสอบสมมุติฐานทางสถิติจะใช้เมื่อนักวิจัยต้องการเปรียบเทียบค่าของตัวแปรผลลัพธ์กับเกณฑ์ หรือเปรียบเทียบตัวแปรผลลัพธ์ระหว่างกลุ่ม เช่น เปรียบเทียบดัชนีมวลกายของประชากรในอำเภอเมือง จังหวัดขอนแก่นว่ามากกว่า 30 กิโลกรัมต่อเมตร^2 หรือไม่ หรือต้องการเปรียบเทียบความชุกของการเป็นโรคอ้วนระหว่างเพศชายและหญิง เป็นต้น

7.1.1 ขั้นตอนการทดสอบสมมุติฐานทางสถิติ

การทดสอบสมมุติฐานทางสถิติมีขั้นตอนดังนี้

1. กำหนดสมมุติฐาน

ขั้นตอนนี้นักวิจัยจะต้องทำการกำหนดสมมุติฐานว่าง (null hypothesis: H_0) และสมมุติฐานทางเลือก (alternative hypothesis: H_a) โดยใช้ค่าพารามิเตอร์เป็นตัวกำหนด เมื่อ H_0 บ่งบอกถึงความไม่แตกต่างกัน ในขณะที่ H_a จะต้องกำหนดให้ตรงกันข้ามกับ H_0 เพื่อเป็นทางเลือกในการตัดสินใจของนักวิจัย เช่น

H_0: ค่าเฉลี่ยของดัชนีมวลกายของประชากร (μ) ในอำเภอเมือง จังหวัดขอนแก่น ไม่แตกต่างจาก 30 กิโลกรัมต่อเมตร^2 (หรือเขียนเป็นสัญลักษณ์ทางคณิตศาสตร์ได้ดังนี้ H_0: μ = 30 กิโลกรัมต่อเมตร^2)

H_a: ค่าเฉลี่ยของดัชนีมวลกายของประชากรในอำเภอเมือง จังหวัดขอนแก่น แตกต่างจาก 30 กิโลกรัมต่อเมตร^2 (หรือเขียนเป็นสัญลักษณ์ทางคณิตศาสตร์ได้ดังนี้ H_a: μ ≠ 30 กิโลกรัมต่อเมตร^2) เป็นต้น

จากการกำหนด H_0 และ H_a ข้างต้นจะเป็นการกำหนดสมมุติฐานแบบสองทาง (two-tailed) คือ นักวิจัยสนใจความแตกต่างระหว่างค่าพารามิเตอร์กับค่าที่กำหนดโดยไม่สนใจทิศทางของความแตกต่าง แต่หากนักวิจัยสนใจทิศทางของความแตกต่างระหว่างค่าดังกล่าวด้วย จะสามารถกำหนดสมมติฐานแบบทางเดียว (one-tailed) ได้ดังนี้

H_0: ค่าเฉลี่ยของดัชนีมวลกายของประชากรในอำเภอเมือง จังหวัดขอนแก่น ไม่เกิน 30 กิโลกรัมต่อเมตร^2 (หรือเขียนเป็นสัญลักษณ์ทางคณิตศาสตร์ได้ดังนี้ H_0: μ ≤ 30 กิโลกรัมต่อเมตร^2)

H_a: ค่าเฉลี่ยของดัชนีมวลกายของประชากรในอำเภอเมือง จังหวัดขอนแก่น มากกว่า 30 กิโลกรัมต่อเมตร^2 (หรือเขียนเป็นสัญลักษณ์ทางคณิตศาสตร์ได้ดังนี้ H_a: μ > 30 กิโลกรัมต่อเมตร^2)

หรืออาจจะกำหนดเป็น

H_0: ค่าเฉลี่ยของดัชนีมวลกายของประชากรในอำเภอเมือง จังหวัดขอนแก่น อย่างน้อย 30 กิโลกรัมต่อเมตร^2 (หรือเขียนเป็นสัญลักษณ์ทางคณิตศาสตร์ได้ดังนี้ H_0: μ ≥ 30 กิโลกรัมต่อเมตร^2)

H^a: ค่าเฉลี่ยของดัชนีมวลกายของประชากรในอำเภอเมือง จังหวัดขอนแก่น น้อยกว่า 30 กิโลกรัมต่อเมตร^2 (หรือเขียนเป็นสัญลักษณ์ทางคณิตศาสตร์ได้ดังนี้ H_a: μ < 30 กิโลกรัมต่อเมตร^2)

การกำหนดว่าจะตั้งสมมุติฐานแบบใดนั้นขึ้นอยู่กับคำถามวิจัยและวัตถุประสงค์ของการวิจัยที่นักวิจัยเป็นผู้กำหนด

2. กำหนดระดับนัยสำคัญ (α)

เป็นการกำหนดความน่าจะเป็นสูงสุดที่นักวิจัยจะตัดสินใจผิดพลาดจากการทดสอบสมมติฐานเมื่อนักวิจัยตัดสินใจปฏิเสธ H_0 เมื่อในสถานการณ์จริง H_0 เป็นจริง ในงานวิจัยทางด้านวิทยาศาสตร์สุขภาพมักนิยมกำหนด α เท่ากับ 0.05

3. คำนวณค่าสถิติทดสอบ

ในขั้นตอนนี้จะเป็นการนำข้อมูลที่ได้เก็บรวบรวมไว้มาทำการคำนวณค่าสถิติจากสถิติทดสอบที่นักวิจัยเลือกใช้

4. ตัดสินใจ

หลังจากที่ได้คำนวณค่าสถิติทดสอบนักวิจัยต้องตัดสินใจว่าจะปฏิเสธ H_0 ที่กำหนดไว้หรือไม่ ในกรณีที่นักวิจัยทำการวิเคราะห์ข้อมูลด้วยโปรแกรม R ผลลัพธ์ที่ได้จากโปรแกรมจะแสดงค่า p-value มาให้ด้วย ซึ่ง p-value จะเป็นค่าที่ได้จากการนำข้อมูลที่นักวิจัยทำการวิเคราะห์มาหาค่าความน่าจะเป็นที่นักวิจัยจะตัดสินใจ

ปฏิเสธ H_0 หรือไม่ ซึ่งในการตัดสินใจปฏิเสธ H_0 จะกระทำเมื่อ p-value น้อยกว่าระดับนัยสำคัญที่กำหนด

5. สรุปผลการทดสอบสมมุติฐาน

ในการสรุปผลการทดสอบสมมุติฐานทางสถิติมีแนวทางในการสรุปดังนี้ ถ้านักวิจัยตัดสินใจปฏิเสธ H_0 จะสรุปว่าพบความแตกต่างอย่างมีนัยสำคัญทางสถิติ ในทางตรงกันข้ามหากนักวิจัยตัดสินใจไม่ปฏิเสธ H_0 หรือยอมรับ H_a จะสรุปว่าพบความแตกต่างอย่างไม่มีนัยสำคัญทางสถิติ

นัยสำคัญทางสถิติแตกต่างจากนัยสำคัญทางชีววิทยาหรือนัยสำคัญทางคลินิก บางครั้งผลการวิจัยที่ได้อาจมีความแตกต่างที่มีนัยสำคัญทางคลินิก แต่อาจไม่มีนัยสำคัญทางสถิติ หากจำนวนกลุ่มตัวอย่างไม่เพียงพอ นอกจากนี้ หากผลการวิจัยพบว่ามีความแตกต่างที่มีนัยสำคัญทางคลินิก โดยเฉพาะผลการวิจัยทางลบ เช่น อาการข้างเคียงจากการทดลองยาใหม่ ให้พิจารณาหยุดการทดลองแม้ว่าจะยังไม่พบความแตกต่างกันทางสถิติ

7.2 การทดสอบสมมุติฐานทางสถิติสำหรับข้อมูลโรคไม่ติดต่อเรื้อรัง

ในหัวข้อนี้เป็นการนำเสนอสถิติทดสอบที่ในการทดสอบสมมุติฐานทางสถิติสำหรับข้อมูลโรคไม่ติดต่อเรื้อรัง ทั้งในกรณีที่ตัวแปรที่ต้องการวิเคราะห์เป็นตัวแปรต่อเนื่อง (continuous variable) และตัวแปรกลุ่ม (categorical variable)

7.2.1 การเปรียบเทียบค่าเฉลี่ยกรณีประชากรกลุ่มเดียว

เมื่อนักวิจัยต้องการเปรียบเทียบค่าเฉลี่ยของตัวแปรต่อเนื่องกับเกณฑ์ ยกตัวอย่างคำถามวิจัยเช่น ค่าเฉลี่ยของดัชนีมวลกายของประชากรในอำเภอเมือง จังหวัดขอนแก่น แตกต่างจาก 30 กิโลกรัมต่อเมตร^2 หรือไม่ เป็นต้น สถิติทดสอบที่ใช้วิเคราะห์เพื่อตอบคำถามวิจัยดังกล่าวคือ สถิติทดสอบ One-sample t ซึ่งเป็นสถิติที่อิงค่าพารามิเตอร์ (parameter statistics) โดยสถิติทดสอบนี้มีเงื่อนไขคือ ข้อมูลที่ต้องการวิเคราะห์ต้องเป็นข้อมูลต่อเนื่องและต้องมีการแจกแจงปกติ ดังนั้น ในการวิเคราะห์นักวิจัยต้องทำการตรวจสอบข้อมูลก่อนว่าข้อมูลเป็นไปตามเงื่อนไขดังกล่าวหรือไม่ สำหรับคำสั่ง R ที่ใช้ในการทดสอบการแจกแจงแบบปกติด้วยสถิติทดสอบ Shapiro-Wilk คือ shapiro.test() ส่วนคำสั่งสำหรับใช้ในการวิเคราะห์สถิติทดสอบ One-sample t คือ t.test()

ตัวอย่างที่ 7.1 นักวิจัยต้องการทราบว่าดัชนีมวลกายของประชากรแตกต่างจาก 30 กิโลกรัมต่อ เมตร^2 หรือไม่ หากข้อมูลดัชนีมวลกายมีการแจกแจงปกติจะสามารถตั้งสมมุติฐานได้ดังนี้

H_0: ค่าเฉลี่ยของดัชนีมวลกายของประชากรไม่แตกต่างจาก 30 กิโลกรัมต่อเมตร^2

H_a: ค่าเฉลี่ยของดัชนีมวลกายของประชากรแตกต่างจาก 30 กิโลกรัมต่อเมตร^2

ทำการวิเคราะห์ด้วยสถิติทดสอบ One-sample t โดยใช้ข้อมูลชื่อ psncd ดังนี้

```
t.test(ps$bmi, mu = 30)

  One Sample t-test

data:  ps$bmi
t = -177.71, df = 34643, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 30
```

```
95 percent confidence interval:
 22.81322 22.97003
sample estimates:
mean of x
 22.89162
```

จากผลลัพธ์ พบว่า p-value < 2.2e-16 ซึ่งน้อยมากๆ และน้อยกว่าระดับนัยสำคัญที่ 0.05 จึงตัดสินใจปฏิเสธ H_0 จึงสามารถสรุปได้ว่าค่าเฉลี่ยของดัชนีมวลกายของประชากรแตกต่างจาก 30 กิโลกรัมต่อเมตร^2 อย่างมีนัยสำคัญทางสถิติ ณ ระดับนัยสำคัญ 0.05 โดยมีค่าเฉลี่ยของดัชนีมวลกายเท่ากับ 22.89 กิโลกรัมต่อเมตร^2 ด้วยช่วงความเชื่อมั่นร้อยละ 95 ตั้งแต่ 22.81 ถึง 22.97 กิโลกรัมต่อเมตร^2 หรือที่ p-value < 0.001

7.2.2 การเปรียบเทียบค่ามัธยฐานกรณีประชากรกลุ่มเดียว

ในกรณีที่ข้อมูลไม่เป็นไปตามเงื่อนไขของสถิติทดสอบ One-sample t คือข้อมูลมีการแจกแจงไม่ปกติ นักวิจัยจะต้องเปลี่ยนมาใช้การทดสอบข้อมูลด้วยสถิติที่ไม่อิงค่าพาราเมทริกซ์ (non-parametric statistics) ที่ชื่อว่า Wilcoxon sign rank แทน ด้วยคำสั่ง wilcox.test() ในโปรแกรม R

ตัวอย่างที่ 7.2 นักวิจัยต้องการทราบว่าดัชนีมวลกายของประชากรแตกต่างจาก 30 กิโลกรัมต่อเมตร^2 หรือไม่ หากข้อมูลดัชนีมวลกายมีการแจกแจงไม่ปกติจะสามารถตั้งสมมุติฐานได้ดังนี้

H_0: ค่ามัธยฐานของดัชนีมวลกายของประชากรไม่แตกต่างจาก 30 กิโลกรัมต่อเมตร^2

H_a: ค่ามัธยฐานของดัชนีมวลกายของประชากรแตกต่างจาก 30 กิโลกรัมต่อเมตร^2

ทำการวิเคราะห์ด้วยสถิติทดสอบ Wilcoxon-sign-rank ดังนี้

```
wilcox.test(ps$bmi, mu = 30)

 Wilcoxon signed rank test with continuity correction

data:  ps$bmi
V = 6306147, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 30
```

จากผลลัพธ์ พบว่า p-value น้อยกว่าระดับนัยสำคัญที่ 0.05 จึงตัดสินใจปฏิเสธ H_0 จึงสามารถสรุปได้ว่าค่ามัธยฐานของดัชนีมวลกายของประชากรแตกต่างจาก 30 กิโลกรัมต่อเมตร^2 อย่างมีนัยสำคัญทางสถิติ

7.2.3 การเปรียบเทียบค่าเฉลี่ยกรณีประชากร 2 กลุ่ม

เมื่อต้องการเปรียบเทียบตัวแปรต่อเนื่องจากประชากร 2 กลุ่ม สามารถทำได้ด้วยการนำค่าเฉลี่ยทั้ง 2 กลุ่มมาเปรียบเทียบกัน เพื่อนำไปสู่การสรุปว่า ค่าเฉลี่ยของประชากร 2 กลุ่มแตกต่างกันหรือไม่ โดยการทดสอบความแตกต่างของค่าเฉลี่ยของประชากร 2 กลุ่ม แบ่งเป็น 2 ประเภท คือ 1) การเปรียบเทียบค่าเฉลี่ยของประชากร 2 กลุ่มที่เป็นอิสระต่อกัน นั่นคือการเก็บข้อมูลที่ไม่ได้มาจากที่เดียวกัน หรือคนเดียวกัน เช่น การเปรียบเทียบค่าเฉลี่ยคะแนนความรู้ของผู้ป่วยเบาหวานที่อาศัยอยู่ในชุมชน A และ B 2) การเปรียบเทียบค่าเฉลี่ยของประชากร 2 กลุ่มที่ไม่เป็นอิสระต่อกัน เช่น การเปรียบเทียบค่าเฉลี่ยคะแนนความรู้ของผู้ป่วยเบาหวานที่อาศัยอยู่ในชุมชน A

ก่อนและหลังดำเนินโครงการอบรม ซึ่งในทางปฏิบัติจะใช้สถิติทดสอบเดียวกันคือ Independence t โดยมีเงื่อนไขคือข้อมูลจากตัวแปรที่สนใจจากทั้งสองกลุ่มต้องมีการแจกแจงปกติ สำหรับคำสั่งในการวิเคราะห์ด้วยสถิติทดสอบ Independence t คือ t.test() ดังตัวอย่างที่ 7.3

ตัวอย่างที่ 7.3 นักวิจัยต้องการทราบว่าดัชนีมวลกายระหว่างเพศชายและเพศหญิงว่าแตกต่างกันหรือไม่ หากข้อมูลดัชนีมวลกายจากทั้งเพศชายและเพศหญิงมีการแจกแจงปกติจะสามารถตั้งสมมุติฐานได้ดังนี้

$H_0: \mu_{male} = \mu_{female}$
$H_a: \mu_{male} \neq \mu_{female}$

เมื่อ μ_{male} คือค่าเฉลี่ยของดัชนีมวลกายของประชากรเพศชาย
μ_{female} คือค่าเฉลี่ยของดัชนีมวลกายของประชากรเพศหญิง

สามารถทำการวิเคราะห์ได้ดังนี้

```
t.test(ps$bmi~ps$gender)

 Welch Two Sample t-test

data:  ps$bmi by ps$gender
t = 2.987, df = 21460, p-value = 0.002821
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.0865173 0.4167898
sample estimates:
mean in group female   mean in group male
            23.00822             22.75656
```

