PROBABILITY AND STATISTICS (103103)

รองศาสตราจารย์ ดร. ไพโรจน์ สัตยธรรม
สาขาวิชาคณิตศาสตร์ สำนักวิชาวิทยาศาสตร์
มหาวิทยาลัยเทคโนโลยีสุรนารี

103103

Probability and Statistics

ส่วนที่สอง Statistics สอนโดย อาจารย์ไพโรจน์ สัตยธรรม

การวัดผล คะแนนเต็มในส่วนนี้ 50 คะแนน ประกอบไปด้วย ปรนัย 50 คะแนน

หนังสือที่ใช้ 1) Murray R. Spiegel, Probability & Statistics ; Schaum's outline series.

2) Hines & Montgomery, Probability and Statistics in Engineer ; John-Wiley & Sons inc. 1990.

เนื้อหาที่จะเรียน

1. Elementary Sampling Theroy	3 hrs.
2. Parameter Estimation	3 hrs.
3. Test of Hypotheses	3 hrs.
4. Chi - square test	3 hrs.
5. Least squares regression	3 hrs.
6. Correlation	2 hrs.

1. Elementary Sampling Theory

1.1 ข้อสังเกตุทั่วไป

ทฤษฎีสุ่มตัวอย่างคือการศึกษาความสัมพันธ์ที่เกิดขึ้นระหว่างประชากร (population) และสิ่งตัวอย่าง (samples) สิ่งที่เราเลือกมาจากประชากรนั้น ความสัมพันธ์ระหว่างประชากรและสิ่งตัวอย่างนั้นมีหลายประเภท ตัวอย่างเช่น การประมาณค่าของประชากรบางประชากร (อาจจะเป็น population mean หรือ variance และอื่น ๆ) จากค่าที่สมนัยกัน ซึ่งหามาได้จากสิ่งตัวอย่าง (เช่น sample mean หรือ variance) จะเรียกค่าที่ได้มาจากประชากรว่า population parameters หรือ **parameters** เฉย ๆ ส่วนค่าที่ได้จากตัวอย่างนั้นจะเรียกว่า sample statistics หรือ **statistics** เฉย ๆ ปัญหาที่เกี่ยวกับการประมาณค่านั้นจะมีการศึกษาต่อไปอีกในหัวข้อที่ 2

นอกเหนือไปจากการใช้ประโยชน์ในการประมาณค่าตัวกล่าวข้างต้นแล้ว ทฤษฎีการสุ่มตัวอย่างยังมีประโยชน์ในการตัดสินว่าความแตกต่างที่สังเกตได้จากสิ่งตัวอย่างสองกลุ่มนั้นเป็นความแตกต่างที่เกิดจากโอกาสในการเลือกสิ่งตัวอย่าง หรือเป็นความแตกต่างที่มีนัยสำคัญจริง ๆ คำถามที่เกี่ยวข้องกับเรื่องราวเหล่านี้ก็คือ การตัดสินใจว่าผลิตภัณฑ์ยาชนิดใหม่ที่ได้พัฒนาขึ้นจะให้ผลในการรักษาดีกว่ายาที่มีอยู่ในปัจจุบันหรือไม่? กระบวนการผลิตสินค้าวิธีใหม่จะดีกว่าวิธีการที่กำลังใช้อยู่หรือไม่? คำตอบต่อคำถามเหล่านี้จะเกี่ยวข้องกับเรื่องราว ในการทดสอบสมมุติฐาน ซึ่งเป็นเรื่องที่สำคัญในทฤษฎีการตัดสินใจและมีการศึกษาในหัวข้อที่ 3

โดยทั่วไปแล้วการศึกษาเกี่ยวกับการทำนายค่าของประชากรโดยอาศัยประโยชน์จากสิ่งตัวอย่างที่เลือกมาจากประชากรดังกล่าว พร้อมกับการบอกระดับความแม่นของการทำนายโดยอาศัยทฤษฎีความน่าจะเป็นเข้าช่วยด้วยนี้จะเรียกว่า การอนุมานเชิงสถิติ (statistical inference)

เพื่อที่จะให้ได้ข้อสรุปของการอนุมานเชิงสถิตินี้มีความถูกต้องพอควร ผู้ที่ทำการสำรวจด้วยตัวอย่าง (sample survey) จะต้องแน่ใจพอควรว่าสิ่งตัวอย่างที่เลือกมาได้นี้เป็นตัวแทนที่ดีของกลุ่มประชากรดังกล่าว

การเลือกสิ่งตัวอย่างเพื่อให้เป็นตัวแทนที่ดีนั้นเป็นภารกิจที่ทำไม่ได้โดยง่ายเสมอไป เช่น ทราบว่าในประเทศไทยมีครัวเรือนอยู่ 10,000,000 ครัวเรือน หากต้องการจะเลือก 100,000 ครัวเรือนให้เป็นสิ่งตัวอย่างเพื่อเป็นตัวแทนครัวเรือนทั้งประเทศ ผู้อ่านควรลองวาดภาพดูว่าจะชักสิ่งตัวอย่างได้อย่างไร

การกำหนดแผนของการชักสิ่งตัวอย่าง (sample design) ซึ่งกำหนดว่าวิธีการชักสิ่งตัวอย่างจะทำได้อย่างไร และการกำหนดขนาดของตัวอย่าง (sample size) เป็นส่วนสำคัญที่สุดในการวางแผนการสำรวจ เพราะทั้งแผนและขนาดของตัวอย่างเป็นตัวกำหนดแผนปฏิบัติงานในขั้นงานสนาม การกำหนดค่าใช้จ่ายในงานสนาม ตลอดไปจนถึงค่าใช้จ่ายในการประมวลข้อมูล คุณภาพของข้อมูลซึ่งเป็นผลมาจากการชักสิ่งตัวอย่างก็จะขึ้นอยู่กับแผนและขนาดของตัวอย่างโดยตรง

แผนของการชักสิ่งตัวอย่างควรจะเป็นแผนซึ่งจะทำให้ได้ตัวอย่างมาโดยหลักเกณฑ์ทางทฤษฎีความน่าจะเป็น (เราเรียกตัวอย่างประเภทนี้ว่า probability sample) ไม่ใช่แผนซึ่งมีการเจาะจงการชักสิ่งตัวอย่างตามใจของผู้วางแผน เพราะผู้วางแผนอาจจะมีความเอนเอียงไปในทางใดทางหนึ่ง สำหรับการชักสิ่งตัวอย่างไม่ว่าจะทำโดยสุจริตใจหรือไม่ก็ตาม

สาระสำคัญของการชักสิ่งตัวอย่างโดยใช้ทฤษฎีความน่าจะเป็นก็คือ สมาชิกทุกตัวในประชากรจะต้องมีโอกาสที่จะถูกสุ่มออกมา (โอกาสนี้วัดได้โดยความน่าจะเป็นซึ่งมีค่าระหว่าง 0 กับ 1) ความน่าจะเป็นของแต่ละสมาชิกที่จะถูกสุ่มออกมาจะมีค่าเท่ากันหรือไม่ก็ได้ แต่นักสถิติจะต้องกำหนดและรู้ว่าความน่าจะเป็นของแต่ละหน่วยแจงนับที่จะถูกสุ่มออกมานั้นเป็นเท่าไร

ในกรณีพิเศษซึ่งเราวางแผนการชักสิ่งตัวอย่างไว้ว่าสมาชิกทุกตัวในกลุ่มประชากรจะต้องมีโอกาสเท่ากันหมดสำหรับการที่จะถูกเลือกเข้ามาอยู่ในสิ่งตัวอย่าง จะเรียกแผนการสุ่มชนิดนี้ว่า แผนการชักสิ่งตัวอย่างแบบธรรมดา (simple random sampling)

ส่วนตัวอย่างที่ได้มานั้นจะเรียกว่า ตัวอย่างเชิงสุ่ม (random sample) สรุปเป็นบทนิยามดังนี้คือ

Random sample (finite population)	A set of observations $x_1, x_2, \ldots, x_n$ constitute a random sample of size n from a finite population of size N, if it is chosen so that each subset of n of the N elements of the population has the same probability of being selected.

1.2 การชักสิ่งตัวอย่างแบบธรรมดา (Simple Random Sampling)

วิธีการให้ได้มาซึ่งสิ่งตัวอย่างสำหรับแผนการชักสิ่งตัวอย่างแบบธรรมดานี้อาจจะทำได้โดยวิธีการจับสลาก หรือใช้ตารางเลขสุ่มก็ได้ รายละเอียดเป็นดังนี้ คือ

วิธีการจับสลาก : สมมติว่าประชากรมีจำนวนสมาชิก N = 500 ต้องการเลือกสิ่งตัวอย่างเชิงสุ่มขนาด n = 10 มา 1 ตัวอย่าง ให้ทำการติดหมายเลขสมาชิกในกลุ่มประชากรตั้งแต่หมายเลข 1 - 500 และต่อจากนั้นให้ทำสลากหมาย 1 - 500 จำนวน 500 ใบ นำมากองรวมกันไว้และกวนสลากให้คละกันดี ต่อจากนั้นให้ทำการหยิบสลากขึ้นมาทีละใบและจดเลขหมายบนสลากไว้ (ก่อนที่จะทำการหยิบครั้งที่ 2 ให้คืนสลากที่หยิบได้ครั้งแรกลงไปด้วย) ทำเช่นนี้ต่อไปเรื่อยๆ จนได้เลขหมายบนสลากที่แตกต่างกันหมดทั้ง 10 ใบ นำเลขหมายที่ได้ไปเลือกสมาชิกในกลุ่มประชากรที่มีเลขหมายตรงกันก็จะได้สิ่งตัวอย่างขนาด 10 ตามต้องการ

วิธีการใช้ตารางเลขสุ่ม ดำเนินการติดหมายเลขตั้งแต่ 001 - 500 เช่นเดียวกันกับวิธีแรกต่อจากนั้นก็ให้เลือกเลขสามหลักที่มีค่าไม่เกิน 500 จากตารางเลขสุ่มให้ได้เลขสามหลักที่แตกต่างกัน 10 จำนวน ต่อจากนั้นก็นำเลข 10 จำนวนดังกล่าวไปชักสิ่งตัวอย่างขนาด 10 ได้ตามต้องการ

วิธีการชักสิ่งตัวอย่างดังที่ได้บรรยายไว้แล้ว ดูอย่างผิวเผินก็ไม่น่ามีปัญหาอะไรนัก แต่ในทางปฏิบัติอาจมีข้ออุปสรรคบางอย่าง ทั้งในกระบวนการชักสิ่งตัวอย่างและวิธีการเก็บสิ่งตัวอย่าง

ประการแรก การเตรียมสลากต้องเสียเวลาและค่าใช้จ่ายไม่น้อย กรณีที่ N มีขนาดใหญ่ (เช่น N = 10,000,000 ครัวเรือน) และจะเตรียมสลากได้ก็ต่อเมื่อเรามีบัญชีรายชื่อสมาชิกทุกตัว ในกลุ่มของประชากรอย่างถูกต้องแล้ว กรณีที่ขนาดของประชากรไม่ใหญ่นัก เช่น ประชากรประกอบไปด้วยนักศึกษาของ มทส. ทั้งหมดประมาณ 1,500 คน รายชื่อของนักศึกษาทั้งหมดอาจจะหาได้ที่แผนกทะเบียน แต่ถ้าประชากรเป็นครัวเรือนทั่วประเทศ ย่อมเป็นการยากที่จะจัดหาบัญชีรายชื่อของครอบครัวทั้งหมดได้อย่างถูกต้องและครบถ้วน เพราะครัวเรือนมีการโยกย้ายเปลี่ยนแปลงอยู่เสมอ และการจัดเตรียมบัญชีรายชื่อของครอบครัวทั้งหมดในกรณีนี้ถึงหากจะทำได้ โดยยอมให้ผิดพลาดบ้างก็ยังคงต้องใช้งบประมาณไม่น้อย บัญชีรายชื่อของสิ่งตัวอย่างเรียกว่า กรอบสิ่งตัวอย่าง (sampling frame) ดังนั้นในการวางแผนการสุ่มตัวอย่างทุกครั้ง สิ่งแรกที่นักสถิติจะต้องพิจารณาก็คือ กรอบตัวอย่างจะหามาได้โดยง่ายหรือไม่? มีความถูกต้องในเชิงสถิติหรือไม่? และค่าใช้จ่ายในการเตรียมกรอบสิ่งตัวอย่างและการเตรียมสลากในการชักสิ่งตัวอย่างจะเป็นเท่าไร

ประการที่สอง สิ่งตัวอย่างที่ได้จากแผนการชักสิ่งตัวอย่างแบบธรรมดาใช้ได้เฉพาะการสำรวจโครงการขนาดเล็กในวงการแคบ ๆ เท่านั้น หากเป็นโครงการใหญ่ระดับชาติ ตัวอย่างที่สุ่มมาได้ (เช่น ครัวเรือนตัวอย่าง) จะมีความกระจัดกระจายกันมาก ซึ่งมองในแง่ประสิทธิภาพของสิ่งตัวอย่างในทางทฤษฎีแล้วอาจจะเป็นลักษณะที่ดี เพราะตัวอย่างยิ่งกระจายกันมากเท่าไร ก็ยิ่งเป็นตัวแทนที่ดีของประชากรเท่านั้น แต่มองจากแง่ของผู้ปฏิบัติ การออกไปเก็บข้อมูลจากสมาชิกของประชากรที่กระจัดกระจายกันอยู่ตามที่ต่าง ๆ จะสิ้นเปลืองเวลาในการเดินทางและค่าพาหนะมาก เกินวิสัยที่จะปฏิบัติ เช่น หากจะเก็บข้อมูลจากครัวเรือน 100,000 แห่ง ครัวเรือนดังกล่าวอาจจะกระจายกันอยู่ในทุกหมู่บ้านรวม 50,000 หมู่บ้าน เวลาที่เสียไป ในการเดินทางของพนักงานในการเคลื่อนย้ายจากหมู่บ้านหนึ่งไปยังอีกหมู่บ้านหนึ่ง จึงอาจจะสูงถึงร้อยละ 80 - 90 ของเวลาปฏิบัติงานในสนามทั้งหมด