จากผลลัพธ์ พบว่า p-value เท่ากับ 0.0028 ซึ่งน้อยกว่าระดับนัยสำคัญที่ 0.05 จึงตัดสินใจปฏิเสธ H_0 จึงสามารถสรุปได้ว่าค่าเฉลี่ยของดัชนีมวลกายของประชากรเพศชายแตกต่างจากค่าเฉลี่ยของดัชนีมวลกายของประชากรเพศหญิง อย่างมีนัยสำคัญทางสถิติ

7.2.4 การเปรียบเทียบค่ามัธยฐานกรณีประชากร 2 กลุ่ม

หากพบว่าข้อมูลไม่เป็นไปตามเงื่อนไขของสถิติทดสอบ Independence t จึงไม่สามารถใช้สถิติพาราเมทริกซ์ได้ จึงต้องใช้สถิตินอนพาราเมทริกซ์ เนื่องจากสถิติดังกล่าวไม่มีข้อตกลงที่เกี่ยวกับการแจกแจงของประชากรนั้นๆ ดังนั้น นักวิจัยต้องปรับมาใช้สถิติทดสอบ Wilcoxon rank sum แทน ด้วยคำสั่ง wilcox.test() ในโปรแกรม R ดังตัวอย่างที่ 7.4

ตัวอย่างที่ 7.4 นักวิจัยต้องการทราบว่าดัชนีมวลกายระหว่างเพศชายและเพศหญิงแตกต่างกันหรือไม่ หากข้อมูลดัชนีมวลกายจากทั้งเพศชายหรือเพศหญิงไม่มีการแจกแจงปกติจะสามารถตั้งสมมุติฐานได้ดังนี้

$H_0: \text{Median}_{male} = \text{Median}_{female}$
$H_a: \text{Median}_{male} \neq \text{Median}_{female}$

เมื่อ $Median_{male}$ คือ มัธยฐานของดัชนีมวลกายของประชากรเพศชาย

$Median_{female}$ คือ มัธยฐานของดัชนีมวลกายของประชากรเพศหญิง

สามารถทำการวิเคราะห์ได้ดังนี้

```
wilcox.test(ps$bmi~ps$gender)

  Wilcoxon rank sum test with continuity correction

 data:  ps$bmi by ps$gender
 W = 158618186, p-value < 2.2e-16
 alternative hypothesis: true location shift is not equal to 0
```

จากผลลัพธ์ พบว่า p-value น้อยกว่าระดับนัยสำคัญที่ 0.05 จึงตัดสินใจปฏิเสธ H_0 จึงสามารถสรุปได้ว่าค่ามัธยฐานของดัชนีมวลกายของประชากรเพศชายแตกต่างจากค่ามัธยฐานของดัชนีมวลกายของประชากรเพศหญิงอย่างมีนัยสำคัญทางสถิติ

7.2.5 การเปรียบเทียบค่าเฉลี่ยกรณีประชากรมากกว่า 2 กลุ่ม

สำหรับสถิติทดสอบที่ใช้ในการเปรียบเทียบตัวแปรต่อเนื่องในกรณีที่มีประชากรที่ต้องการศึกษามากกว่า 2 กลุ่ม และมีตัวแปรกลุ่ม 1 ตัวแปร คือ สถิติทดสอบ One-way Analysis of Variance (ANOVA) สำหรับเงื่อนไขของ One-way ANOVA คือ ข้อมูลจากทุกกลุ่มต้องมีการแจกแจงปกติและทุกกลุ่มต้องมีความแปรปรวนไม่แตกต่างกัน สำหรับคำสั่งในการวิเคราะห์ One-way ANOVA ในโปรแกรม R คือ คำสั่ง aov() และ summary() ดังตัวอย่างที่ 7.5

ตัวอย่างที่ 7.5 นักวิจัยต้องการทราบว่าประชากรที่มีช่วงอายุต่างกันจะมีดัชนีมวลกายแตกต่างกันหรือไม่ หากข้อมูลเป็นไปตามเงื่อนไขของ One-way ANOVA จะสามารถตั้งสมมุติฐานได้ดังนี้

H_0: ค่าเฉลี่ยดัชนีมวลกายจากประชากรทุกกลุ่มอายุไม่แตกต่างกัน หรือ $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a: มีค่าเฉลี่ยดัชนีมวลกายจากประชากรอย่างน้อย 1 กลุ่มอายุ ที่แตกต่างจากกลุ่มอายุอื่น

สามารถทำการวิเคราะห์ได้ดังนี้

```
tab1(ps$age.gr)
 ps$age.gr :
          Frequency Percent Cum. percent
 (35,40]       1104     3.2         3.2
 (40,50]       8177    23.6        26.8
 (50,60]      10920    31.5        58.3
 (60,110]     14470    41.7       100.0
   Total      34671   100.0       100.0
summary(aov(ps$bmi~ps$age.gr))
                Df  Sum Sq Mean Sq F value Pr(>F)
 ps$age.gr       3   12535    4178   75.86 <2e-16 ***
 Residuals   34640 1907785      55
 ---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
27 observations deleted due to missingness
```

จากผลลัพธ์ ของคำสั่ง tab1 พบว่าตัวแปรกลุ่มอายุมีทั้งหมด 4 กลุ่ม และเมื่อทำการวิเคราะห์ One-way ANOVA พบว่า p-value น้อยมากๆ (p-value < 0.001) ซึ่งน้อยกว่าระดับนัยสำคัญที่ 0.05 จึงตัดสินใจปฏิเสธ H_0 และสามารถสรุปได้ว่ามีค่าเฉลี่ยดัชนีมวลกายจากประชากรอย่างน้อย 1 กลุ่มอายุ ที่แตกต่างจากกลุ่มอายุอื่นอย่างมีนัยสำคัญทางสถิติ

เมื่อผลการวิเคราะห์ด้วย One-way ANOVA ปฏิเสธ H_0 จะต้องทำการวิเคราะห์เพื่อตรวจสอบว่ามีกลุ่มใดแตกต่างกันบ้างด้วยการเปรียบเทียบพหุ (Multiple comparison) ซึ่งสามารถทำได้หลายวิธี แต่ในบทนี้จะใช้การเปรียบเทียบเพื่อหาคู่ต่างด้วยวิธี Tukey's Honest Significant Difference (Tukey HSD) ด้วยคำสั่ง TukeyHSD() ดังนี้

```
TukeyHSD(aov(ps$bmi~ps$age.gr))
  Tukey multiple comparisons of means
    95% family-wise confidence level

 Fit: aov(formula = ps$bmi ~ ps$age.gr)

 $`ps$age.gr`
                        diff        lwr        upr     p adj
 (40,50]-(35,40]   0.06632329 -0.5452614  0.6779079 0.9924553
 (50,60]-(35,40]   0.15203822 -0.4503447  0.7544212 0.9161426
 (60,110]-(35,40] -1.10838483 -1.7039343 -0.5128353 0.0000104
 (50,60]-(40,50]   0.08571493 -0.1932307  0.3646605 0.8593043
 (60,110]-(40,50] -1.17470813 -1.4385729 -0.9108433 0.0000000
 (60,110]-(50,60] -1.26042305 -1.5021958 -1.0186503 0.0000000
```

จากผลลัพธ์ ข้างต้นจะพบว่า มีจำนวน 3 คู่เปรียบเทียบที่ให้ p adj น้อยกว่าระดับนัยสำคัญ 0.05 แสดงว่าคู่เปรียบเทียบเหล่านั้นมีความแตกต่างกันทางสถิติ ในการสรุปผลการวิเคราะห์จะขอยกตัวอย่างเพียง 1 คู่เปรียบเทียบที่แตกต่างกันดังนี้ กลุ่มคนที่มีอายุมากกว่า 60 ปี ถึง 110 ปี มีค่าเฉลี่ยดัชนีมวลกายแตกต่างจากกลุ่มคนที่มีอายุมากกว่า 35 ปี ถึง 40 ปี อย่างมีนัยสำคัญทางสถิติ (p-value < 0.0001) โดยต่างกันอยู่ที่ 1.1 กิโลกรัมต่อเมตร^2

7.2.6 การเปรียบเทียบค่ามัธยฐานกรณีประชากรมากกว่า 2 กลุ่ม

ในกรณีที่ข้อมูลไม่สอดคล้องกับเงื่อนไขของสถิติทดสอบ One-way ANOVA นั้น นักวิจัยสามารถเปลี่ยนมาใช้สถิตินอนพารามิเตอร์คือ สถิติทดสอบ Kruskal-Wallis ได้โดยใช้คำสั่ง kruskal.test() ดังตัวอย่างที่ 7.6

ตัวอย่างที่ 7.6 นักวิจัยต้องการทราบว่าประชากรที่มีช่วงอายุต่างกันจะมีดัชนีมวลกายแตกต่างกันหรือไม่

จากการทดสอบด้วยสถิติทดสอบ Bartlett ดังผลลัพธ์ข้างล่างพบว่าข้อมูลดัชนีมวลกายของกลุ่มอายุต่างๆ มีความแปรปรวนแตกต่างกัน (p-value < 0.001) ดังปรากฏในผลลัพธ์ของคำสั่ง bartlett.test() ซึ่งไม่เป็นไปตามเงื่อนไขของการใช้ One-way ANOVA ดังนั้นนักวิจัยจะต้องทำวิเคราะห์ด้วยสถิติทดสอบ Kruskal-Wallis ซึ่งสามารถตั้งสมมุติฐานได้ดังนี้

H_0: ค่ามัธยฐานของดัชนีมวลกายจากประชากรทุกกลุ่มอายุไม่แตกต่างกัน

H_a: มีค่ามัธยฐานของดัชนีมวลกายจากประชากรอย่างน้อย 1 กลุ่มอายุ ที่แตกต่างจากกลุ่มอายุอื่น

เมื่อทำการวิเคราะห์ด้วยโปรแกรม R ได้ผลลัพธ์ดังนี้

```
bartlett.test(ps$bmi~ps$age.gr)

  Bartlett test of homogeneity of variances

 data:  ps$bmi by ps$age.gr
 Bartlett's K-squared = 10017, df = 3, p-value < 2.2e-16
kruskal.test(ps$bmi~ps$age.gr)

  Kruskal-Wallis rank sum test

 data:  ps$bmi by ps$age.gr
 Kruskal-Wallis chi-squared = 781.94, df = 3, p-value < 2.2e-16
```

จากผลลัพธ์ การทดสอบด้วยสถิติทดสอบ Kruskal-Wallis พบว่า p-value < 0.001 น้อยกว่าระดับนัยสำคัญที่ 0.05 จึงตัดสินใจปฏิเสธ H_0 จึงสามารถสรุปได้ว่ามีค่ามัธยฐานของดัชนีมวลกายจากประชากรอย่างน้อย 1 กลุ่มอายุ ที่แตกต่างจากกลุ่มอายุอื่น อย่างมีนัยสำคัญทางสถิติ

7.2.7 การทดสอบความสัมพันธ์ระหว่างตัวแปรต่อเนื่อง 2 ตัวแปร

หัวข้อนี้จะแนะนำสถิติที่ใช้บอกขนาดความสัมพันธ์เชิงเส้นตรงระหว่างตัวแปรต่อเนื่อง 2 ตัวแปรคือ ค่าสัมประสิทธิ์สหสัมพันธ์ (correlation) และสถิติทดสอบชื่อ Pearson's correlation และ Spearman's rank correlation ในการทดสอบความสัมพันธ์เชิงเส้นตรงระหว่างตัวแปรต่อเนื่อง 2 ตัวแปร เมื่อกำหนดให้ r และ r แทนค่าสัมประสิทธิ์สหสัมพันธ์ของประชากรและตัวอย่างตามลำดับ โดย r และ r มีค่าตั้งแต่ -1 ถึง 1 หาก r และ r มีค่าเท่ากับ 0 แสดงว่าตัวแปรทั้งสองไม่มีความสัมพันธ์เชิงเส้นตรงต่อกัน และหากพบว่าค่าสัมประสิทธิ์สหสัมพันธ์มีค่าบวกแสดงว่าทั้งสองตัวแปรมีความสัมพันธ์กันในทิศทางบวกหรือทิศทางเดียวกัน แต่ถ้าสัมประสิทธิ์สหสัมพันธ์มีค่าลบแสดงว่าทั้งสองตัวแปรมีความสัมพันธ์กันในทิศทางลบหรือทิศทางตรงกันข้ามต่อกัน

สำหรับสถิติทดสอบ Pearson's correlation จะใช้ได้เมื่อข้อมูลจากทั้งสองตัวแปรมีการแจกแจงปกติร่วมกัน ซึ่งเบื้องต้นอาจตรวจสอบว่าข้อมูลทั้งสองตัวแปรมีการแจกแจงปกติหรือไม่ หากไม่ก็แสดงว่าข้อมูลไม่เป็นไปตามเงื่อนไขดังกล่าว นักวิจัยจะต้องทำการวิเคราะห์ด้วยสถิติทดสอบ Spearman's rank correlation แทน สำหรับคำสั่งที่ใช้ในการวิเคราะห์ Pearson's correlation และ Spearman's rank correlation คือ cor.test() เพียงแต่ในกรณีที่เป็นการวิเคราะห์ด้วย Spearman's rank correlation จะต้องระบุ method = "spearman" หลังชื่อตัวแปรที่ทำการวิเคราะห์ดังตัวอย่างที่ 7.7 และ 7.8 ตามลำดับ

ตัวอย่างที่ 7.7 นักวิจัยต้องการทราบว่าอายุกับดัชนีมวลกายของประชากรที่ทำการศึกษามีความสัมพันธ์กันหรือไม่

หากพบว่าข้อมูลเป็นไปตามเงื่อนไขของสถิติทดสอบ Pearson's correlation จะสามารถตั้งสมมุติฐานได้ดังนี้

H_0: อายุกับดัชนีมวลกายของประชากรที่ทำการศึกษาไม่มีความสัมพันธ์เชิงเส้นตรงต่อกัน (หรือ $H_0: \rho = 0$)

H_a: อายุกับดัชนีมวลกายของประชากรที่ทำการศึกษามีความสัมพันธ์เชิงเส้นตรงต่อกัน (หรือ $H_a: \rho \neq 0$)

เมื่อทำการวิเคราะห์ด้วยโปรแกรม R ได้ผลลัพธ์ดังนี้

```
cor.test(ps$bmi, ps$age)

  Pearson's product-moment correlation

 data:  ps$bmi and ps$age
 t = -19.263, df = 34642, p-value < 2.2e-16
 alternative hypothesis: true correlation is not equal to 0
 95 percent confidence interval:
  -0.1133511 -0.0925139
 sample estimates:
        cor
 -0.1029438
```

จากผลลัพธ์ พบว่า p-value < 0.001 น้อยกว่าระดับนัยสำคัญที่ 0.05 จึงตัดสินใจปฏิเสธ H_0 จึงสามารถสรุปได้ว่าอายุกับดัชนีมวลกายของประชากรที่ทำการศึกษามีความสัมพันธ์เชิงเส้นตรงต่อกันในทิศทางตรงกันข้ามอย่างมีนัยสำคัญทางสถิติ ด้วยขนาดความสัมพันธ์เท่ากับ -0.10 ด้วยช่วงความเชื่อมั่นร้อยละ 95 ตั้งแต่ -0.11 ถึง -0.09

ตัวอย่างที่ 7.8 นักวิจัยต้องการทราบว่าอายุกับดัชนีมวลกายของประชากรที่ทำการศึกษามีความสัมพันธ์กันหรือไม่

หากพบว่าข้อมูลไม่เป็นไปตามเงื่อนไขของสถิติทดสอบ Pearson's correlation จะสามารถตั้งสมมุติฐานได้ดังนี้

H_0: อายุกับดัชนีมวลกายของประชากรที่ทำการศึกษาไม่มีความสัมพันธ์เชิงเส้นตรงต่อกัน (หรือ $H_0: \rho = 0$)

H_a: อายุกับดัชนีมวลกายของประชากรที่ทำการศึกษามีความสัมพันธ์เชิงเส้นตรงต่อกัน (หรือ $H_a: \rho \neq 0$)

เมื่อทำการวิเคราะห์ด้วยโปรแกรม R ได้ผลลัพธ์ดังนี้

```
cor.test(ps$bmi, ps$age, method = "spearman")
 Warning in cor.test.default(ps$bmi, ps$age, method = "spearman"): Cannot compute
 exact p-value with ties

  Spearman's rank correlation rho
```

```
data:  ps$bmi and ps$age
S = 8.0845e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1666023
```

จากผลลัพธ์ พบว่า p-value น้อยกว่าระดับนัยสำคัญที่ 0.05 จึงตัดสินใจปฏิเสธ H_0 จึงสามารถสรุปได้ว่า อายุกับดัชนีมวลกายของประชากรที่ทำการศึกษามีความสัมพันธ์เชิงลบต่อกัน อย่างมีนัยสำคัญทางสถิติ ด้วยขนาดความสัมพันธ์เท่ากับ -0.16

7.2.8 การเปรียบเทียบค่าสัดส่วนกรณีประชากรกลุ่มเดียว

เมื่อต้องการเปรียบค่าสัดส่วนของตัวแปรแจงนับจากประชากรกลุ่มเดียวกับเกณฑ์สามารถทำได้โดยใช้สถิติทดสอบ Z ด้วยคำสั่ง prop.test() ในโปรแกรม R ดังตัวอย่างที่ 7.9

ตัวอย่างที่ 7.9 นักวิจัยต้องการทราบว่าในประชากรที่ทำการศึกษามีคนเป็นโรคเบาหวานมากกว่าร้อยละ 10 หรือไม่

สามารถตั้งสมมุติฐานได้ดังนี้

H_0: สัดส่วนการเป็นโรคเบาหวานในประชากรไม่เกิน 0.1

H_a: สัดส่วนการเป็นโรคเบาหวานในประชากรมากกว่า 0.1

เมื่อทำการวิเคราะห์ด้วยโปรแกรม R ได้ผลลัพธ์ดังนี้

```
table(ps$dm)

     DM normal
   4051  28050
length(ps$dm)
 [1] 34671
prop.test(x=4051, n= 34671,p = 0.1, alternative = "greater")

  1-sample proportions test with continuity correction

 data:  4051 out of 34671, null probability 0.1
 X-squared = 109.07, df = 1, p-value < 2.2e-16
 alternative hypothesis: true p is greater than 0.1
 95 percent confidence interval:
  0.1140191 1.0000000
 sample estimates:
         p
 0.1168412
```

จากผลลัพธ์ของคำสั่ง table() จะเห็นว่าตัวแปร dm มีจำนวน 3 กลุ่มคือ DM NA และ normal โดยมีคนเป็นเบาหวาน 4,051 คน จากทั้งหมด 34,671 คน เมื่อทำการวิเคราะห์ด้วยคำสั่ง prop.test() พบว่าสัดส่วนการเป็นโรคเบาหวานในตัวอย่างเท่ากับ 0.1168 หรือร้อยละ 11.68 เมื่อทำการทดสอบในประชากรพบว่าสัดส่วนการเป็นโรคเบาหวานในประชากรมากกว่าร้อยละ 10 (หรือสัดส่วน 0.1) อย่างมีนัยสำคัญทางสถิติ (p-value < 0.001)

7.2.9 การเปรียบเทียบค่าสัดส่วนกรณีประชากร 2 กลุ่ม

เมื่อต้องการเปรียบเทียบค่าสัดส่วนของตัวแปรแจงนับจากประชากร 2 กลุ่ม สามารถทำได้โดยใช้สถิติทดสอบ Z ด้วยคำสั่ง prop.test() ในโปรแกรม R ดังตัวอย่างที่ 7.10

ตัวอย่างที่ 7.10 นักวิจัยต้องการทราบว่าในประชากรที่ทำการศึกษา เพศชายและเพศหญิงมีคนเป็นโรคเบาหวานแตกต่างกันหรือไม่

สามารถตั้งสมมุติฐานได้ดังนี้

H_0: สัดส่วนการเป็นโรคเบาหวานในประชากรเพศชายและเพศหญิงไม่แตกต่างกัน

H_a: สัดส่วนการเป็นโรคเบาหวานในประชากรเพศชายและเพศหญิงแตกต่างกัน

เมื่อทำการวิเคราะห์ด้วยโปรแกรม R ได้ผลลัพธ์ดังนี้

```
dat <- ps[ps$dm != "NA", c("gender","dm")]
tabpct(dat$gender, dat$dm, percent = "row")

 Row percent
           dat$dm
 dat$gender      DM  normal  Total
     female    2296   14999  17295
            (13.3)  (86.7)  (100)
     male      1755   13051  14806
            (11.9)  (88.1)  (100)
prop.test(x=c(2296, 1755), n= c(17295, 14806))

  2-sample test for equality of proportions with continuity correction

 data:  c(2296, 1755) out of c(17295, 14806)
 X-squared = 14.504, df = 1, p-value = 0.0001399
 alternative hypothesis: two.sided
 95 percent confidence interval:
  0.006901282 0.021542927
 sample estimates:
    prop 1    prop 2
 0.1327551 0.1185330
```

จากผลลัพธ์ เป็นสร้างกรอบข้อมูล (data frame) ใหม่ชื่อ dat ซึ่งมีแค่ 2 ตัวแปรคือ ตัวแปร gender และ dm และไม่เอากลุ่ม NA ในตัวแปร dm และทำการวิเคราะห์ด้วยคำสั่ง tabpct() พบว่าเพศหญิงเป็นโรคเบาหวานจำนวน 2,296 คน จาก 17,295 คน และเพศชายเป็นโรคนี้จำนวน 1,755 คน จาก 14,806 คน เมื่อทำการวิเคราะห์

เพื่อตอบโจทย์ด้วยคำสั่ง prop.test() พบว่า สัดส่วนการเป็นโรคเบาหวานในประชากรเพศชายและเพศหญิงแตกต่างกัน อย่างมีนัยสำคัญทางสถิติ (p-value = 0.001)

7.2.10 การทดสอบความสัมพันธ์ระหว่างตัวแปรกลุ่ม 2 ตัวแปร

7.2.10.1 การทดสอบไคว์สแควร์ (Chi-squared test)

การทดสอบความสัมพันธ์ระหว่างตัวแปรกลุ่ม 2 ตัวแปร สามารถทำได้โดยใช้สถิติทดสอบ Chi-squared หรือ Fisher's exact โดยสถิติทดสอบ Chi-squared โดยจะใช้เมื่อมีตัวอย่างขนาดใหญ่เพียงพอ หรือพิจารณาจากความถี่ค่าคาดหมาย (expected value) ไม่ควรต่ำกว่า 5 หรือความถี่คาดหวังที่มีค่าน้อยกว่า 5 ต้องไม่เกินร้อยละ 20 ของจำนวนเซลล์ทั้งหมดในตารางสองทางที่สร้างจากตัวแปรทั้งสอง นั่นคือข้อตกลงที่จะใช้สถิติทดสอบ Chi-squared ได้ หากค่าความถี่คาดหมาย (expected value) น้อยกว่า 5 น้อยกว่าร้อยละ 20 ของจำนวนเซลล์ทั้งหมดในตารางดังกล่าว จะต้องทำการวิเคราะห์ด้วยสถิติทดสอบ Fisher's exact สำหรับคำสั่งที่ใช้ในการหาค่าความถี่คาดหมาย คือ chisq.test()$expected ส่วนคำสั่ง chisq.test() ใช้ในการวิเคราะห์ Chi-squared ดังตัวอย่างที่ 7.11 และ คำสั่ง fisher.test() จะใช้ในการวิเคราะห์ Fisher's exact

ตัวอย่างที่ 7.11 นักวิจัยต้องการทราบว่าเพศกับการเป็นโรคเบาหวานในประชากรที่ทำการศึกษามีความสัมพันธ์กันหรือไม่

สามารถตั้งสมมุติฐานได้ดังนี้

H_0 : เพศกับการเป็นโรคเบาหวานในประชากรที่ทำการศึกษาไม่มีความสัมพันธ์กัน

H_a : เพศกับการเป็นโรคเบาหวานในประชากรที่ทำการศึกษามีความสัมพันธ์กัน

เมื่อทำการวิเคราะห์ด้วยโปรแกรม R ได้ผลลัพธ์ดังนี้

```
chisq.test(table(dat))$expected
         dm
 gender        DM   normal
   female 2182.55 15112.45
   male   1868.45 12937.55
chisq.test(table(dat))

  Pearson's Chi-squared test with Yates' continuity correction

 data:  table(dat)
 X-squared = 14.504, df = 1, p-value = 0.0001399
```

จากผลลัพธ์ เมื่อทำการหาค่าความถี่คาดหมายพบว่าค่าความถี่คาดหมายมากกว่า 5 ทั้ง 4 เซลล์ จึงทำการวิเคราะห์ด้วยสถิติทดสอบ Chi-squared ได้ p-value = 0.001 จึงตัดสินใจปฏิเสธ H_0 สรุปได้ว่าเพศกับการเป็นโรคเบาหวานในประชากรที่ทำการศึกษามีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติ ณ ระดับนัยสำคัญ 0.05

7.2.10.2 การทดสอบอัตราส่วนเสี่ยง (Risk ratio)

อัตราส่วนเสี่ยง (Risk ratio) หรือความเสี่ยงสัมพัทธ์ (Relative risk) เรียกย่อว่า RR เป็นการหาอัตราส่วนระหว่างอุบัติการณ์ของการเกิดเหตุการณ์ที่สนใจ ในกลุ่มที่ได้รับปัจจัยเสี่ยงต่อการเกิดอุบัติการณ์ของกลุ่มที่ไม่ได้

ตารางที่ 7.1 ความถี่ระหว่างปัจจัยเสี่ยงและผล

	Disease +	Disease -
Exposure +	a	b
Exposure -	c	d

รับปัจจัยเสี่ยง ซึ่งเป็นตัวบ่งบอกว่ากลุ่มที่ได้รับปัจจัยเสี่ยงมีโอกาสเกิดเหตุการณ์ที่สนใจเป็นกี่เท่าของกลุ่มที่ไม่ได้รับปัจจัยเสี่ยง โดยความถี่ระหว่างปัจจัยเสี่ยงและผล ดังแสดงในตารางที่ 7.1

ตัวอย่างที่ 7.12 นักวิจัยต้องการทราบว่าผู้มีดัชนีมวลกายเกินเกณฑ์มาตรฐาน (bmigr: obesity) จะมีความเสี่ยงต่อการเป็นโรคเบาหวานหรือไม่

```
#RR : Risk ratio

table(ps$dm)
     DM normal
   4051  28050
table(ps$bmigr)
  normal obesity
   33366    1278
ps$bmigr <- relevel(ps$bmigr, ref = "obesity")

dat <- ps[ps$dm != "NA", c("bmigr","dm")]
tabpct(dat$bmigr, dat$dm, percent = "row")

 Row percent
          dat$dm
 dat$bmigr      DM  normal  Total
   obesity     257     968   1225
              (21)    (79)  (100)
   normal     3792   27070  30862
            (12.3)  (87.7)  (100)
csi(257,968,3792,27070)
           Exposure
 Outcome    Non-exposed Exposed Total
   Negative 27070        968     28038
   Positive 3792         257     4049
   Total    30862        1225    32087

            Rne          Re      Rt
   Risk     0.12         0.21    0.13
```

```
                                    Estimate Lower95ci Upper95ci
  Risk difference (attributable risk)    0.09     0.07      0.1
  Risk ratio                             1.71     1.52      1.92
  Attr. frac. exp. -- (Re-Rne)/Re        0.41
  Attr. frac. pop. -- (Rt-Rne)/Rt*100 %  2.63
  Number needed to harm (NNH)            11.5     9.55      14.96
    or 1/(risk difference)
```

จากการวิเคราะห์ข้อมูล ผู้วิจัยได้กำหนดให้กลุ่มที่สนใจศึกษาเป็นกลุ่มที่มีดัชนีมวลกายเกิน เป็นกลุ่มอ้างอิง โดยการใช้คำสั่ง relavel() จากนั้นหาความถี่เพื่อคำนวณค่า RR โดยใช้คำสั่ง csi() ซึ่งใช้ในกรณีที่เป็นค่าความถี่ 2 x 2 เท่านั้น แต่หากข้อมูลที่ระบุเป็นรายคน สามารถคำนวณโดยใช้คำสั่ง cs() ซึ่งอยู่ในไลบรารี่ epiDisplay เช่นกัน ผลการคำนวณมีค่า RR เท่ากับ 1.71 และมีช่วงความเชื่อมั่นเท่ากับ 1.52 – 1.92 โดยช่วงความเชื่อมั่นของค่า RR ไม่คร่อม 1 จึงสรุปได้ว่า กลุ่มที่มีดัชนีมวลกายอยู่ในภาวะอ้วนมีความเสี่ยงต่อการเป็นโรคเบาหวานสูงกว่ากลุ่มปกติเป็น 1.71 เท่า

7.2.10.2 การหาอัตราส่วนออดส์ (Odds ratio)

อัตราส่วนออดส์ (Odds ratio) เรียกย่อว่า OR คำนวณจากอัตราส่วนความน่าจะเป็น 2 ค่า คือ ความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจต่อความน่าจะเป็นของการเกิดเหตุการณ์ที่ไม่สนใจ

ตัวอย่างที่ 17.13 ผู้วิจัยต้องการหาความสัมพันธ์ระหว่างการสูบบุหรี่และการป่วยเป็นโรคความดันโลหิตสูง จึงทำการทดสอบความสัมพันธ์ดังนี้

```
#OR
ps$smoke <- factor(ps$smoke,
                   levels = c("no","rarely","regular","sometime"),
                   labels = c("no","yes", "yes","yes"))
ps$smoke <- relevel(ps$smoke, ref = "yes")

ps$rbp <- relevel(ps$rbp, ref = "abnormal")

dat <- ps[ps$rbp != "NA", c("smoke","rbp")]
tabpct(dat$smoke, dat$rbp, percent = "row")

 Row percent
          dat$rbp
 dat$smoke  abnormal  normal  Total
      yes        85     275    360
             (23.6)  (76.4)  (100)
      no       354    1474   1828
             (19.4)  (80.6)  (100)
cc(dat$smoke, dat$rbp)

          dat$rbp
```

```
dat$smoke abnormal normal Total
    yes         85    275   360
    no         354   1474  1828
    Total      439   1749  2188

OR =  1.29
95% CI =  0.98, 1.69
Chi-squared = 3.38, 1 d.f., P value = 0.066
Fisher's exact test (2-sided) P value = 0.072
```

จากผลการคำนวณค่า OR โดยใช้คำสั่ง cc() ในกรณีที่ข้อมูลเป็นความถี่แล้วนั้น พบว่า ค่า OR มีค่าเท่ากับ 1.29 และค่าช่วงความเชื่อมั่นเท่ากับ 0.98 – 1.69 ซึ่งสามารถสรุปได้ว่า คนที่สูบบุหรี่มีโอกาสเป็นโรคความดันโลหิตสูงเป็น 1.29 เท่าเมื่อเทียบกับคนที่ไม่สูบบุหรี่ แต่อย่างไรก็ตาม พบว่าความดันโลหิตสูงและการสูบบุหรี่ไม่มีนัยสำคัญทางสถิติที่ 0.05

7.3 แบบฝึกหัดท้ายบท

1. ให้ทำการทดสอบสมมติฐานว่าดัชนีมวลกาย (bmi) ในผู้ป่วยที่เป็นโรคเบาหวานและไม่มีโรคเบาหวาน (dm) แตกต่างกันหรือไม่ โดยให้ทำการเขียนวิธีการและสมมติฐานที่ใช้ พร้อมทั้งอธิบายวิธีการวิเคราะห์อย่างละเอียด
2. ให้ทำการทดสอบสมมติฐานว่าน้ำหนักตัว (weight) ในทุกสถานภาพสมรส (mstatus) แตกต่างกันหรือไม่ โดยให้ทำการเขียนวิธีการและสมมติฐานที่ใช้ พร้อมทั้งอธิบายวิธีการวิเคราะห์อย่างละเอียด

7.4 บรรณานุกรม

1. เชษฐา งามจรัส. ชีวสถิติและการวิเคราะห์ข้อมูลด้วย R. ครั้งที่ 2. ขอนแก่น: โรงพิมพ์ มหาวิทยาลัยขอนแก่น; 2553.
2. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2019.