1.3 การชักสิ่งตัวอย่างแบบกลุ่ม (cluster sampling)

ในทางปฏิบัติจริง ๆ แล้ว จึงมักไม่ใช้สิ่งตัวอย่างที่เลือกมาได้ตามแผนการชักสิ่งตัวอย่างแบบธรรมดา แต่นักสถิติจะหาแผนแบบการชักสิ่งตัวอย่างที่ไม่ต้องยุ่งยากและประหยัดค่าใช้จ่ายในการเตรียมกรอบตัวอย่าง ขณะเดียวกันแผนก็ต้องอำนวยความสะดวกในขั้นปฏิบัติงานเก็บข้อมูลด้วย

มาตรการที่ช่วยแก้ปัญหาทั้งสองข้อดังกล่าวได้ ก็โดยการเลือกหน่วยสิ่งตัวอย่าง (sampling unit) ที่เหมาะสม เท่าที่ได้พิจารณามาแล้ว หน่วยสิ่งตัวอย่างกับสมาชิกของประชากรเป็นสิ่งเดียวกัน

แต่ในความเป็นจริงแล้วไม่จำเป็นต้องเป็นเช่นนั้นเสมอไป เราอาจจะจัดสมาชิกของประชากรหลายหน่วยมารวมกันเป็นหนึ่งหน่วยสิ่งตัวอย่างก็ได้ เช่น ในการสุ่มครัวเรือน แทนที่จะใช้ครัวเรือนเป็นหน่วยสิ่งตัวอย่าง เราอาจจะใช้หมู่บ้าน (ซึ่งก็คือกลุ่มของครัวเรือน) เป็นหน่วยสิ่งตัวอย่าง เมื่อเลือกได้หมู่บ้านใดเป็นหมู่บ้านตัวอย่างแล้ว ก็จะถือว่าครัวเรือนทุกครัวเรือนในหมู่บ้านเป็นครัวเรือนตัวอย่างด้วย ดังนั้น หากต้องการตัวอย่างขนาด 100,000 ครัวเรือน เราอาจจะสุ่มหมู่บ้านตัวอย่างประมาณ 830 หมู่บ้าน จากจำนวน 50,000 หมู่บ้าน เพราะเป็นที่ทราบกันว่าหมู่บ้านตัวอย่างโดยเฉลี่ยจะมีประมาณ 120 ครัวเรือน วิธีการที่ใช้หมู่บ้านเป็นหน่วยสุ่มตัวอย่างเช่นนี้ กรอบสิ่งตัวอย่างคือบัญชีรายชื่อหมู่บ้านที่มีอยู่ทั่วประเทศ (มีประมาณ 50,000 หมู่บ้าน) ซึ่งจะหามาได้ง่ายจากกระทรวงมหาดไทย และไม่ต้องเสียค่าใช้จ่ายมากนัก หากมีการเปลี่ยนแปลงในการแบ่งเขตหมู่บ้านหรือมีหมู่บ้านใหม่เกิดขึ้นในสารบบ ก็ไม่เป็นการยากที่จะปรับแก้ไขกรอบตัวอย่างให้ถูกต้องได้อย่างทันท่วงที

แผนแบบการชักสิ่งตัวอย่างโดยใช้หน่วยสุ่มตัวอย่างเป็นกลุ่มของหน่วยแจงนับเช่นนี้ มีชื่อเรียกว่า ตัวอย่างแบบกลุ่ม (cluster sampling) และจำนวนหน่วยแจงนับในแต่ละกลุ่มเรียกว่า ขนาดของกลุ่ม (cluster size) แต่ละกลุ่มอาจจะมีจำนวนหน่วยแจงนับเท่ากันหรืออาจจะไม่เท่ากันก็ได้ เช่น กรณีสุ่มหมู่บ้าน จำนวนครัวเรือนในหมู่บ้านไม่เท่ากันในทุกหมู่บ้าน เราจะเห็นได้ต่อไปว่าตัวอย่างแบบกลุ่ม อำนวยความสะดวกและเพิ่มประสิทธิภาพในการทำงานของพนักงานสนาม เพราะหน่วยแจงนับรวมกันเป็นกลุ่มก้อน ไม่ต้องเสียเวลาในการเดินทางมากนัก เมื่อเข้าไปในหมู่บ้านครั้งหนึ่งก็จะแจงนับบ้านทุกบ้าน เสร็จแล้วจึงจะเดินทางไปยังหมู่บ้านถัดไป เวลาที่เสียไปเพราะเดินทางจึงเป็นส่วนน้อย เมื่อเทียบกับเวลาในการเก็บข้อมูล

ถึงแม้การชักสิ่งตัวอย่างแบบกลุ่มจะมีส่วนดีแต่ก็มีส่วนเสียอยู่ตรงที่ตัวอย่างแบบกลุ่มมักจะมีความเคลื่อนคลาดในการสุ่มตัวอย่างสูง เมื่อเทียบกับตัวอย่างแบบธรรมดา ซึ่งมีขนาดตัวอย่างเท่ากัน หากจำเป็นจะต้องลดความเคลื่อนคลาดลงก็จะต้องเพิ่มจำนวนกลุ่มตัวอย่าง ซึ่งก็หมายความว่า จะต้องเสียค่าใช้จ่ายเพิ่มขึ้น นักสถิติผู้กำหนดแผนแบบการชักสิ่งตัวอย่างจึงต้องพิจารณาดูว่าความสะดวกหรือค่าใช้จ่ายที่ทุ่มไปเพราะใช้ตัวอย่างแบบเป็นกลุ่ม จะคุ้มกับความคลาดเคลื่อนที่เพิ่มขึ้นหรือไม่ นอกจากนั้น นักสถิติยังอาจจะพิจารณาเลือกขนาดของกลุ่มได้ตามความเหมาะสม (เช่น แทนที่จะใช้ทั้งหมู่บ้านเป็นกลุ่มๆ ละ 120 ครัวเรือน อาจจะใช้เพียงกลุ่มละ 40 ครัวเรือน ซึ่งการจัดกลุ่มอาจจะทำได้โดยเส้นแบ่งเขตทางภูมิศาสตร์ตามแผนที่) ขนาดของกลุ่มที่ไม่เท่ากันจะมีผลต่อการจัดงานสนาม ทำให้ค่าใช้จ่ายแตกต่างกัน และขณะเดียวกันความเคลื่อนคลาดของข้อมูลก็จะแตกต่างกันด้วย

1.4 การแบ่งชั้นภูมิก่อนชักสิ่งตัวอย่าง (Stratification in Sampling)

ในบางครั้งก่อนที่จะดำเนินการชักสิ่งตัวอย่างจากประชากร นักสถิติจำเป็นจะต้องพิจารณาแบ่งประชากรออกเป็นหลายๆ ส่วน เรียกว่า ชั้นภูมิ (stratum) หลาย ๆ ชั้นภูมิ แล้วจึงแยกชักสิ่งตัวอย่างออกมาจากแต่ละชั้นภูมิ โดยกำหนดขนาดของสิ่งตัวอย่างจากแต่ละชั้นภูมิหรือกำหนดอัตราการชักสิ่งตัวอย่างตามความเหมาะสม การชักสิ่งตัวอย่างแยกเป็นชั้นภูมิเช่นนี้ อาจจะทำให้สิ่งตัวอย่างกระจายกันดีกว่าการชักสิ่งตัวอย่างแบบธรรมดา เพราะสามารถแน่ใจได้ว่ามีตัวแทนจากทุกชั้นภูมิ สิ่งตัวอย่างที่ได้จึงมีประสิทธิภาพสูงกว่า ชั้นภูมิที่ต่างกันอาจจะมีความสำคัญไม่เท่ากันต่อยอดข้อมูลสถิติที่จะผลิตได้ อัตราการชักสิ่งตัวอย่างในชั้นภูมิที่สำคัญ อาจจะกำหนดให้สูงกว่าอัตราการสุ่มในชั้นภูมิที่มีความสำคัญน้อยลงไป เช่น หากจะสุ่มครัวเรือน 100,000 ครัวเรือนจากทั่วประเทศไทย อาจจะแบ่งประชากรออกเป็น 4 ชั้นภูมิตามภาคเหนือ ใต้ ตะวันออก และกลาง และภายในแต่ละภาคอาจจะแบ่งออกเป็นอีก 2 ชั้นภูมิ คือ ในเขตเทศบาล และนอกเขตเทศบาล รวมทั้งสิ้นเป็น 8 ชั้นภูมิ หรือนักสถิติอาจจะเห็นว่ากรุงเทพมหานครเป็นเขตสำคัญอาจจะจัดให้เป็นชั้นภูมิต่างหากก็ได้ อัตราการสุ่มตัวอย่างอาจจะให้แตกต่างกันได้ เช่น การต้องการจะเก็บสถิติเรื่องรายได้ของครัวเรือนก็อาจจะใช้อัตราการสุ่มในชั้นภูมิที่เป็นเขตเทศบาลที่สูงกว่าอัตรานอกเขตเทศบาล เป็นต้น การรู้จักใช้อัตราการสุ่มที่แตกต่างกันตามความสำคัญของชั้นภูมินี้ อาจจะเพิ่มประสิทธิภาพของตัวอย่างได้

การสุ่มแบบชั้นภูมิยังจะให้ประโยชน์อย่างอื่น เช่น อาจจะแสดงข้อมูลแยกเป็นรายชั้นภูมิได้ยิ่งการแบ่งชั้นภูมิทำตามเขตภูมิศาสตร์ด้วยแล้ว ก็อาจจะแสดงข้อมูลแยกเป็นรายเขตภูมิศาสตร์ได้โดยง่าย นักสถิติอาจจะเลือกใช้แผนแบบการสุ่มที่แตกต่างกันภายในชั้นภูมิต่างๆ ตามความเหมาะสมของสภาพในชั้นภูมินั้น ๆ เช่น สภาพทางภูมิศาสตร์ในเขตเทศบาลแตกต่างกับนอกเขตเทศบาลมาก อาจจะใช้กรอบตัวอย่างที่ไม่ได้มาจากแหล่งเดียวกันหรือใช้หน่วยสุ่มตัวอย่างที่แตกต่างกันได้ นอกจากนั้นการแบ่งจักรวาลออกเป็นชั้นภูมิ ยังจะอำนวยความสะดวกในการบริหารการเก็บข้อมูล และการประมวลข้อมูล เช่น อาจจะใช้ชั้นภูมิเป็นเขตควบคุมการปฏิบัติงานสนามพร้อมกันไป

การพิจารณาเกณฑ์ในการแบ่งชั้นภูมิ การกำหนดอัตราการสุ่มตัวอย่างในแต่ละชั้นภูมิจะต้องพิจารณาควบคู่กันไปกับแผนแบบการสุ่มตัวอย่างที่จะใช้ในแต่ละชั้นภูมิ เหล่านี้เป็นหน้าที่ของนักสถิติผู้วางแผนการสำรวจจะต้องทำทุกครั้ง

1.5 ข้อควรระวังในการชักสิ่งตัวอย่าง

ในการวางแผนการชักสิ่งตัวอย่าง นักสถิติจะต้องตระหนักในความสำคัญของคุณภาพของข้อมูลและจะต้องวิเคราะห์ / คำนวณว่า ข้อมูลที่จะประมาณได้จากสิ่งตัวอย่างมีความเคลื่อนคลาดและมีความเอนเอียงเป็นเท่าไร หากไม่ใช้ตัวแบบ probability sample แล้ว จะไม่สามารถคำนวณปริมาณดังกล่าวได้เลย ปกติแล้วการรวบรวมข้อมูลสถิติจากสิ่งตัวอย่าง จะมีความเคลื่อนคลาดในข้อมูลทุกครั้ง ความเคลื่อนคลาดของข้อมูลรายการใด ขึ้นอยู่กับความแตกต่าง (variation) ของข้อมูลเบื้องต้นระหว่างหน่วยแจงนับหนึ่ง ๆ ปริมาณความแตกต่างนี้เราวัดได้ด้วยค่าความแปรปรวนของข้อมูลในประชากร หากมีความแตกต่างกันมาก ค่าความแปรปรวนสูง ความเคลื่อนคลาดจะสูง แต่จะทำให้ ลดลงได้โดยเพิ่มขนาดของตัวอย่าง และหากข้อมูลมีการจำแนกรายละเอียดมากขึ้น ข้อมูลนั้น ๆ ซึ่งเป็นยอดแต่ละยอดในตารางสถิติก็จะมีความเคลื่อนคลาดเพิ่มขึ้นด้วย ความเอนเอียงจากสิ่งตัวอย่างไม่จำเป็นจะต้องมีเสมอไปทุกครั้ง ความเอนเอียงจากสิ่งตัวอย่างเกิดจากสองแฟกเตอร์ที่สำคัญ คือ

(1) การใช้สูตรในการประมาณที่มีความเอนเอียง ในกรณีนี้จะสามารถคำนวณได้ว่าองศาแห่งความเอนเอียงเป็นเท่าไร

(2) เกิดจากการที่สิ่งตัวอย่างไม่ได้เป็นตัวแทนที่ดีของประชากร ซึ่งเป็นเพราะ

2.1 ไม่ได้ใช้วิธีการสุ่มตามทฤษฎีความน่าจะเป็น

2.2 กรอบตัวอย่างไม่สมบูรณ์

2.3 มีอัตราการไม่ตอบสูง ข้อมูลที่ได้มาจากผู้ตอบ ไม่ได้เป็นข้อมูลที่เป็นตัวแทนของประชากร

การจะเพิ่มขนาดสิ่งตัวอย่างหรือการเพิ่มรายละเอียดในการจัดจำแนกข้อมูลอาจจะมีผลหรือไม่มีผลเพิ่ม หรือลดความเอนเอียง