บทที่ 8

การวิเคราะห์การถดถอยเชิงเส้น
(Linear regression analysis)

บทที่ 8

การวิเคราะห์การถดถอยเชิงเส้น (Linear regression analysis)

รศ.ดร.อภิรดี แซ่ลิ่ม
E-mail: api_45@hotmail.com

การถดถอยเชิงเส้น (Linear regression) เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรตามที่เป็นตัวแปรแบบต่อเนื่อง กับตัวแปรอิสระ (หรือตัวแปรต้น) หนึ่งตัวหรือมากกว่าหนึ่งตัวขึ้นไป โดยตัวแปรอิสระสามารถเป็นได้ทั้งตัวแปรแบบต่อเนื่องหรือกลุ่ม หรืออีกนัยหนึ่งการถดถอยเชิงเส้น เป็นการทำนายค่าของตัวแปรตามจากค่าของตัวแปรอิสระ หากตัวแบบการถดถอยเชิงเส้นที่สร้างขึ้นมีตัวแปรอิสระเพียงตัวเดียว เรียกตัวแบบการถดถอยนั้นว่า การถดถอยเชิงเส้นอย่างง่าย (Simple linear regression) และหากตัวแบบการถดถอยเชิงเส้นที่สร้างขึ้นมีตัวแปรอิสระตั้งแต่สองตัวแปรขึ้นไป เรียกตัวแบบการถดถอยนั้นว่า การถดถอยเชิงเส้นพหุคูณ (Multiple linear regression)

8.1 การสร้างแผนภาพแสดงความสัมพันธ์

ก่อนการสร้างตัวแบบการถดถอยเชิงเส้น ควรมีการสำรวจความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระทีละตัวในเบื้องต้นก่อน

8.1.1 แผนภาพการกระจาย (Scatter plot)

หากตัวแปรอิสระเป็นตัวแปรแบบต่อเนื่อง การสำรวจความสัมพันธ์ได้จากการสร้างแผนภาพการกระจาย (Scatter plot) ซึ่งเป็นวิธีการนำเอาค่าของข้อมูลในตัวแปรตาม และตัวแปรอิสระ โดยทั้งสองตัวแปรนี้เป็นตัวแปรแบบต่อเนื่อง มาลงตำแหน่งบนกราฟที่นำเสนอเป็นจุด เพื่อแสดงให้เห็นถึงความสัมพันธ์ระหว่างสองตัวแปรนี้ โดยค่าจากตัวแปรอิสระกำหนดให้อยู่บนแกน X (ในแนวนอน) ส่วนค่าจากตัวแปรตามกำหนดให้อยู่บนแกน Y (ในแนวตั้ง) ความสัมพันธ์อาจเป็นเชิงบวก หรือเชิงลบ หรือไม่มีความสัมพันธ์กันก็ได้ การสร้างแผนภาพการกระจายเป็นวิธีการที่ทำให้เห็นภาพความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามในเบื้องต้น ก่อนการดำเนินการวิเคราะห์การถดถอยเชิงเส้น

8.1.2 แผนภาพกล่อง (Boxplot)

หากตัวแปรอิสระเป็นตัวแปรแบบกลุ่ม การสำรวจความสัมพันธ์ได้จากการสร้างแผนภาพกล่อง โดยจำนวนของกล่องเท่ากับจำนวนกลุ่มที่มีในตัวแปรอิสระนั้นๆ แผนภาพกล่องสามารถนำเสนอได้ทั้งในแนวนอนและแนวตั้ง

ข้อมูลที่ใช้เป็นตัวอย่างในบทนี้คือ ข้อมูล blevel.csv จำนวน 34,671 แถว จำนวนตัวแปร 9 ตัวแปร โดยดึงมาเฉพาะตัวแปรที่จะใช้งานจากข้อมูล psncd ดังรายละเอียดคำอธิบายชุดข้อมูลในตารางที่ 1

ตารางที่ 8.1 คำอธิบายตัวแปร และค่าตัวแปรของข้อมูลผู้เข้ารับบริการตรวจคัดกรองสุขภาพเกี่ยวกับโรคเรื้อรังไม่ติดต่อ

ตัวแปร	คำอธิบายตัวแปร	คำอธิบายค่าตัวแปร
id	รหัส	
gender	เพศ	1 = ชาย 2 = หญิง
age	อายุ (ปี)	
mstatus	สถานภาพสมรส	1 = โสด 2 = แต่งงาน 3 = หม้าย/หย่าร้าง/อื่น ๆ
edu	ระดับการศึกษา	1 = ไม่ได้เรียน 2 = ประถมศึกษาหรือต่ำกว่า 3 = มัธยมศึกษา 4 = ปวช/ปวส 5 = ปริญญาตรีหรือสูงกว่า
bmi	ดัชนีมวลกาย	
sbp	ความดันโลหิตซิสโตลิก (mmHg)	
dbp	ความดันโลหิตไอแอสโตลิก (mmHg)	
bslevel	ระดับน้ำตาลในเลือด (mg/dL)	

แสดงการวิเคราะห์ข้อมูลดังนี้

เปิดไลบรารี่ epiDisplay

```
library(epiDisplay)
```

ตั้งค่า working directory

```
setwd("D:\\ncddata")
```

ตั้งค่าให้ R อ่านภาษาไทยได้

```
Sys.setlocale(locale="Thai")
 [1] "LC_COLLATE=Thai_Thailand.874;LC_CTYPE=Thai_Thailand.874;LC_MONETARY=Thai_
Thailand.874;LC_NUMERIC=C;LC_TIME=Thai_Thailand.874"
```

เปิดอ่านไฟล์ bslevel.csv

```
bs <- read.csv("blevel.csv")
des(bs)

  No. of observations =  34671
    Variable      Class           Description
 1  id            integer
 2  gender        factor
 3  age           numeric
 4  mstatus       factor
 5  edu           factor
 6  bmi           numeric
 7  sbp           numeric
 8  dbp           numeric
 9  bslevel       integer
```

8.2 วิเคราะห์สถิติเชิงพรรณนาเบื้องต้น

8.2.1 คำนวณค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน สำหรับตัวแปรต่อเนื่อง จำนวนและร้อยละสำหรับตัวแปรกลุ่ม

```
codebook(bs[,c(2:9)])
 gender    :
        Frequency Percent
 female     18611    53.7
 male       16060    46.3

  ==================
 age    :
  obs.  mean    median  s.d.   min.   max.
  34671 58.674 57.492  12.453 38.031 103.001

  ==================
 mstatus    :
                 Frequency Percent
```

```
Married             22470   66.30
Single               9221   27.21
Widowes/divoced      2200    6.49

 ==================
edu   :
                Frequency Percent
Bachelor/higher       690    3.48
Colledge              945    4.76
Illiterate           1320    6.65
Mathayom             3901   19.65
Pathom              12996   65.46

 ==================
bmi   :
 obs.  mean   median  s.d.   min.   max.
 34623 22.775 22.432  3.616  5.933  64.516

 ==================
sbp   :
 obs.  mean    median  s.d.   min.   max.
 34516 122.089 120     14.787 81     254

 ==================
dbp   :
 obs.  mean   median  s.d.   min.   max.
 34440 75.441 75      9.642  41     158

 ==================
bslevel   :
 obs.  mean   median  s.d.   min.   max.
 32101 89.9   88      17.325 10     350
```

8.2.2 คำนวณค่าทางสถิติในภาพรวม เพื่อแสดงจำนวนชุดข้อมูลที่มีอยู่ในแต่ละตัวแปร

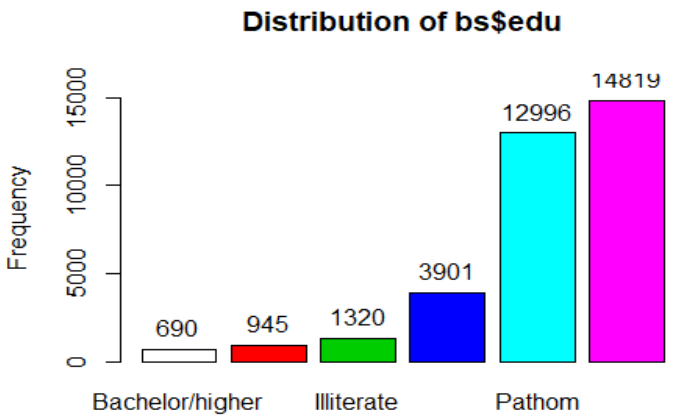
```
summ(bs)
 No. of observations = 34671

   Var. name obs.  mean   median  s.d.    min.   max.
 1 id        34671 17336  17336   10008.8 1      34671
 2 gender    34671 1.463  1       0.499   1      2
 3 age       34671 58.67  57.49   12.45   38.03  103
 4 mstatus   33891 1.402  1       0.608   1      3
```

```
 5 edu       19852 4.389  5       1.036   1      5
 6 bmi       34623 22.77  22.43   3.62    5.93   64.52
 7 sbp       34516 122.09 120     14.79   81     254
 8 dbp       34440 75.44  75      9.64    41     158
 9 bslevel   32101 89.9   88      17.33   10     350
```

8.2.3 จัดกลุ่มตัวแปร edu กลุ่มที่เป็น missing ให้เป็นกลุ่ม unknown

```
bs$edu <- ifelse(is.na(bs$edu),"Unknown",as.character(bs$edu))
bs$edu <- factor(bs$edu)
tab1(bs$edu)
```




```
 bs$edu :
                 Frequency Percent Cum. percent
 Bachelor/higher       690     2.0          2.0
 Colledge              945     2.7          4.7
 Illiterate           1320     3.8          8.5
 Mathayom             3901    11.3         19.8
 Pathom              12996    37.5         57.3
 Unknown             14819    42.7        100.0
   Total             34671   100.0        100.0
```

8.2.4. เลือกเฉพาะผู้เข้ารับบริการตรวจสุขภาพที่มีผลระดับน้ำตาลในเลือด

```
bs <- bs[!is.na(bs$bslevel),]
summ(bs)

 No. of observations = 32101
```

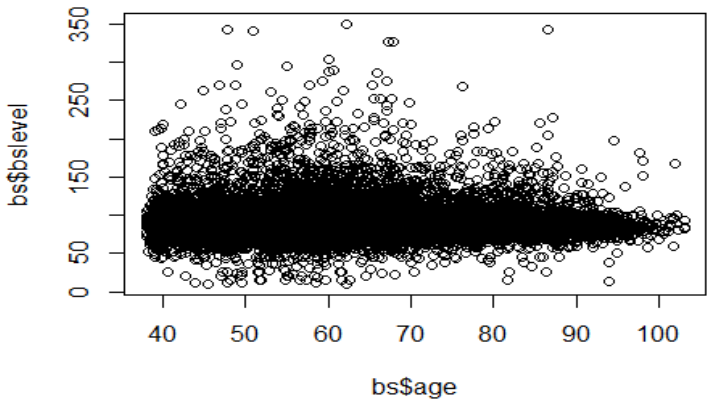
```
  Var. name obs.  mean      median  s.d.      min.   max.
1 id        32101 17555.39 17677   10007.65 1      34671
2 gender    32101 1.461    1       0.499    1      2
3 age       32101 58.77    57.57   12.43    38.03  103
4 mstatus   31396 1.403    1       0.608    1      3
5 edu       32101 5.094    5       1.109    1      6
6 bmi       32068 22.82    22.48   3.63     5.93   55.4
7 sbp       31962 122.24   120     14.94    81     242.5
8 dbp       31890 75.45    75      9.72     41     158
9 bslevel   32101 89.9     88      17.33    10     350
```

ผลการวิเคราะห์ข้อมูลเบื้องต้น พบว่า ผู้เข้ารับบริการเป็นเพศหญิง ร้อยละ 53.7 มีอายุเฉลี่ย 57.5 ปี (SD 12.5 ปี) ส่วนใหญ่สถานภาพแต่งงานแล้ว ร้อยละ 66.3 จบการศึกษาในระดับประถมศึกษาหรือต่ำกว่า ร้อยละ 65.5 มีดัชนีมวลการเฉลี่ย 22.8 (SD 3.6) ความดันโลหิตซิสโตลิกเฉลี่ย 122.1 mmHg (SD 14.8 mmHg) ความดันโลหิตไดแอสโตลิกเฉลี่ย 75.4 mmHg (SD 9.6 mmHg) มีระดับน้ำตาลในเลือดเฉลี่ย 89.9 mg/dL (SD 17.3 mg/dL)

ทำการสำรวจความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระในเบื้องต้น โดยตัวแปรแบบต่อเนื่องประกอบด้วย อายุ ดัชนีมวลกาย ความดันซิสโตลิก ความดันไอแอสโตลิก และระดับน้ำตาลในเลือด สำหรับในบทนี้คำถามในการวิจัย คือ มีปัจจัยใดบ้างที่มีความสัมพันธ์หรือมีอิทธิพลต่อระดับน้ำตาลในเลือด ดังนั้น ตัวแปรที่กำหนดให้เป็นตัวแปรตาม คือ ระดับน้ำตาลในเลือด ส่วนตัวแปรอื่นๆ ที่เหลือเป็นตัวแปรอิสระ

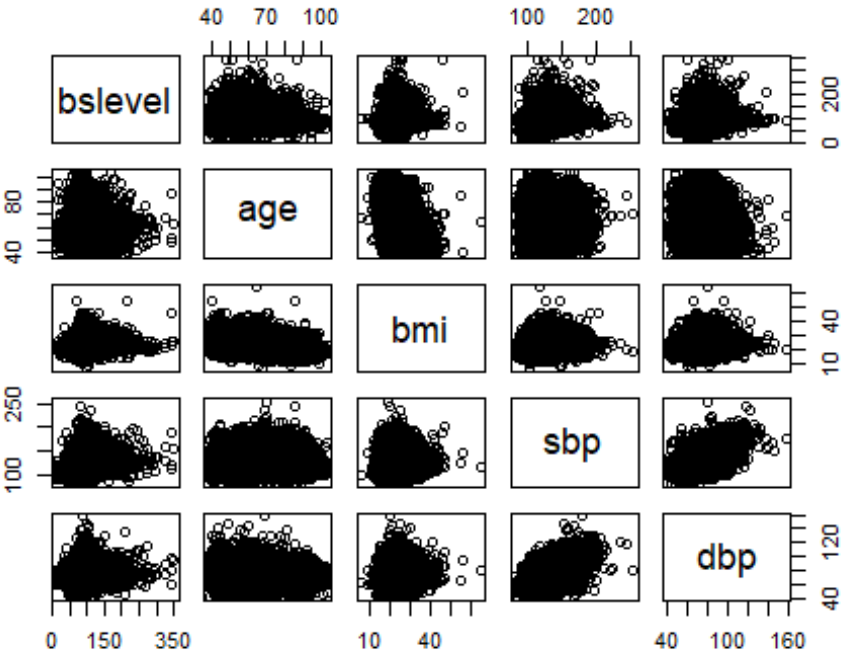
8.2.5 สร้างแผนภาพการกระจายเพื่อสำรวจความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระที่เป็นตัวแปรต่อเนื่อง

```
plot(bs$age,bs$bslevel)
```



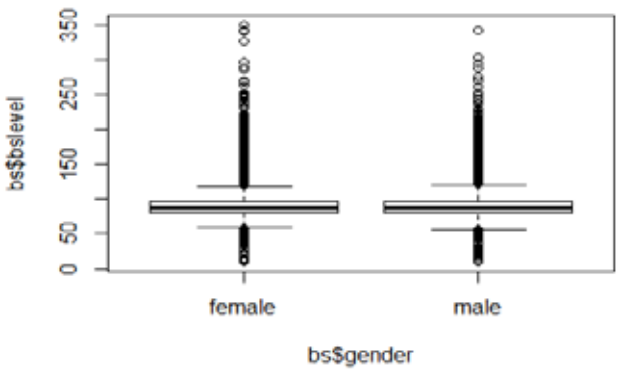
8.2.6 สร้างแผนภาพการกระจายเมทริกซ์เพื่อสำรวจความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระที่เป็นตัวแปรต่อเนื่องเป็นคู่ทีละหลายคู่

```
pairs(bs[,c(9,3,6:8)])
```