ตัวอย่างคลาสสิคที่แสดงความเอนเอียงในข้อมูล ซึ่งเกิดจากความบกพร่องของสิ่งตัวอย่าง ก็คือการสำรวจเพื่อทำนายผลการเลือกตั้งประธานาธิบดีในสหรัฐฯ เมื่อปี 1936 ซึ่งจัดทำโดยนิตยสาร Literary Digest ขนาดของประชากร (ผู้มีสิทธิออกเสียง) มีประมาณ 70,000,000 หน่วย ชักสิ่งตัวอย่างออกมา 10,000,000 หน่วย โดยให้บุคคลทุกคนที่มีนามปรากฏในสมุดโทรศัพท์เป็นหน่วยแจงนับตัวอย่าง ในจำนวน 10 ล้านคนที่สอบถามไป ได้คำตอบกลับมาประมาณ 2,000,000 คน และในจำนวนนี้ร้อยละ 60 จะลงคะแนนให้ นาย Landon ส่วนอีกร้อยละ 40 จะลงคะแนนให้ นาย Roosevelt Literary Digest ได้ทำนายผลการเลือกตั้งว่า นาย Landon ซึ่งสังกัดพรรค Republican จะได้ชัยชนะแบบแผ่นดินจะถล่ม (land-slide victory) แต่ผลกลับปรากฏเป็นตรงกันข้าม นาย Roosevelt กลับชนะอย่างแผ่นดินถล่ม เหตุที่เป็นเช่นนี้เพราะแผนแบบการชักสิ่งตัวอย่างไม่ได้เป็นไปตามทฤษฎีความน่าจะเป็น

การชักสิ่งตัวอย่างเฉพาะแต่ผู้มีโทรศัพท์มุ่งแต่จะให้ได้ความสะดวกในการชักสิ่งตัวอย่างเป็นสำคัญ แต่ปรากฏ (ในภายหลัง) ว่า ผู้มีโทรศัพท์เป็นกลุ่มคนรวยซึ่งนิยมพรรค Republican ไม่ได้เป็นตัวแทนที่ดีของผู้มีสิทธิออกเสียงทั้งหมด ยิ่งไปกว่านั้นผู้ตอบเพียงร้อยละ 20 ที่ส่งคำตอบกลับมาก็ยังไม่สามารถบอกได้ว่าเป็นตัวแทนที่ดีของคน 10,000,000 คนที่สุ่มออกมาหรือไม่ ข้อมูลที่เอามาทำนายผลการเลือกตั้งจึงมีความเอนเอียงอย่างใช้ประโยชน์ไม่ได้ (กรณีนี้ทำให้เกิดความเสียหายแก่นิตยสาร Literary Digest ถึงกับต้องเลิกกิจการไปในที่สุด) แต่นั่นเป็นเหตุการณ์ที่เกิดขึ้นเกือบ 40 ปีมาแล้ว ในขณะนั้นนักสถิติยังไม่สู้จะเข้าใจระเบียบวิธีการสุ่มตัวอย่างดีนักและยังไม่ตระหนักในอันตรายของการใช้ข้อมูลที่มีความเอนเอียง ในปัจจุบันการทำนายผลการเลือกตั้งโดยใช้ตัวอย่างผู้ตอบเพียง 4,000 - 5,000 ราย (สุ่มออกมาได้ตามทฤษฎีความน่าจะเป็น) ก็สามารถทำนายผลได้อย่างแม่นยำ โดยมีความเคลื่อนคลาดไม่เกิน 1 เปอร์เซ็นต์ โดยไม่มีความเอนเอียงจากการใช้ตัวอย่างผิด ๆ แต่อาจจะมีความเอนเอียงบ้างจากแหล่งอื่น เช่น เอนเอียงจากผู้ตอบ (response bias)

1.6 <u>ศึกษาการชักสิ่งตัวอย่างกรณีประชากรมีจำนวนมากเป็นอนันต์</u>

เมื่อต้องทำงานกับปัญหาที่ประชากรมีจำนวนมากเป็นอนันต์ วิธีการชักสิ่งตัวอย่างก็จะแตกต่างจากวิธีที่เคยเรียนมาบ้าง ทั้งนี้เพราะในเชิงฟิสิกส์แล้วเราไม่สามารถที่จะติดหมายเลขให้กับสมาชิกของประชากรทุกตัวได้ ดังนั้นการเลือกสิ่งตัวอย่างเชิงสุ่มจึงกระทำในทำนองที่ดูว่ามีความเอนเอียงน้อยที่สุด ตัวอย่างเช่นการเลือกสิ่งตัวอย่างจากรายการผลิต (อาจจะเป็นหลอดไฟก็ได้) เพื่อมาตรวจสอบว่าผลิตภัณฑ์ดังกล่าวได้มาตราฐานหรือไม่? อาจจะดำเนินการได้โดยการหยิบสิ่งตัวอย่างจากรายการผลิตในทุก ๆ ครึ่งชั่วโมงเป็นต้น อีกตัวอย่างหนึ่งก็คือน้ำเสียที่ปล่อยออกมาจากโรงงานต่าง ๆ การที่จะตรวจสอบว่าน้ำเสียที่ถูกปล่อยออกมานั้นมีคุณภาพมาตรฐานหรือไม่? วิธีการตรวจสอบนั้นจะทำได้โดยการชักสิ่งตัวอย่างซึ่งอาจจะเป็นการไปตักน้ำเสียดังกล่าวมาทดสอบทุก ๆ อาทิตย์เป็นต้น ในเชิงคณิตศาสตร์นิยามของ random sample กรณี infinite population จะเป็นดังนี้คือ

Random sample (infinite population)	A set of observation $X_1, X_2, \ldots, X_n$ constitutes a random sample of size n from the infinite population f(x) if : 1. Each x_i is a value of a random variable whose distribution is given by f(x) 2. These n random variables are independent We also apply the term random sample to the random variables are $X_1, \ldots, X_n$.

1.7 Sampling distribution

พิจารณาสิ่งตัวอย่างขนาด n ทั้งหมดที่เลือกมาจากประชากรกลุ่มหนึ่งที่กำหนดให้ (เลือกแบบคืนที่ในกรณีที่ประชากรมีจำนวนจำกัด) สำหรับสิ่งตัวอย่างแต่ละชุดที่เลือกมาได้ เราสามารถคำนวณหาค่า statistic เช่น mean หรือ standard deviation หรืออื่น ๆ ซึ่งค่าเหล่านี้จะแตกต่างกันไปในแต่ละสิ่งตัวอย่าง ด้วยวิธีการเช่นนี้เราจะได้การกระจายของค่า statistic ซึ่งจะเรียกว่า sampling distribution

ตัวอย่างเช่น สมมุติว่า statistic ที่เราสนใจคือ sample mean การกระจายของ mean ก็จะมีชื่อเรียกว่า sampling distribution of means เป็นต้น ทำนองเดียวกันนักศึกษาคงจะให้คำจำกัดความของคำเหล่านี้ได้ด้วยตนเองคือ

- sampling distribution of standard deviation
- sampling distribution of variance
- sampling distribution of medians
- sampling distribution of proportions.

ตัวอย่าง กำหนดให้ประชากรประกอบไปด้วยจำนวน 5 จำนวน คือ 2, 3, 6, 8, 11 พิจารณาการชักสิ่งตัวอย่างขนาด 2 ที่เป็นไปได้ทั้งหมดจากประชากรนี้ (ชักตัวอย่างแบบคืนที่)

(ก) จงหา mean ของประชากร

(ข) จงหา standard deviation ของประชากร

(ค) จงหา mean of sampling distribution of menas

(ง) จงหา standard deviation ของ sampling distribution of means นั่นคือจงหา standard error of means

วิธีทำ (ก) $$\mu = \frac{2 + 3 + 6 + 8 + 11}{5} = \frac{30}{5} = 6.0$$

(ข) $$\sigma^2 = \frac{(2-6)^2 + (3-6)^2 + (6-6)^2 + (8-6)^2 + (11-6)^2}{5}$$

$$= \frac{16 + 9 + 0 + 4 + 25}{5} = 10.8$$

และ

$$\sigma = \sqrt{10.8} = 3.29$$

(ค) มีสิ่งตัวอย่างขนาด 2 ที่เป็นได้ทั้งหมด คือ 5(5) = 25 (ชักสิ่งตัวอย่างแบบคืนที่) สิ่งตัวอย่างเหล่านั้นประกอบด้วย

(2, 2)	(2, 3)	(2, 6)	(2, 8)	(2, 11)
(3, 2)	(3, 3)	(3, 6)	(3, 8)	(3, 11)
(6, 2)	(6, 3)	(6, 6)	(6, 8)	(6, 11)
(8, 2)	(8, 3)	(8, 6)	(8, 8)	(8, 11)
(11, 2)	(11, 3)	(11, 6)	(11, 8)	(11, 11)

sample mean ที่สมนัยกับสิ่งตัวอย่างข้างต้นคือ

(1)

2.0	2.5	4.0	5.0	6.5
2.5	3.0	4.5	5.5	7.0
4.0	4.5	6.0	7.0	8.5
5.0	5.5	7.0	8.0	9.5
6.5	7.0	8.5	9.5	11.0

ดังนั้น mean of sampling distribution of means คือ

$$\mu_{\bar{X}} = \frac{\text{ผลบวกของ sample means ใน (1) ข้างต้น}}{25}$$

$$= \frac{150}{25} = 6.0$$

ขอให้สังเกตว่า $\mu_{\bar{X}} = \mu$

(ฅ)
$$\sigma^2_{\bar{X}} = \frac{(2.0-6.0)^2 + (2.5-6.0)^2 + \ldots (11.0-6.0)^2}{25}$$

$$= 135/25 = 5.40$$

ข้อให้สังเกตว่า $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$ (ทั้งนี้เพราะ $\frac{\sigma^2}{n} = \frac{10.8}{2} = 5.40$)

ดังนั้น

Standard error of means $\sqrt{\sigma^2_{\bar{X}}} = \sqrt{5.40} = 2.32$

ข้อสังเกตที่ได้เขียนไว้ข้างต้นนี้ไม่ใช่เรื่องบังเอิญ เป็นเรื่องที่สามารถพิสูจน์เป็นทฤษฎีบทดังนี้

<u>ทฤษฎีบท 1</u> ให้ $X_1, X_2, ..., X_n$ เป็น random sample ขนาด n ซึ่งถูกเลือกมาจากประชากรกลุ่มหนึ่ง ซึ่งการแจกแจงของประชากรดังกล่าวมี mean = μ และมี variance = σ^2 [ขอให้ระลึกว่าคำว่า "random samyle" หมายความว่า (1) $X_1, X_2, ..., X_n$ ถูกเลือกมาจากประชากรที่มีการแจกแจงเหมือนกัน และ (2) $X_1, X_2, ..., X_n$ ต่างเป็นอิสระต่อกัน] ต่อไปให้ $\bar{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n)$

จะได้ว่า

(ก) ถ้าประชากรที่กำหนดให้มีการแจกแจงแบบปรกติ $N(\mu, \sigma^2)$ แล้ว $\bar{x}$ จะมีการแจกแจงเป็นแบบปรกติ $N(\mu, \frac{\sigma^2}{n})$ (ด้วยค่า n เท่าใดก็ได้)

(ข) ถ้าประชากรที่กำหนดให้มีการแจกแจงเป็นแบบอื่น ๆ ที่ไม่ใช่ปรกติ (เช่น uniform, exponential, และอื่น ๆ) แล้วการแจกแจงของ $\bar{x}$ จะเข้าใกล้การแจกแจงแบบปรกติ $N(\mu, \frac{\sigma^2}{n})$ เมื่อ n มีค่ามากพอ (โดยทั่วไปแล้วจะถือว่า $n \geq 30$ คือค่าที่มากพอ)

<u>พิสูจน์</u> (ก) เป็นผลจากทฤษฎีบทในวิชา mathematical statistics ว่าผลบวกเชิงเส้นของตัวแปรสุ่มที่มีการแจกแจงแบบปรกติก็ยังคงมีการแจกแจงแบบปรกติอยู่ โดยที่ถ้า $X_1, X_2, ..., X_n$ มีการแจกแจงแบบ $N(\mu, \sigma^2)$ แล้ว $\bar{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n)$ จะมีการแจกแจงแบบ $N(\mu, \frac{\sigma^2}{n})$ เมื่อ n มีค่า = 2, 3, 4, . . . เป็นต้นไป

(ข) สำหรับกรณีที่ random sample ไม่ได้ถูกเลือกมาจากประชากรที่มีการแจกแจงแบบปรกติ ในกรณีเช่นนี้การแจกแจงของ $\bar{X}$ อาจจะไม่อยู่ในแบบปรกติก็ได้ แต่อย่างไรก็ดี The central limit theorem ก็ได้บอกแก่เราว่าการแจกแจงของ $\bar{X}$ ก็ยังคงเข้าใกล้การแจกแจงแบบปรกติ $N(\mu, \frac{\sigma^2}{n})$ เมื่อ n มีค่ามากพอสำหรับค่า n ที่ถือว่ามากพอนั้นจะใช้ $n \geq 30$ [เนื้อความของ Central limit Theorem เป็นดังนี้คือ :

If $\bar{X}$ is the mean of a random sample $X_1, X_2, ..., X_n$ from a distribution with mean μ and finite variance $\sigma^2 > 0$ then the distribution of $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ approach a distribution that is $N(0, 1)$ as $n \to \infty$]

ข้อสังเกต (ก) จากทฤษฎีบท 1 ทำให้เราสรุปได้ว่า mean และ standard deviation ของ sampling distribution of means คือ

(2)
$$\mu_{\bar{X}} = E(\bar{X}) = \mu$$
$$\sigma_{\bar{X}} = Var(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