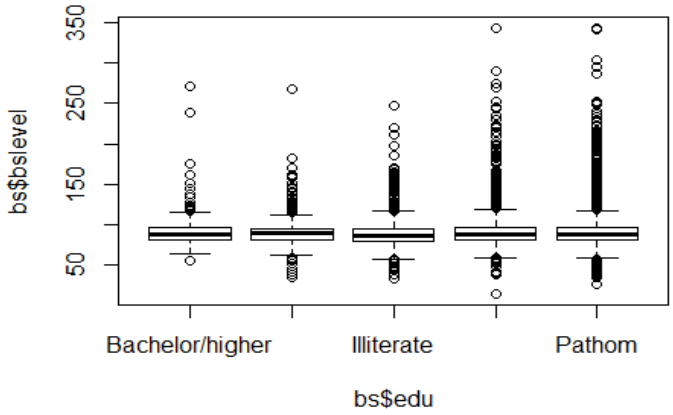


8.2.7 สร้างแผนภาพกล่องแสดงความสัมพันธ์ระหว่าตัวแปรตามและตัวแปรอิสระที่เป็นตัวแปรแบบกลุ่ม

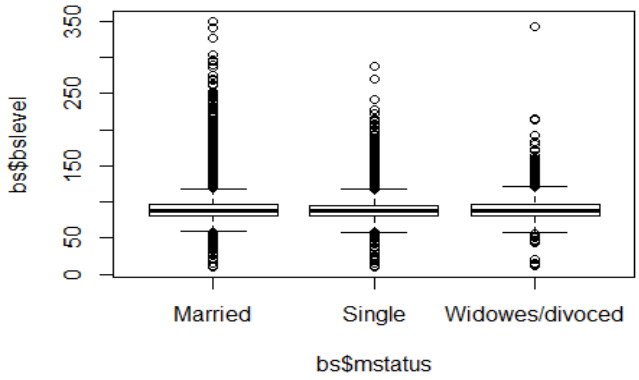
```
boxplot(bs$bslevel~bs$gender)
```



```r
boxplot(bs$bslevel~bs$edu)
```




```r
boxplot(bs$bslevel~bs$mstatus)
```



จากแผนภาพการกระจายแสดงความสัมพันธ์ระหว่างระดับน้ำตาลในเลือดและอายุ พบว่าลักษณะกราฟมีแนวโน้มให้เห็นความสัมพันธ์ในเชิงบวก นั่นคือ เมื่ออายุเพิ่มขึ้น ระดับน้ำตาลในเลือดเพิ่มขึ้นด้วยเช่นเดียวกัน แต่อย่างไรก็ตาม ไม่สามารถสรุปได้ว่ามีความสัมพันธ์กันอย่างมีนัยสำคัญหรือไม่ จนกว่าจะมีการทดสอบความสัมพันธ์ทางสถิติด้วยการถดถอยเชิงเส้น

ส่วนแผนภาพการกระจายเมทริกซ์ เป็นแผนภาพแสดงความสัมพันธ์ระหว่างตัวแปรต่อเนื่องสองตัวแปรทีละคู่จนครบจำนวนการจับคู่ที่เป็นไปได้ สำหรับแผนภาพกระจายเมทริกซ์ข้างต้น ให้พิจารณาเฉพาะแผนภาพการกระจายในแถวแรกสุด ทั้งนี้ แผนภาพดังกล่าว มีตัวแปรระดับน้ำตาลในเลือดอยู่ในแกน Y ส่วนตัวแปรอายุ ดัชนีมวลกาย ความดันโลหิตซิสโตลิก และความดันไดแอสโตลิกอยู่ในแกน X แผนภาพคู่แรก คือ ระหว่างระดับน้ำตาลในเลือดและอายุ ซึ่งได้อธิบายไปแล้ว แผนภาพถัดมา คือ แผนภาพระหว่างระดับน้ำตาลในเลือดและดัชนีมวลกาย แผนภาพที่สาม คือ แผนภาพระหว่างระดับน้ำตาลในเลือดและความดันซิสโตลิก และแผนภาพสุดท้าย

ในแถวแรก คือ ระดับน้ำตาลในเลือดและความดันไอแอสโตลิก จะเห็นได้ว่าความสัมพันธ์ระหว่างระดับน้ำตาลในเลือด และดัชนีมวลกาย ความดันโลหิตซิสโตลิก ความดันโลหิตไดแอสโตลิก มีความสัมพันธ์เชิงบวกจึงต้องมีการทดสอบทางสถิติในขั้นถัดไป

ในส่วนของความสัมพันธ์ระหว่างระดับน้ำตาลในเลือดกับตัวแปรอิสระแบบกลุ่ม จากแผนภาพกล่องพบว่าเส้นมัธยฐานที่อยู่กลางกล่องในแต่ละกลุ่มใกล้เคียงกัน ทำให้มีแนวโน้มไม่มีความสัมพันธ์กัน หรือมีความสัมพันธ์กันน้อย

8.3 การถดถอยเชิงเส้นอย่างง่าย (Simple linear regression)

การวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรตามที่เป็นตัวแปรแบบต่อเนื่องกับตัวแปรอิสระหนึ่งตัว หรือการทำนายค่าของตัวแปรตาม โดยอาศัยค่าจากตัวแปรอิสระหนึ่งตัว สมการถดถอยเชิงเส้นอย่างง่าย คือ

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

โดยที่ i แทนค่าของตัวแปรอิสระ x_i และตัวแปรตาม y_i แต่ละคู่ (x_i,y_i) ส่วน β_0 และ β_1 เรียกว่าสัมประสิทธิ์การถดถอยเมื่อ β_0 คือ จุดตัดบนแกน Y (Intercept) และ β_1 คือ ความชัน (Slope) และ ε_i คือ ความผิดพลาดหรือความคลาดเคลื่อน (Error)

ขั้นตอนถัดไปหลังจากการสำรวจความสัมพันธ์เบื้องต้นด้วยการสร้างแผนภาพ คือ การวิเคราะห์ตัวแปรเดียว (Univariate analysis) นั่นก็คือ การหาความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระทีละหนึ่งตัว เป็นคู่จนครบตามจำนวนตัวแปรอิสระที่มีอยู่ โดยในขั้นตอนนี้เป็นการสำรวจตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามอย่างมีนัยสำคัญทางสถิติ เพื่อประโยชน์ในการคัดเลือกตัวแปรอิสระมาสร้างตัวแบบหลายตัวแปร สำหรับการวิเคราะห์หลายตัวแปร (Multivariate analysis) ในขั้นถัดไป

8.3.1 การวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย

```
m1 <- lm(data=bs,bslevel~age)
summary(m1)

 Call:
 lm(formula = bslevel ~ age, data = bs)

 Residuals:
     Min      1Q  Median      3Q     Max
 -80.073  -9.113  -1.923   5.940 259.928

 Coefficients:
              Estimate Std. Error t value Pr(>|t|)
 (Intercept) 87.001590   0.467018 186.292  < 2e-16 ***
 age          0.049325   0.007775   6.344 2.26e-10 ***
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1
```

```
 Residual standard error: 17.31 on 32099 degrees of freedom
 Multiple R-squared:  0.001252,   Adjusted R-squared:  0.001221
 F-statistic: 40.25 on 1 and 32099 DF,  p-value: 2.263e-10

m2 <- lm(data=bs,bslevel~gender)
summary(m2)

 Call:
 lm(formula = bslevel ~ gender, data = bs)

 Residuals:
     Min      1Q  Median      3Q     Max
 -80.091  -9.091  -1.678   5.909 259.909

 Coefficients:
             Estimate Std. Error t value Pr(>|t|)
 (Intercept)  90.0905     0.1317 683.892   <2e-16 ***
 gendermale   -0.4122     0.1940  -2.125   0.0336 *
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

 Residual standard error: 17.32 on 32099 degrees of freedom
 Multiple R-squared:  0.0001407,  Adjusted R-squared:  0.0001095
 F-statistic: 4.517 on 1 and 32099 DF,  p-value: 0.03357
m3 <- lm(data=bs,bslevel~edu)
summary(m3)

 Call:
 lm(formula = bslevel ~ edu, data = bs)

 Residuals:
     Min      1Q  Median      3Q     Max
 -79.414  -9.414  -1.960   6.040 260.586

 Coefficients:
               Estimate Std. Error t value Pr(>|t|)
 (Intercept)   90.05107    0.70268 128.154   <2e-16 ***
 eduColledge    0.56587    0.92272   0.613   0.5397
 eduIlliterate -0.96458    0.86060  -1.121   0.2624
 eduMathayom    1.61394    0.75916   2.126   0.0335 *
 eduPathom     -0.09098    0.72052  -0.126   0.8995
 eduUnknown    -0.63705    0.71774  -0.888   0.3748
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1
```

```
 Residual standard error: 17.31 on 32095 degrees of freedom
 Multiple R-squared:  0.001651,	Adjusted R-squared:  0.001495
 F-statistic: 10.62 on 5 and 32095 DF,  p-value: 3.307e-10

m4 <- lm(data=bs,bslevel~mstatus)
summary(m4)

 Call:
 lm(formula = bslevel ~ mstatus, data = bs)

 Residuals:
     Min      1Q  Median      3Q     Max 
 -80.364  -8.631  -1.631   5.636 259.636 

 Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
 (Intercept)             90.3644     0.1201 752.441  < 2e-16 ***
 mstatusSingle           -1.7336     0.2219  -7.813 5.77e-15 ***
 mstatusWidowes/divoced   0.2544     0.4034   0.631    0.528    
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

 Residual standard error: 17.31 on 31393 degrees of freedom
   (705 observations deleted due to missingness)
 Multiple R-squared:  0.002057,	Adjusted R-squared:  0.001993 
 F-statistic: 32.35 on 2 and 31393 DF,  p-value: 9.205e-15
m5 <- lm(data=bs,bslevel~bmi)
summary(m5)

 Call:
 lm(formula = bslevel ~ bmi, data = bs)

 Residuals:
     Min      1Q  Median      3Q     Max 
 -79.757  -8.920  -1.860   5.889 259.073 

 Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
 (Intercept) 79.91334    0.61390  130.17   <2e-16 ***
 bmi          0.43785    0.02657   16.48   <2e-16 ***
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1
```

```
 Residual standard error: 17.25 on 32066 degrees of freedom
   (33 observations deleted due to missingness)
 Multiple R-squared:  0.008396,   Adjusted R-squared:  0.008365
 F-statistic: 271.5 on 1 and 32066 DF,  p-value: < 2.2e-16


m6 <- lm(data=bs,bslevel~sbp)
summary(m6)

 Call:
 lm(formula = bslevel ~ sbp, data = bs)

 Residuals:
     Min      1Q  Median      3Q     Max
 -81.408  -8.666  -1.821   5.702 254.087

 Coefficients:
              Estimate Std. Error t value Pr(>|t|)
 (Intercept) 65.947995   0.788294   83.66   <2e-16 ***
 sbp          0.195848   0.006401   30.59   <2e-16 ***
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

 Residual standard error: 17.09 on 31960 degrees of freedom
   (139 observations deleted due to missingness)
 Multiple R-squared:  0.02845,    Adjusted R-squared:  0.02842
 F-statistic:   936 on 1 and 31960 DF,  p-value: < 2.2e-16
m7 <- lm(data=bs,bslevel~dbp)
summary(m7)

 Call:
 lm(formula = bslevel ~ dbp, data = bs)

 Residuals:
     Min      1Q  Median      3Q     Max
 -80.718  -8.907  -1.907   5.904 256.746

 Coefficients:
              Estimate Std. Error t value Pr(>|t|)
 (Intercept) 76.229467   0.756687  100.74   <2e-16 ***
 dbp          0.181109   0.009947   18.21   <2e-16 ***
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1
```

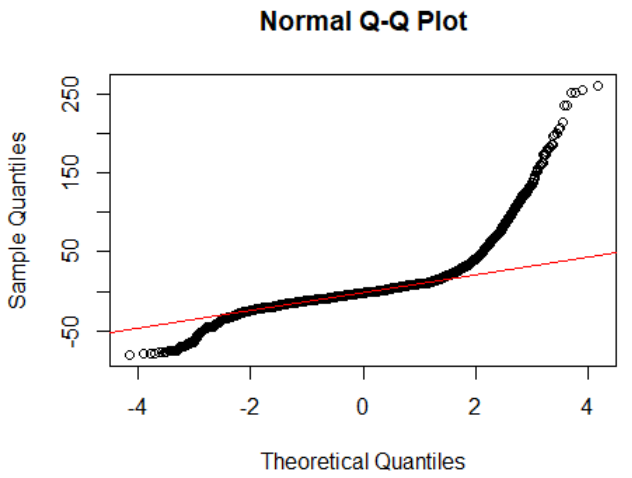
```
Residual standard error: 17.26 on 31888 degrees of freedom
  (211 observations deleted due to missingness)
Multiple R-squared:  0.01029,    Adjusted R-squared:  0.01026
F-statistic: 331.5 on 1 and 31888 DF,  p-value: < 2.2e-16
```

8.3.2 การตรวจสอบข้อตกลง (Assumption)

ก่อนที่จะทำการสรุปผลการวิเคราะห์ข้อมูล จำเป็นต้องมีการตรวจสอบข้อตกลง (Assumption) ก่อนเสมอ ทั้งนี้ การวิเคราะห์การถดถอยเชิงเส้น มีข้อตกลงในการวิเคราะห์ คือ 1) ความคลาดเคลื่อนมีการแจกแจงปกติ 2) ตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์เชิงเส้นตรง 3) ความแปรปรวนของความคลาดเคลื่อนมีค่าคงที่

การตรวจสอบว่าความคลาดเคลื่อนมีการแจกแจงปกติหรือไม่ สามารถตรวจสอบได้จากกราฟ Q-Q plot การตรวจสอบความสัมพันธ์เชิงเส้นตรง ตรวจสอบได้จากการสร้างแผนภาพการกระจายระหว่างตัวแปรตามและตัวแปรอิสระ ส่วนการตรวจสอบความคลาดเคลื่อนที่มีค่าคงที่ ตรวจสอบจากการสร้างกราฟระหว่างค่าความเคลื่อนและค่าจากการทำนาย

```
res1 <- residuals(m1)
qqnorm(res1)
qqline(res1,col="red")
```




```
res2 <- residuals(m2)
qqnorm(res2)
qqline(res2,col="red")
res3 <- residuals(m3)
qqnorm(res3)
qqline(res3,col="red")
res4 <- residuals(m4)
qqnorm(res4)
qqline(res4,col="red")
```

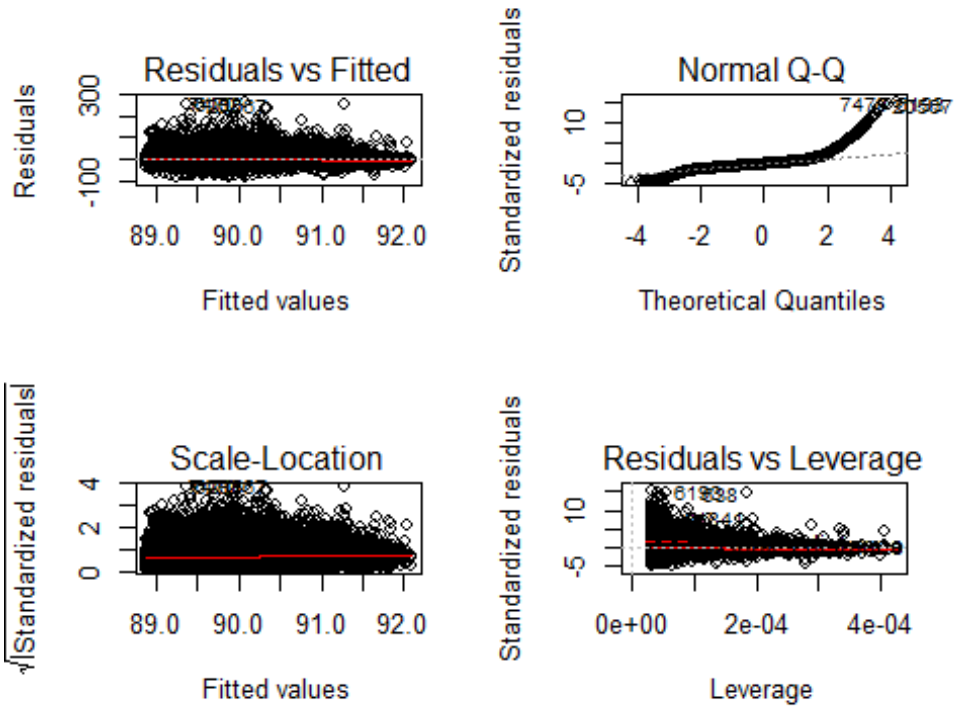
```
res5 <- residuals(m5)
qqnorm(res5)
qqline(res5,col="red")
res6 <- residuals(m6)
qqnorm(res6)
qqline(res6,col="red")
res7 <- residuals(m6)
qqnorm(res7)
qqline(res7,col="red")
```