ถ้า n มีค่ามากพอ โดยทั่วไปแล้วจะถือว่า $n \geq 30$ คือค่าที่มากพอ

(ข) จะเรียก standard deviation ของ sampling distribution ของตัว statistic ตัวหนึ่ง ๆ ว่า standard error เพราะฉะนั้นขณะนี้เรามี standard error อยู่ 1 ตัวแล้วคือ

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

ตัวอย่างที่ 1 สมมุติว่าความสูงของนักเรียน 3,000 ในโรงเรียนแห่งหนึ่งมีการแจกแจงแบบปรกติด้วย mean 68.0 นิ้ว และ satndard deviation 3.0 นิ้ว ถ้ามีการเลือกสิ่งตัวอย่างนักเรียนขนาด 25 คนมา 80 ตัวอย่าง จงหา mean และ standard deviation ของ sampling distribution of means ถ้าการเลือกตัวอย่างเป็นแบบคืนที่

วิธีทำ จำนวนสิ่งตัวอย่างที่เป็นไปทั้งหมดคือ $(3,000)^{25}$ ซึ่งมีค่ามากกว่า 80 มาก ดังนั้นจากการทดลองเลือกมา 80 ตัวอย่าง เราคงจะสรุปจากการทดลองได้ลำบาก แต่จากทฤษฎีเราพอจะสรุปได้ว่า การแจกแจงค่าเฉลี่ยของการชักสิ่งตัวอย่าง จะเข้าใกล้การแจกแจงแบบปรกติด้วย

$$\mu_{\bar{X}} = \mu = 68.0 \text{ นิ้ว}$$
$$\sigma_{\bar{X}} = \sigma/\sqrt{n} = 3.0/\sqrt{25} = 0.6 \text{ นิ้ว}$$

ตัวอย่างที่ 2 มีสิ่งตัวอย่างในตัวอย่างที่ 1 กี่สิ่งตัวอย่างที่ผู้อ่านคาดว่าจะมีค่าเฉลี่ย

(ก) อยู่ระหว่าง 66.8 และ 68.3 นิ้ว

(ข) น้อยกว่า 66.4 นิ้ว

วิธีทำ ค่าของ mean ของ sample ในหน่วยมาตราฐานถูกกำหนดด้วยสูตร

$$z = \frac{\bar{x} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{x} - 68.0}{0.6}$$

ศูนย์บรรณสารและสื่อการศึกษา
มหาวิทยาลัยเทคโนโลยีสุรนารี

(ก) 66.8 ในหน่วยมาตราฐานคือ (66.8 - 68.0) / 0.6 = -2.0

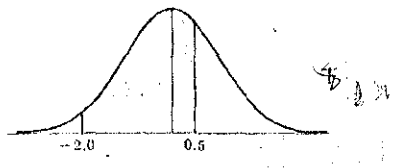
68.3 ในหน่วยมาตราฐานคือ (68.3 - 68.0) / 0.6 = 0.5

ความน่าจะเป็นที่สิ่งตัวอย่างจะมีค่าเฉลี่ยระหว่าง 66.8 และ 68.3 นิ้ว

= (พื้นที่ภายใต้เส้นโค้งปรกติระหว่าง z = -2.0 และ z = 0.5)

= (พื้นที่ระหว่าง z = -2 และ z = 0) + (พื้นที่ระหว่าง z = 0 และ z = 0.5)

= 0.4772 + 0.1915 = 0.6687



ดังนั้นจำนวนของตัวอย่างตามความคาดหมาย = (80)(0.6687) หรือ 53.

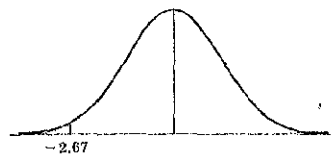
(ข) 66.4 ในหน่วยมาตราฐาน = (66.4 - 68.0) / 0.6 = -2.67

ความน่าจะเป็นที่สิ่งตัวอย่างจะมีค่าเฉลี่ยน้อยกว่า 66.4 นิ้ว

= (พ.ท. ภายใต้เส้นโค้งปรกติจากทางซ้ายถึง z = -2.67)

= (พ.ท. จากทางซ้ายถึง z = 0) - (พ.ท. ระหว่าง z = -2.67 และ z = 0)

= 0.5 - 0.4962 = 0.0038



ดังนั้นจำนวนของตัวอย่างตามความคาดหมาย = (80)(0.0038) = 0.304

1.8 Sampling Distribution of Proportions

ขอให้พิจารณาประชากรกลุ่มหนึ่งซึ่งมีการแจกแจงแบบทวินามและให้

p = ความน่าจะเป็นที่เหตุการณ์หนึ่งจะเกิดขึ้น (success)

q = ความน่าจะเป็นที่เหตุการณ์ดังกล่าวจะไม่เกิดขึ้น (failure)

= 1 - p

ต่อไปให้พิจารณาสิ่งตัวอย่างขนาด n ทั้งหมดซึ่งถูกเลือกมาจากประชากรกลุ่มนี้ และสำหรับสิ่งตัวอย่างขนาด n แต่ละชุดดังกล่าวให้คำนวณหาอัตราส่วน $P = \frac{X}{n}$ ของ success. [ตัวอย่างเช่นถ้าเราโยนเหรียญเที่ยงตรง 1 อันไปเรื่อย ๆ และให้ประชากรของเราคือบรรดา outcome ที่เกิดจากการโยนเหรียญดังกล่าว ในกรณีที่มีการเลือกตัวอย่างขนาด n มาจากประชากรกลุ่มนี้ P ก็จะเป็นอัตราส่วนที่จะเกิด success ของการโยนเหรียญ n ครั้ง] ทำเช่นนี้ต่อไปกับทุก ๆ ชุดของสิ่งตัวอย่างที่เลือกมาได้ เราก็จะได้ sampling distribution of proportion ซึ่ง mean μ_p และ standard deviation σ_p หาได้ดังนี้

$$\mu_p = E(P) = E(\frac{X}{n}) = \frac{1}{n}E(x) = \frac{1}{n} \cdot np = p$$

$$\sigma_p^2 = Var(P) = Var\ \frac{X}{n} = \frac{1}{n^2}Var(X) = \frac{1}{n^2} \cdot npq = \frac{pq}{n}$$

ดังนั้น $\sigma_p = \sqrt{\frac{pq}{n}}$

จึงสรุปได้ว่า mean μ_p และ standard deviation σ_p ของ sampling distribution of proportions คือ

(3)
$$\boxed{\begin{aligned} \mu_p &= p \\ \sigma_p &= \sqrt{\frac{pq}{n}} \end{aligned}}$$

เช่นเดียวกันกับในกรณีของ mean สูตรนี้จะใช้ได้กับกรณีที่ประชากรมีมากเป็นอนันต์หรือประชากรมีจำนวนจำกัดแต่การชักสิ่งตัวอย่างเป็นแบบคืนที่ เนื่องจากประชากรมีการแจกแจงแบบทวินาม (ซึ่งไม่ใช่การแจกแจงแบบปรกติ) ดังนั้นการแจกแจงของ sampling distribution of proportion จึงอาจจะไม่เป็นการแจกแจงแบบปรกติก็ได้ แต่โดย Central Limit Theorem เราทราบว่าการแจกแจงของ sampling distribution of proportion จะเข้าใกล้ $N(\mu_p, \sigma_p)$ ถ้า n มีค่ามากพอนั่นคือ $n \geq 30$