ผลจากการสร้าง Q-Q plot พบว่า ค่าความคลาดเคลื่อนจากทุกตัวแบบไม่เป็นไปตามข้อตกลงของการถดถอยเชิงเส้น นั่นคือ ค่าความคลาดเคลื่อนมีการแจกแจงไม่เป็นปกติ

การตรวจสอบข้อตกลงทั้งสามข้อสามารถตรวจสอบได้จากการ plot ผลการวิเคราะห์ที่ได้จากการถดถอยเชิงเส้น ด้วยคำสั่ง plot() และต้องกำหนดจำนวนรูปที่มีในหน้าต่างเป็นขนาด 2 แถว 2 คอลัมน์ ด้วยคำสั่ง par() ซึ่งเป็น การนำผลการวิเคราะห์จากตัวแบบที่เก็บไว้ในชื่อใดชื่อหนึ่ง จากนั้นทำการสร้างกราฟจำนวน 4 กราฟลงในหน้าต่างเดียวกันดังตัวอย่างคำสั่งดังต่อไปนี้

การตรวจสอบข้อตกลงของการถดถอยเชิงเส้นจากการสร้างกราฟจำนวนสี่กราฟ

```
par(mfrow=c(2,2))
plot(m1)
```



```
plot(m2)
plot(m3)
plot(m4)
plot(m5)
plot(m6)
```

สำหรับคำสั่งในการสร้าง plot จำนวน 4 plot เพื่อประโยชน์ในการตรวจสอบตัวแบบที่สร้างขึ้นเป็นไปตามข้อตกลงของการทดสอบทางสถิติของตัวแบบนั้นๆ หรือไม่

กราฟรูปแรกบนสุดด้านซ้ายมือ เป็นการ plot ระหว่าง ค่าที่ได้จากการทำนายจากตัวแบบที่สร้างขึ้นกับค่าความคลาดเคลื่อนหรือค่าเศษเหลือ เป็นกราฟสำหรับตรวจสอบว่าความคลาดเคลื่อนมีความแปรปรวนคงที่หรือไม่ ถ้าหากมีความแปรปรวนคงที่ เส้นในแนวนอนจะมีลักษณะเป็นเส้นตรง จุดที่อยู่บนเส้นแนวนอนและใต้เส้นแนวนอนมีการกระจายเท่าๆ กัน

กราฟรูปที่ 2 ด้านบนสุดขวามือ เป็น Q-Q plot หากความคลาดเคลื่อนมีการแจกแจงปกติ ค่าในแต่ละจุดจะตกอยู่บนเส้นทแยงมุม

กราฟรูปที่ 3 ด้านล่างสุดซ้ายมือ เป็นกราฟที่ plot ระหว่างค่าจากการทำนายกับค่ารากที่สองของความคลาดเคลื่อนมาตรฐาน กราฟนี้จะคล้ายกับกราฟแรก ถ้าหากความคลาดเคลื่อนมีความแปรปรวนคงที่ เส้นในแนวนอนจะมีลักษณะเป็นเส้นตรง

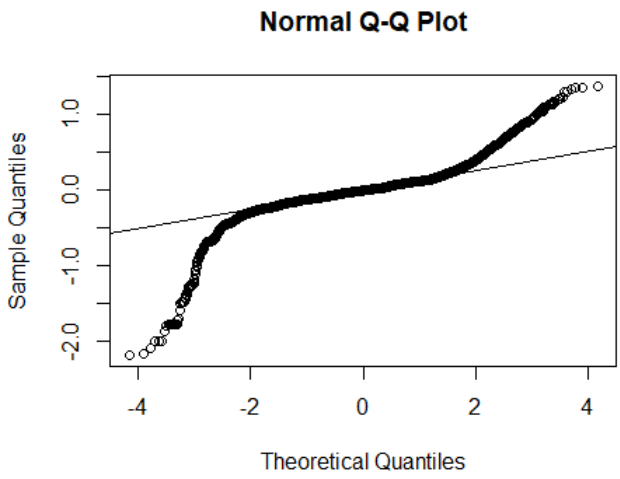
กราฟรูปที่ 4 ด้านล่างสุดขวามือ เป็นกราฟที่ plot ระหว่างค่า Leverage และค่าความคลาดเคลื่อนมาตรฐาน เป็นกราฟที่ใช้ในการพิจารณาว่ามีข้อมูลนอกเกณฑ์หรือไม่ (Outlier)

ผลจากการวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย พบว่า ตัวแบบไม่เป็นไปตามข้อตกลงในทุกตัวแบบที่สร้างขึ้น ดังนั้น จึงยังไม่สามารถนำผลการวิเคราะห์มาสรุปผลได้ จนกว่าจะมีการปรับให้ตัวแบบเป็นไปตามข้อตกลง วิธีการหนึ่งที่นิยมใช้ในการแก้ปัญหานี้ คือ การแปลงข้อมูล (Data transformation)

8.4 การแปลงข้อมูล (Data transformation)

หากตัวแบบที่สร้างขึ้นแล้วไม่เป็นไปตามข้อตกลง จำเป็นต้องปรับให้ตัวแบบเป็นไปตามข้อตกลง นั่นก็คือ การแปลงข้อมูล ซึ่งมีวิธีการแปลงที่หลากหลาย เช่น การแปลงด้วยค่ารากที่สอง การแปลงด้วยค่ารากที่สาม การแปลงด้วยค่ากำลังสอง การแปลงด้วยค่าลอกการิทึม เป็นต้น ตัวอย่างของข้อมูลชุดนี้ทำการแปลงค่าของข้อมูลระดับน้ำตาลในเลือดด้วยค่าลอการิทึมฐานธรรมชาติ

```
bs$lbs <- log(bs$bslevel)
m1 <- lm(data=bs,lbs~age)
```




```
summary(m1)

 Call:
 lm(formula = lbs ~ age, data = bs)

 Residuals:
      Min       1Q   Median       3Q      Max
 -2.18213 -0.09228 -0.00566  0.07951  1.37323

 Coefficients:
              Estimate Std. Error t value Pr(>|t|)
 (Intercept) 4.455e+00  4.718e-03 944.307  < 2e-16 ***
 age         4.770e-04  7.854e-05   6.074 1.26e-09 ***
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

 Residual standard error: 0.1749 on 32099 degrees of freedom
 Multiple R-squared:  0.001148,   Adjusted R-squared:  0.001117
 F-statistic: 36.89 on 1 and 32099 DF,  p-value: 1.262e-09
qqnorm(residuals(m1))
qqline(residuals(m1))
m2 <- lm(data=bs,lbs~gender)
summary(m2)
qqnorm(residuals(m2))
qqline(residuals(m2))
m3 <- lm(data=bs,lbs~edu)
summary(m3)
qqnorm(residuals(m3))
```

```
qqline(residuals(m3))
m4 <- lm(data=bs,lbs~mstatus)
summary(m4)
qqnorm(residuals(m4))
qqline(residuals(m4))
m5 <- lm(data=bs,lbs~bmi)
summary(m5)
qqnorm(residuals(m5))
qqline(residuals(m5))
m6 <- lm(data=bs,lbs~sbp)
summary(m6)
qqnorm(residuals(m6))
qqline(residuals(m6))
m7 <- lm(data=bs,lbs~dbp)
summary(m7)
qqnorm(residuals(m7))
qqline(residuals(m7))
```

แม้ว่าจะมีการแปลงข้อมูลระดับน้ำตาลในเลือดด้วยค่าลอการิทึมแล้วก็ตาม แต่ค่าความคลาดเคลื่อนยังมีการแจกแจงที่ไม่ปกติ ทั้งนี้อาจเป็นเพราะกลุ่มตัวอย่างเป็นคนละกลุ่มกัน เช่น กลุ่มตัวอย่างที่เป็นโรคเบาหวานแล้ว และกลุ่มตัวอย่างปกติที่ยังไม่เป็นโรคเบาหวาน การแปลงข้อมูล จึงยังไม่สามารถแก้ปัญหาตัวแบบที่ไม่เป็นไปตามข้อตกลงได้ ดังนั้น จึงอาจจะต้องทำการวิเคราะห์แยกกลุ่มระหว่างกลุ่มสองกลุ่มนี้ออกจากกัน

8.5 การวิเคราะห์แยกกลุ่ม (Stratified analysis)

วิธีการแก้ปัญหาของตัวแบบที่ไม่เป็นไปตามข้อตกลงถัดมา คือ การวิเคราะห์แยกกลุ่ม โดยปกติแล้วระดับน้ำตาลในเลือด มีเกณฑ์การพิจารณา ดังนี้ คือ กลุ่มปกติ คือ กลุ่มที่มีระดับน้ำตาลในเลือดปกติ อยู่ระหว่าง 54-99 mg/dL กลุ่มเสี่ยง คือ กลุ่มที่มีระดับน้ำตาลในเลือด อยู่ระหว่าง 100-126 mg/dL และกลุ่มผิดปกติ คือ กลุ่มที่มีภาวะระดับน้ำตาลในเลือด 126 mg/dL ขึ้นไป หรือ มีค่าน้อยกว่า 54 mg/dL ในที่นี้ ทำการแบ่งข้อมูลเฉพาะกลุ่มปกติ คือ รวมกลุ่มปกติและกลุ่มเสี่ยงเข้าด้วยกัน

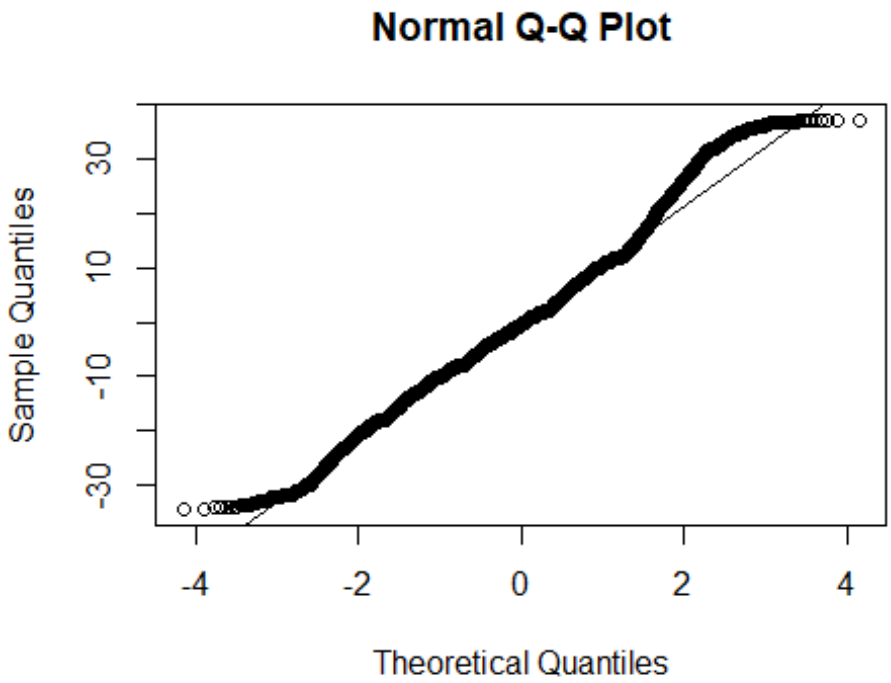
การวิเคราะห์การถดถอยอย่างง่ายในกลุ่มที่มีระดับน้ำตาลในเลือดอยู่ในกลุ่มปกติ

```
bsn <- bs[bs$bslevel>53 & bs$bslevel<126,]
mn1 <- lm(data=bsn,bslevel~age)
summary(mn1)
 Call:
 lm(formula = bslevel ~ age, data = bsn)

 Residuals:
     Min      1Q  Median      3Q     Max
 -34.508  -7.803  -0.400   6.892  37.227
```

```
 Coefficients:
             Estimate Std. Error t value Pr(>|t|)
 (Intercept) 86.949029   0.301980 287.930  < 2e-16 ***
 age          0.021339   0.005034   4.239 2.25e-05 ***
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

 Residual standard error: 11.01 on 30999 degrees of freedom
 Multiple R-squared:  0.0005794,  Adjusted R-squared:  0.0005471
 F-statistic: 17.97 on 1 and 30999 DF,  p-value: 2.25e-05
qqnorm(residuals(mn1))
qqline(residuals(mn1))
```




```
mn2 <- lm(data=bsn,bslevel~gender)
summary(mn2)
qqnorm(residuals(mn2))
qqline(residuals(mn2))
mn3 <- lm(data=bsn,bslevel~edu)
summary(mn3)
qqnorm(residuals(mn3))
qqline(residuals(mn3))
mn4 <- lm(data=bsn,bslevel~mstatus)
```

```
summary(mn4)
qqnorm(residuals(mn4))
qqline(residuals(mn4))
mn5 <- lm(data=bsn,bslevel~bmi)
summary(mn5)
qqnorm(residuals(mn5))
qqline(residuals(mn5))
mn6 <- lm(data=bsn,bslevel~sbp)
summary(mn6)
qqnorm(residuals(mn6))
qqline(residuals(mn6))
mn7 <- lm(data=bsn,bslevel~dbp)
summary(mn7)
qqnorm(residuals(mn7))
qqline(residuals(mn7))
```

จาก Q-Q plot พบว่า ค่าความคลาดเคลื่อนจากตัวแบบที่สร้างขึ้นทั้งหมดจำนวน 7 ตัวแบบ มีการแจกแจงปกติ ซึ่งเป็นไปตามข้อตกลงของตัวแบบ ดังนั้น จึงสามารถดำเนินการวิเคราะห์หลายตัวแปรต่อไปได้

8.6 การวิเคราะห์ความสัมพันธ์แบบตัวแปรเชิงเดี่ยว (Univariate analysis)

การวิเคราะห์หาความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระทีละหนึ่งตัว นอกจากวิเคราะห์ด้วยการถดถอยเชิงเส้นอย่างง่ายแล้ว ยังสามารถวิเคราะห์ความสัมพันธ์ในเบื้องต้น เพื่อคัดเลือกตัวแปรอิสระเก็บไว้ในตัวแบบหลายตัวแปร เช่น การทดสอบ t-test สำหรับกรณีที่ตัวแปรตามแบบต่อเนื่องและตัวแปรอิสระมีสองกลุ่ม และการทดสอบ ANOVA สำหรับตัวแปรตามแบบต่อเนื่องและตัวแปรอิสระที่มีมากกว่าสองกลุ่มขึ้นไป สามารถทำการวิเคราะห์ความสัมพันธ์ดังกล่าวทีละคู่ จนครบตามจำนวนตัวแปรอิสระแบบกลุ่มที่มีอยู่ ด้วยคำสั่ง statStack() ที่อยู่ในไลบรารี epiDisplay โดยคำสั่งนี้ จะเลือกใช้ชนิดของสถิติตามความเหมาะสม หากตัวแปรตามมีการแจกแจงไม่ปกติ สถิติที่ใช้ในการทดสอบจึงเป็นสถิตินอนพาราเมทริกซ์แทน สำหรับคำสั่งนี้เป็นคำสั่งที่ง่าย ใช้สะดวก แต่ข้อเสีย คือ ตัวแปรอิสระจะต้องเป็นตัวแปรแบบกลุ่มเท่านั้น ในที่นี้ตัวแปรอิสระที่เป็นตัวแปรกลุ่ม คือ gender mstatus และ edu ในกรณีที่ตัวแปรอิสระเป็นตัวแปรแบบต่อเนื่อง คำสั่งนี้ไม่สามารถใช้ได้ ซึ่งการหาความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระที่เป็นตัวแปรแบบต่อเนื่อง ทำได้ด้วยการวิเคราะห์การถดถอยอย่างง่าย หรือสามารถใช้การวิเคราะห์สหสัมพันธ์ ด้วยคำสั่ง cor()

```
> statStack(data=bsn,bslevel,by=c(gender,mstatus,edu))
    Total     bslevel    mean(SD)    Test                        P value
gender                            t-test  (30999 df) = 1.99 0.046
female          16677               88.3 (11.1)
male            14324               88.1 (10.9)
```

```
      mstatus                    Kruskal-Wallis test       < 0.001
Married          20020            88 (81,95)
Single           8370             87 (80,95)
Widowes/divoced 1935              88 (81,96)

      edu                      Kruskal-Wallis test       < 0.001
Bachelor/higher 597               89 (82,95)
Colledge         807              89 (82,95)
Illiterate       1167             87 (80,95)
Mathayom         3481             89 (82,96)
Pathom           11477            88 (81,95)
Unknown          13472            87 (80,95)
```

ผลการวิเคราะห์ข้างต้น พบว่า เพศ สถานภาพสมรส และ ระดับการศึกษามีความสัมพันธ์กับระดับน้ำตาลในเลือดอย่างมีนัยสำคัญทางสถิติ โดยผลการวิเคราะห์จากคำสั่ง statStack() แสดงจำนวนตัวอย่างทั้งหมด ค่าเฉลี่ยและส่วนเบี่ยงมาตรฐานแยกตามแต่ละกลุ่มของตัวแปรอิสระ สถิติที่ทดสอบและค่า p-value ส่วนกรณีตัวแปรตามมีการแจกแจงไม่ปกติ ผลการวิเคราะห์จะแสดงเป็นค่ามัธยฐาน และช่วงพิสัยควอไทล์มาให้แทน

8.7 การถดถอยเชิงเส้นพหุคูณ (Multiple linear regression)

การถดถอยเชิงเส้นอย่างง่ายเป็นการสร้างตัวแบบสำหรับการพิจารณาความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระในเบื้องต้น หรือ ที่เรียกว่าการวิเคราะห์ความสัมพันธ์แบบตัวแปรเชิงเดี่ยว (Univariate analysis) เพื่อประกอบการพิจารณาในการคัดเลือกตัวแปรอิสระมาสร้างตัวแบบหลายตัวแปรสำหรับการวิเคราะห์แบบหลายตัวแปร (Multivariate analysis) การทำนายการเปลี่ยนแปลงของตัวแปรตามจากตัวแปรอิสระที่มีหลายตัวแปรสามารถอธิบายความสัมพันธ์เชิงเส้นด้วยสมการ ดังนี้

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i$$

เมื่อ β_0 คือ Intercept หรือจุดตัดบนแกน y และ β_j คือ ความชันบนพื้นผิวการถดถอย (Regression surface) โดย j = 1,2,3,...,k และ ε_i คือ ความคลาดเคลื่อน (Error or residuals)

การวิเคราะห์ตัวแปรเดียว สำหรับการศึกษาเชิงสำรวจ หรือเชิงสังเกต จึงยังไม่ใช่ข้อสรุปสุดท้าย จึงควรต้องมีการวิเคราะห์หลายตัวแปร เพื่อนำตัวแปรอิสระต่างๆ มาใส่ในตัวแบบเดียวกัน เพื่อวัตถุประสงค์ของการปรับลดอิทธิพลซึ่งกันและกัน ตัวแปรอิสระไหนที่มีอิทธิพลต่อตัวแปรตามน้อย ความสัมพันธ์นั้นจากเดิมที่มีนัยสำคัญ

จากการวิเคราะห์ตัวแปรเดียว แต่เมื่อทำการวิเคราะห์หลายตัวแปร นัยสำคัญก็อาจจะหายไปได้ ในที่นี้ ผลจากการวิเคราะห์ตัวแปรเดียว พบว่า ตัวแปรอิสระทุกตัวแปรมีความสัมพันธ์กับตัวแปรตามอย่างมีนัยสำคัญทางสถิติ จึงใส่ตัวแปรอิสระเหล่านี้ทุกตัวไว้ในตัวแบบ

การวิเคราะห์การถดถอยเชิงเส้นพหุคูณในกลุ่มที่มีระดับน้ำตาลในเลือดอยู่ในกลุ่มปกติ

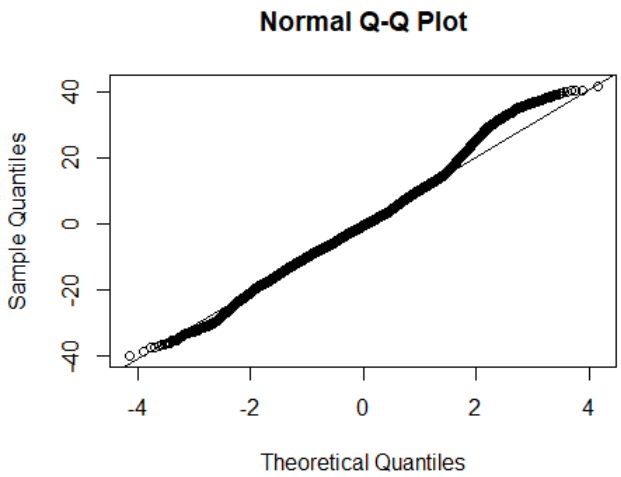
```
mfn <- lm(data=bsn,bslevel~age+gender+edu+mstatus+bmi+sbp+dbp)
summary(mfn)

 Call:
 lm(formula = bslevel ~ age + gender + edu + mstatus + bmi + sbp +
     dbp, data = bsn)

 Residuals:
     Min      1Q  Median      3Q     Max
 -39.775  -7.092  -0.484   6.611  41.692

 Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
 (Intercept)            67.103192   0.859828  78.043  < 2e-16 ***
 age                     0.021646   0.005537   3.909 9.27e-05 ***
 gendermale             -0.078928   0.138263  -0.571 0.568101
 eduColledge            -0.070042   0.595265  -0.118 0.906334
 eduIlliterate          -1.593851   0.556964  -2.862 0.004217 **
 eduMathayom             0.233896   0.484625   0.483 0.629361
 eduPathom              -0.685382   0.461130  -1.486 0.137208
 eduUnknown             -0.893244   0.459237  -1.945 0.051777 .
 mstatusSingle          -0.619404   0.159590  -3.881 0.000104 ***
 mstatusWidowes/divoced -0.066643   0.264924  -0.252 0.801388
 bmi                     0.224948   0.018019  12.484  < 2e-16 ***
 sbp                     0.102890   0.005249  19.601  < 2e-16 ***
 dbp                     0.040351   0.007938   5.083 3.74e-07 ***
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

 Residual standard error: 10.84 on 30074 degrees of freedom
   (914 observations deleted due to missingness)
 Multiple R-squared:  0.03736,    Adjusted R-squared:  0.03698
 F-statistic: 97.26 on 12 and 30074 DF,  p-value: < 2.2e-16
qqnorm(residuals(mfn))
qqline(residuals(mfn))
```



การทดสอบ F-test เพื่อตัดตัวแปรที่ไม่มีนัยสำคัญออกจากตัวแบบ

```
anova(mfn)
 Analysis of Variance Table

 Response: bslevel
             Df  Sum Sq Mean Sq  F value    Pr(>F)
 age          1    2100    2100  17.8800 2.360e-05 ***
 gender       1     523     523   4.4526   0.03486 *
 edu          5    8892    1778  15.1432 6.890e-15 ***
 mstatus      2    3458    1729  14.7216 4.070e-07 ***
 bmi          1   32133   32133 273.6047 < 2.2e-16 ***
 sbp          1   86934   86934 740.2250 < 2.2e-16 ***
 dbp          1    3034    3034  25.8370 3.737e-07 ***
 Residuals 30074 3531989     117
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1
```

ผลจากการตรวจสอบข้อตกลงด้วย Q-Q plot พบว่า ค่าความคลาดเคลื่อนมีการแจกแจงปกติ อย่างไรก็ตาม ผลการวิเคราะห์หลายตัวแปร พบตัวแปรอายุที่มีค่า p-value มากกว่าระดับนัยสำคัญทางสถิติ คือ 0.05 จึงควรทำการทดสอบด้วย F-test เพื่อใช้ประกอบการพิจารณาตัดตัวแปรที่ไม่มีนัยสำคัญออกจากตัวแบบ ผลจากการวิเคราะห์ พบว่า ไม่สามารถตัดตัวแปรใดออกจากตัวแบบได้ (ค่า p-value จาก F-test มีค่าน้อยกว่า 0.05 ทุกตัวแปร) จึงสรุปผลจากการวิเคราะห์การถดถอยเชิงเส้นพหุคูณข้างต้นได้ดังนี้ ระดับน้ำตาลในเลือดมีความสัมพันธ์กับอายุ การศึกษา สถานภาพสมรส ดัชนีมวลกาย ความดันโลหิตซิสโตลิก และความดันโลหิตไดแอสโตลิก โดย อายุ ดัชนีมวลกาย ความดันโลหิตซิสโตลิก และความดันโลหิตไดแอสโตลิกที่เพิ่มขึ้นมีผลต่อระดับน้ำตาล

ในเลือดที่เพิ่มขึ้น ส่วนระดับการศึกษา พบว่า คนที่ไม่ได้เรียนหนังสือมีระดับน้ำตาลในเลือดต่ำกว่าคนที่มีการศึกษาในระดับปริญญาตรีขึ้นไป โดยเฉลี่ย 1.6 mg/dL ตามลำดับ คนที่มีสถานภาพโสดมีระดับน้ำตาลในเลือดต่ำกว่าคนที่มีสถานภาพสมรส โดยเฉลี่ย 0.6 mg/dL

8.8 แบบฝึกหัดท้ายบท

จากแฟ้มข้อมูล psncd ให้ทำการวิเคราะห์หาว่าปัจจัยเหล่านี้ (age, mstatus, edu, smoke, alcohol, dm) มีผลต่อความสัมพันธ์ของการเพิ่มขึ้นของดัชนีมวลกาย (bmi) หรือไม่ โดยใช้วิธีการวิเคราะห์การถดถอยเชิงเส้นแบบตัวแปรเดียวและหลายตัวแปร

8.9 บรรณานุกรม

1. อภิรดี แซ่ลิ่ม (2559), การจัดการข้อมูล กราฟ และการวิเคราะห์ทางสถิติด้วยโปรแกรม R, บริษัท ไอคิวมีเดียดีไซน์: สงขลา, 434 หน้า.
2. Maindonald, J. and Braum, W.J. (2010). Data Analysis and Graphics Using R - An example-Based Approach Third Edition. New York: Cambridge University Press.
3. R Core Team. (2019). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
4. Venables, W.N. and Ripley, B.D. (2002). Modern Applied Statistics with S-PLUS. Fourth Edition. New York: Springer.

บทที่ 9

การวิเคราะห์การถดถอยโลจิสติก
(Logistic regression analysis)

บทที่ 9

การวิเคราะห์การถดถอยโลจิสติก (Logistic regression analysis)

ดร.นิรันดร์ อินทรัตน์
E-mail: nirun.i@msu.ac.th

การวิเคราะห์ข้อมูลโดยใช้ตัวแบบทางสถิติที่ชื่อว่า การถดถอยพหุโลจิสติก หรือ Multiple logistic regression จะใช้ในกรณีที่ผลลัพธ์การศึกษา (ตัวแปรตาม : y) มีลักษณะเป็นตัวแปรกลุ่มจำนวนสองกลุ่มเท่านั้น เช่น การเป็นโรคเบาหวาน (เป็นโรค กับ ไม่เป็นโรค) หรือ การเป็นโรคความดันโลหิตสูง (เป็นโรค กับ ไม่เป็นโรค) เป็นต้น ในขณะที่ปัจจัย (ตัวแปรต้น : x) จะมีลักษณะเป็นตัวแปรกลุ่ม หรือตัวแปรต่อเนื่องก็ได้ และสามารถมีได้มากกว่าหนึ่งตัวแปร โดยที่ตัวแบบทางสถิตินี้มีข้อดีคือ สามารถควบคุมตัวแปรกวน (ตัวแปรที่มีผลกระทบต่อความสัมพันธ์ของปัจจัยที่เราต้องการศึกษากับผลลัพธ์) ได้หลายตัวแปร

การวิเคราะห์ข้อมูลจะต้องทำ 2 ขั้นตอนหลักๆ คือ การวิเคราะห์แบบสองตัวแปร โดยจะทำการวิเคราะห์เพื่อหาความสัมพันธ์แบบหยาบๆ ของปัจจัยที่ต้องการศึกษา กับ ผลลัพธ์ จากนั้นเป็นขั้นตอนที่ 2 ทำโดยการเลือกตัวแปรจากขั้นตอนที่ 1 ที่มีค่า p-value น้อยกว่า 0.05 เข้ามาสู่การวิเคราะห์แบบหลายตัวแปร จากนั้นจึงทำการคัดเลือกตัวแปรที่มีอิทธิพลต่อตัวแปรตาม ทั้งนี้ เพื่อให้ได้ตัวแบบที่เหมาะสมที่สุด โดยสามารถทำนายผลลัพธ์ของการศึกษาแม่นยำที่สุด การคัดเลือกตัวแบบหลักๆ มี 2 ขั้นตอนคือ

1. แบบ forward หมายถึง เป็นการสร้างตัวแบบโดยการเริ่มต้นตัวแบบแรกเป็นตัวแบบว่าง (Null model) จากนั้นทำการนำตัวแปรที่คัดเลือกมาเข้าตัวแบบทีละตัวแปร

2. แบบ backward หมายถึง เป็นการสร้างตัวแบบโดยการเริ่มต้นตัวแบบแรกเป็นตัวแบบที่ประกอบไปด้วยตัวแปรที่คัดเลือกมาแล้วจากการวิเคราะห์สองตัวแปร การนั้นทำการดึงตัวแปรออกทีละตัวแปรโดยเลือกเอาตัวแปรที่มีค่า p-value มากที่สุดออก จนได้ตัวแบบสุดท้ายจึงทำการรายงานผลการวิเคราะห์เป็นค่า adjusted odds ratio

9.1 การวิเคราะห์ความสัมพันธ์แบบตัวแปรเชิงเดี่ยว (Univariable analysis)

ในโปรแกรม R เราสามารถใช้ epiDisplay package ใน package นี้จะมีคำสั่งที่ชื่อว่า tableStack ซึ่งใช้ในการวิเคราะห์ข้อมูลสองตัวแปรได้อย่างรวดเร็ว

ตัวอย่าง การวิเคราะห์ปัจจัยที่มีความสัมพันธ์ต่อการเป็นโรคความดันโลหิตสูง (rbp) โดยใช้ข้อมูลจากแฟ้ม psncd จำนวน 34,671 แถว

```
#table 1
tb1csv <- tableStack(vars=c(age.gr,gender,mstatus,edu,
                  smoke,alcohol,dmfamily,htfamily,bmigr,dm),by=rbp
                 ,total.column = TRUE,decimal = 2,data=data)
```

วิเคราะห์ข้อมูลโดยใช้คำสั่ง tableStack จากนั้นเก็บข้อมูลผลการวิเคราะห์เก็บไว้ในตัวแบบชื่อว่า tb1csv ซึ่งสามารถเรียกดูตัวแบบโดยการพิมพ์ชื่อตัวแบบดังกล่าว จากนั้นจะทำการส่งออกผลลัพธ์เพื่อเก็บไว้ในรูปแบบ csv โดยใช้คำสั่ง write.csv(tb1csv,”table1.csv”) นั่นคือ ต้องการส่งออกผลลัพธ์ที่ชื่อว่า tb1 เก็บไว้เป็นไฟล์ที่ชื่อว่า table8.1 ซึ่งผลลัพธ์จะเก็บไว้ในแฟ้มเดียวกับข้อมูลที่อ่านเข้ามา ทั้งนี้ ในตัวอย่างคำสั่งด้านบน เราจะได้ผลลัพธ์การวิเคราะห์ดังหน้าถัดไป

```
tb1csv
         normal   abnormal    Total  Test stat.     P value
 Total   1238        295     1533
 age.gr                              Chisq. (3 df) = 26.627 < 0.001
    (35,40]          38 (3.07)    5 (1.69)    43 (2.8)
    (40,50]          293 (23.67)  45 (15.25)  338 (22.05)
    (50,60]          383 (30.94)  72 (24.41)  455 (29.68)
    (60,110]         524 (42.33)  173 (58.64) 697 (45.47)

 gender                              Chisq. (1 df) = 3.182  0.0744
    female           684 (55.25)  146 (49.49) 830 (54.14)
    male             554 (44.75)  149 (50.51) 703 (45.86)

 mstatus                             Chisq. (2 df) = 3.122  0.21
    single           164 (13.25)  37 (12.54)  201 (13.11)
    married          968 (78.19)  223 (75.59) 1191 (77.69)
    widowed/divorced 106 (8.56)   35 (11.86)  141 (9.2)

 edu                                 Chisq. (4 df) = 11.251 0.0239
    Iilliterate      130 (10.5)   26 (8.81)   156 (10.18)
    Pathom           515 (41.6)   100 (33.9)  615 (40.12)
    Mathayom         423 (34.17)  131 (44.41) 554 (36.14)
    Colledge         105 (8.48)   24 (8.14)   129 (8.41)
    Bachelor/higher  65 (5.25)    14 (4.75)   79 (5.15)
```

```
 smoke                                Chisq. (1 df) = 7.396  0.0065
    no              1061 (85.7)  234 (79.32) 1295 (84.47)
    yes             177 (14.3)   61 (20.68)  238 (15.53)

 alcohol                            Chisq. (1 df) = 0      0.9998
    no              982 (79.32)  234 (79.32) 1216 (79.32)
    yes             256 (20.68)  61 (20.68)  317 (20.68)

 dmfamily                           Chisq. (1 df) = 4.673  0.0306
    no              1126 (90.95) 256 (86.78) 1382 (90.15)
    yes             112 (9.05)   39 (13.22)  151 (9.85)

 htfamily                           Chisq. (1 df) = 3.339  0.0677
    no              1108 (89.5)  253 (85.76) 1361 (88.78)
    yes             130 (10.5)   42 (14.24)  172 (11.22)

 bmigr                              Chisq. (1 df) = 10.147 0.0014
    normal          1202 (97.09) 275 (93.22) 1477 (96.35)
    obesity         36 (2.91)    20 (6.78)   56 (3.65)

 dm                                 Chisq. (1 df) = 25.17  < 0.001
    normal          1067 (86.19) 219 (74.24) 1286 (83.89)
    DM              171 (13.81)  76 (25.76)  247 (16.11)

write.csv(tb1csv,"table1.csv")
```

9.2 การสร้างตัวแบบการถดถอยโลจิสติก

จากผลลัพธ์การวิเคราะห์ข้อมูล ข้อ 9.1 เราจะทำการเลือกตัวแปรที่มีค่า p-value น้อยกว่า 0.05 เข้ามาวิเคราะห์ในขั้นตอนการวิเคราะห์หลายตัวแปร ประกอบไปด้วยตัวแปร age.gr, edu, smoke, dmfamily, bmigr, dm ในขั้นตอนนี้เราจะใช้วิธีการคัดเลือกตัวแปรที่มีอิทธิพลต่อตัวแปรตาม โดยใช้วิธีการคัดเลือกตัวแปรแบบย้อนกลับ (Backward elimination) ดังนี้

```
#backward elimination

m0<-glm(rbp~age.gr+edu+smoke+dmfamily+bmigr+dm,family = "binomial",
data=data)
summary(m0)

 Call:
 glm(formula = rbp ~ age.gr + edu + smoke + dmfamily + bmigr +
     dm, family = "binomial", data = data)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2379  -0.7045  -0.5559  -0.4535   2.2419

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.4771     0.5326  -4.651 3.31e-06 ***
age.gr(40,50]        0.2058     0.5090   0.404 0.685995
age.gr(50,60]        0.4295     0.5055   0.850 0.395488
age.gr(60,110]       0.9074     0.4970   1.826 0.067901 .
eduPathom            0.0485     0.2475   0.196 0.844667
eduMathayom          0.3026     0.2427   1.247 0.212464
eduColledge          0.2585     0.3278   0.788 0.430471
eduBachelor/higher   0.3000     0.3771   0.796 0.426215
smokeyes             0.4100     0.1696   2.417 0.015638 *
dmfamilyyes          0.3131     0.2075   1.509 0.131397
bmigrobesity         0.8563     0.2960   2.893 0.003813 **
dmDM                 0.5518     0.1671   3.302 0.000958 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1501.5  on 1532  degrees of freedom
Residual deviance: 1438.1  on 1521  degrees of freedom
AIC: 1462.1

Number of Fisher Scoring iterations: 4
```

จากผลการวิเคราะห์ดังกล่าวเราจะดึงตัวแปร edu ออกเพราะมีค่า p value มากที่สุด จากนั้นประมวลตัวแบบถัดไป

```
#remove edu
m1<-glm(rbp~age.gr+   smoke+dmfamily+bmigr+dm,family = “binomial”,data=data)

lrtest(m0,m1)
 Likelihood ratio test for MLE method
 Chi-squared 4 d.f. =  3.582907 , P value =  0.4653846

summary(m1)
```

```
Call:
glm(formula = rbp ~ age.gr + smoke + dmfamily + bmigr + dm, family = "binomial",
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2188  -0.6611  -0.5242  -0.4827   2.1790

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.2763     0.4843  -4.700  2.6e-06 ***
age.gr(40,50]    0.1851     0.5078   0.365 0.715417
age.gr(50,60]    0.3610     0.4996   0.723 0.469894
age.gr(60,110]   0.8669     0.4909   1.766 0.077448 .
smokeyes         0.4193     0.1687   2.485 0.012968 *
dmfamilyyes      0.3181     0.2053   1.549 0.121295
bmigrobesity     0.8919     0.2946   3.028 0.002464 **
dmDM             0.6145     0.1620   3.793 0.000149 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1501.5  on 1532  degrees of freedom
Residual deviance: 1441.6  on 1525  degrees of freedom
AIC: 1457.6

Number of Fisher Scoring iterations: 4
```

จากผลการวิเคราะห์ดังกล่าวเราจะทำการทดสอบว่าตัวแปรที่เราดึงออก (edu) มีผลต่อตัวแบบหรือไม่ โดยการทดสอบ lrtest(m0,m1) ซึ่งจากการทดสอบพบว่า ได้ค่า p value เท่ากับ 0.4654 ซึ่งมากกว่า 0.05 ดังนั้น จึงดึงตัวแปรนี้ออกได้ ขั้นตอนต่อมาทำการพิจารณาตัวแบบที่ 1 จะเห็นว่าตัวแปร dmfamily มีค่ามากที่สุดและมากกว่า 0.05 จึงดึงตัวแปรนี้ออก และประมวลผลอีกครั้ง

```
#remove dmfamily
m2<-glm(rbp~age.gr+    smoke+          bmigr+dm,family = "binomial",data=data)

lrtest(m2,m1)
 Likelihood ratio test for MLE method
 Chi-squared 1 d.f. =  2.31329 , P value =  0.1282723
```

```
summary(m2)

 Call:
 glm(formula = rbp ~ age.gr + smoke + bmigr + dm, family = "binomial",
     data = data)

 Deviance Residuals:
     Min       1Q   Median       3Q      Max
 -1.2475  -0.6673  -0.5348  -0.4881   2.1728

 Coefficients:
                Estimate Std. Error z value Pr(>|z|)
 (Intercept)     -2.2615     0.4848  -4.665 3.08e-06 ***
 age.gr(40,50]    0.1943     0.5082   0.382  0.70223
 age.gr(50,60]    0.3888     0.4996   0.778  0.43642
 age.gr(60,110]   0.8728     0.4914   1.776  0.07573 .
 smokeyes         0.4263     0.1686   2.529  0.01144 *
 bmigrobesity     0.9051     0.2952   3.066  0.00217 **
 dmDM             0.6470     0.1606   4.028 5.63e-05 ***
 ---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 (Dispersion parameter for binomial family taken to be 1)

     Null deviance: 1501.5  on 1532  degrees of freedom
 Residual deviance: 1444.0  on 1526  degrees of freedom
 AIC: 1458

 Number of Fisher Scoring iterations: 4
```

พิจารณาตัวแปรที่เราดึงออก (dmfamily) มีผลต่อการทำนายตัวแปรตามหรือไม่ โดยการทดสอบ lrtest(m2,m1) ซึ่งจากการทดสอบพบว่า ได้ค่า p value เท่ากับ 0.1283 ดังนั้น จึงดึงตัวแปรนี้ออก ขั้นตอนต่อมาทำการพิจารณาตัวแบบที่ 2 จะเห็นว่าตัวแปร age.gr มีค่ามากที่สุด จึงดึงตัวแปรนี้ออก และประมวลผลตัวแบบถัดไป

```
#remove age.gr
m3<-glm(rbp~          smoke+         bmigr+dm,family = "binomial",data=data)

lrtest(m3,m2)
 Likelihood ratio test for MLE method
 Chi-squared 3 d.f. =  19.84581 , P value =  0.0001827
```

```
summary(m3)

 Call:
 glm(formula = rbp ~ smoke + bmigr + dm, family = "binomial",
     data = data)

 Deviance Residuals:
     Min       1Q   Median       3Q      Max
 -1.3365  -0.5789  -0.5789  -0.5789   1.9333

 Coefficients:
              Estimate Std. Error z value Pr(>|z|)
 (Intercept)  -1.70130    0.08297 -20.506  < 2e-16 ***
 smokeyes      0.47465    0.16736   2.836  0.00457 **
 bmigrobesity  0.83340    0.29250   2.849  0.00438 **
 dmDM          0.75973    0.15803   4.808 1.53e-06 ***
 ---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 (Dispersion parameter for binomial family taken to be 1)

     Null deviance: 1501.5  on 1532  degrees of freedom
 Residual deviance: 1463.8  on 1529  degrees of freedom
 AIC: 1471.8

 Number of Fisher Scoring iterations: 4
```

พิจารณาตัวแปรที่เราดึงออก (age.gr) ว่ามีผลต่อตัวแปรตามหรือไม่ โดยการทดสอบ lrtest(m3,m2) ซึ่งพบว่าได้ค่า p value เท่ากับ 0.0001827 จากผลการวิเคราะห์ดังกล่าวตัวแปร age.gr มีผลต่อโมเดล จึงไม่ควรตัดออก ดังนั้น ตัวแบบสุดท้ายที่มีความเหมาะสมคือ ตัวแบบที่มีตัวแปรอิสระประกอบไปด้วย age.gr, smoke, bmigr และ dm นั้นคือ ผลการวิเคราะห์ตัวแบบที่ 2

นอกจากการคัดเลือกตัวแบบด้วยวิธีการดังกล่าวแล้ว ยังสามารถใช้คำสั่ง step ("modelname", direction =" วิธีการที่ต้องการ") เพื่อให้โปรแกรมคัดเลือกตัวแบบที่เหมาะสมให้ โดยตัวแบบที่เหมาะสมที่สุดมาจากตัวแบบที่มีค่า AIC น้อยที่สุดและจะแสดงให้เป็นตัวแบบลำดับท้ายสุด ในการยกตัวอย่างในครั้งนี้ จะใช้วิธีการย้อนกลับ ดังนั้น ตัวแบบที่จะมาใช้เริ่มต้นจะเป็นตัวแบบที่มีตัวแปรทุกตัวที่สนใจ ดังนี้

```
step(m0, direction = "backward")
 Start:  AIC=1462.05
 rbp ~ age.gr + edu + smoke + dmfamily + bmigr + dm

            Df Deviance    AIC
```

```
- edu       4   1441.6 1457.6
<none>          1438.0 1462.0
- dmfamily  1   1440.2 1462.2
- smoke     1   1443.7 1465.7
- bmigr     1   1445.8 1467.8
- dm        1   1448.6 1470.6
- age.gr    3   1457.5 1475.5

Step:  AIC=1457.64
rbp ~ age.gr + smoke + dmfamily + bmigr + dm

           Df Deviance    AIC
<none>          1441.6 1457.6
- dmfamily  1   1444.0 1458.0
- smoke     1   1447.6 1461.6
- bmigr     1   1450.0 1464.0
- dm        1   1455.4 1469.4
- age.gr    3   1462.1 1472.1

Call:  glm(formula = rbp ~ age.gr + smoke + dmfamily + bmigr + dm, family =
"binomial",
    data = data)

Coefficients:
   (Intercept)   age.gr(40,50]   age.gr(50,60]  age.gr(60,110]        smokeyes
       -2.2763          0.1851          0.3610          0.8669          0.4193
   dmfamilyyes    bmigrobesity            dmDM
        0.3181          0.8919          0.6145

Degrees of Freedom: 1532 Total (i.e. Null);  1525 Residual
Null Deviance:       1502
Residual Deviance: 1442  AIC: 1458
```

หากพิจารณา จะพบว่าตัวแบบสุดท้ายที่ได้จากการใช้คำสั่ง step() ประกอบด้วย age.gr, smoke, dmfamily, bmigr และ dm ซึ่งพบว่า มีข้อแตกต่างจากวิธีก่อนหน้า คือ ตัวแบบสุดท้ายไม่มี dmfamily อยู่ด้วย ดังนั้น ขึ้นอยู่กับผู้วิเคราะห์ที่จะพิจารณาว่าในทางการแพทย์ ตัวแปร dmfamily มีผลต่อการเป็นโรคความดันโลหิตสูงหรือไม่

เมื่อเราได้ตัวแบบสุดท้ายแล้ว จะทำการแสดงผลการวิเคราะห์ข้อมูลโดยแสดงค่า crude odds ratio และ adjusted odds ratio โดยใช้คำสั่ง logistic.display()

```
resultLR<-logistic.display(m2)
resultLR

 Logistic regression predicting rbp : abnormal vs normal

          crude OR(95%CI adj. OR(95%CI) P(Wald's test)    P(LR-test)
 age.gr: ref.=(35,40]
    (40,50]  1.17 (0.44,3.12) 1.21 (0.45,3.29) 0.702                 <0.001
    (50,60]  1.43 (0.54,3.75) 1.48 (0.55,3.93) 0.436
    (60,110] 2.51 (0.97,6.48) 2.39 (0.91,6.27) 0.076

 smoke:
       yes vs no    1.56 (1.13,2.16) 1.53 (1.1,2.13    0.011               0.013

 bmigr:
      obesity vs normal 2.43 (1.38,4.26) 2.47 (1.39,4.41) 0.002  0.003

 dm: DM vs normal 2.17 (1.59,2.94) 1.91 (1.39,2.62) <0.001   <0.001

 Log-likelihood = -721.9752
 No. of observations = 1533
 AIC value = 1457.9504
```

การแปลผล จากผลการวิเคราะห์ พบว่า คนที่มีอายุเพิ่มมากขึ้นจะทำให้มีโอกาสเสี่ยงต่อการเป็นโรคความดันโลหิตมากยิ่งขึ้น และคนที่สูบบุหรี่มีโอกาสเสี่ยงต่อการเป็นโรคความดันโลหิตสูง 1.56 เท่า เมื่อเทียบกันคนที่ไม่สูบบุหรี่ นอกจากนั้นแล้ว พบว่า คนที่เป็นโรคอ้วนและมีโรคเบาหวาน จะทำให้มีโอกาสเป็นโรคความดันโลหิตสูงเป็น 2.43 และ 2.17 เท่า เมื่อเทียบกับคนที่ไม่ได้เป็นโรคดังกล่าว

ข้อสังเกต หากผลการวิเคราะห์พบว่า ค่า crude OR เป็นศูนย์ เช่น 0.46 จะทำให้อธิบายผลได้ยาก แนะนำให้เปลี่ยนกลุ่มอ้างอิง และทำการประมวลผลอีกครั้ง

9.3 แบบฝึกหัด

1. จากชุดข้อมูลแฟ้ม psncd ให้ทำการวิเคราะห์ข้อมูล โดยใช้วิธีการวิเคราะห์ตัวแปรเดียว (Univariate analysis) โดยกำหนดให้ตัวแปรตามคือ bmigr (obisity และ normal)
2. ให้หาปัจจัยใดที่มีผลต่อการเกิดโรคอ้วน โดยทำการวิเคราะห์ผลต่อเนื่องจากข้อที่ 1 โดยใช้วิธีการถดถอยโลจิสติก (logistic regression) พร้อมทั้งอธิบายผลลัพธ์ที่ได้

วพส.

สถาบันวิจัยและพัฒนาสุขภาพภาคใต้
ชั้น 6 อาคารบริหารคณะแพทยศาสตร์ หน่วยระบาดวิทยา
คณะแพทยศาสตร์ มหาวิทยาลัยสงขลานครินทร์
ตำบลคอหงส์ อำเภอหาดใหญ่ จังหวัดสงขลา
โทร 075-455150
Email : southern.rdh@gmail.com
www.rdh.psu.ac.th

ISBN 978-616-271-574-7
9 786162 715747