ประโยชน์อันดับแรกของการศึกษาในเรื่องราวข้างต้นก็คือการหาความน่าจะเป็นของ mean และ proportion ดังตัวอย่างต่อไปนี้

<u>ตัวอย่างที่ 3</u> ได้เป็นที่ทราบกันแล้วว่าเครื่องจักรอันหนึ่งจะผลิตสินค้าที่ไม่ได้มาตรฐานประมาณ 2% ได้มีการชักสิ่งตัวอย่างผลิตภัณฑ์ของเครื่องจักรชิ้นนี้มา 400 ชิ้น

(ก) จงหาความน่าจะเป็นที่ว่าจะมีผลิตภัณฑ์ไม่ได้มาตราฐานปนมาด้วย 3% หรือมากกว่า

(ข) จงหาความน่าจะเป็นที่ว่าจะมีผลิตภัณฑ์ไม่ได้มาตราฐานปนมาด้วย 2% หรือน้อยกว่า

<u>วิธีทำ</u> : ในที่นี้

$$\mu_p = p = 0.02 \quad \text{และ} \quad \sigma_p = \sqrt{pq/n} = \sqrt{0.02(0.98)/400}$$

$$= 0.14 / 20 = 0.007$$

(ก) ใช้ตัวปรับสำหรับกรณีตัวแปรเต็มหน่วย $1/2n = 1/800 = 0.00125$ เรามี

$$(0.03 - 0.00125) \text{ ในหน่วยมาตราฐาน} = \frac{0.03 - 0.00125 - 0.02}{0.007}$$

$$= 1.25$$

ความน่าจะเป็นที่ต้องการ = (พื้นที่ภายใต้เส้นโค้งปรกติไปทางขวาของ $z = 1.25$)

$$= 0.1056$$

ถ้าไม่ใช่ตัวปรับจะได้ค่าความน่าจะเป็นเท่ากับ 0.0764

(ข) $(0.02 + 0.00125)$ ในหน่วยมาตราฐาน $= \frac{0.02 + 0.00125 - 0.02}{0.007}$

$$= 0.18$$

ความน่าจะเป็นที่ต้องการ = (พื้นที่ภายใต้เส้นโค้งปรกติจากทางซ้ายคือ $z = 0.18$)

$$= 0.5000 + 0.0714 = 0.5714$$

ถ้าไม่ใช่ตัวปรับจะได้ค่าความน่าจะเป็นเท่ากับ 0.5000.

แบบฝึกหัด

1. Explain why the following will not lead to random sampling of the desired populations.
   (a) To determine what the average person spends on a vacation, a researcher interviews pressengers on a lnxury cruise.
   (b) To determine the average income of its graduates 10 years after graduation, the alumni office of a university sent questionnaires in 1994 to all members of the class of 1984 and bases its estimate on the questionaires returned.
   (c) To determine public sentiment about import restrictions, an interviewer asks voters : "Do you feel that this unfair practice should be stopped?"

2. The weights of 1500 ball bearing are normally distributed with mean of 22.40 ounces and a standard deviation of .048 ounces. If 300 random samples of size 36 are drawn from this population, determine the expected mean and standard deviation of the sampling distribution of means if the sampling is done with replacement

   Ans $\mu_{\bar{X}} = 22.40$ oz, $\sigma_{\bar{X}} = .008$ oz.

3. In problem 2, how many of the random samples would have their means
   (a) between 22.39 and 22.41 oz,
   (b) greater than 22.42 oz,
   (c) less than 22.37 oz,
   (d) less than 22.38 or more than 22.41 oz.

   Ans (a) 237, (b) 2, (c) none, (d) 34

4. Find the probability that next 200 children born
   (a) less than 40% will be boys.
   (b) between 43% and 57% will be girls
   (c) more than 54% will be boys. Assume equal probablilties for birth of boys ans girls

   Ans (a) .0019, (b) .9596 (c) .1151

2. การประมาณค่าพารามิเตอร์

การอนุมานเชิงสถิติ (statistical inference) เป็นกระบวนการใช้ข้อมูลจากสิ่งตัวอย่าง เพื่อหาข้อสรุปเกี่ยวกับข้อมูลของประชากร เทคนิคของการอนุมานเชิงสถิตินั้น สามารถแบ่งออกได้เป็นสองวิธีใหญ่ ๆ ดังนี้ คือ

(ก) <u>การประมาณค่าพารามิเตอร์</u> (parameter estimation)

วิธีการนี้จะเกี่ยวข้องกับการประมาณค่าของ parameter ด้วยค่าของ statistics ที่สมนัยกัน สำหรับการประมาณค่าของพารามิเตอร์นั้น ยังแบ่งออกเป็นวิธีย่อย ๆ ได้อีกดังนี้ คือ

(ก.1) การประมาณค่าแบบจุด (point estimation)

(ก.2) การประมาณค่าแบบช่วง (interval estimation)

(ข) <u>การทดสอบสมมุติฐาน</u> (hypothesis testing)

วิธีการนี้จะเกี่ยวข้องกับการทดสอบว่าค่าของประชากรที่เราได้คาดหมายไว้นั้นเชื่อถือได้หรือไม่? ด้วยความน่าจะเป็นเท่ากับเท่าไร? การศึกษาเรื่องนี้จะอยู่ในหัวข้อที่ 3 สำหรับหัวข้อนี้จะเน้นการอภิปรายไปที่วิธีการแรกโดยจะเริ่มต้นที่ (ก.1) ก่อน

2.1 การประมาณค่าแบบจุด (point estimation)

การประมาณค่าของพารามิเตอร์ ด้วยค่าของจำนวนจริงเพียงค่าเดียวนั้น จะเรียกว่าเป็น <u>การประมาณค่าแบบจุด</u> ส่วนการประมาณค่าของพารามิเตอร์โดยใช้ช่วงของจำนวนจริงนั้นจะเรียกว่าเป็น <u>การประมาณค่าแบบช่วง</u>

การประมาณค่าแบบช่วงนั้นโดยทั่วไปแล้วจะมีความถูกต้องมากกว่าจึงเป็นที่นิยมใช้ในการประมาณค่าของประชากรมากกว่าแบบจุดโดยส่วนใหญ่แล้วการประมาณค่าแบบจุดจะใช้ในการประมาณค่าของ population mean, หรือ population median โดยอาศัยข้อมูลจาก sample mean หรือ sample median เป็นต้น

ตัวอย่างของการประมาณค่าแบบจุดจะเป็นดังนี้คือ สมมุติว่าคะแนนสอบวิชา Prob. & Stat. ของนักศึกษา 500 คน มีการแจกแจงแบบปรกติด้วยค่าของ mean = μ และ variance σ^2 ที่ยังไม่ทราบค่า แต่ผู้ตรวจกระดาษคำตอบต้องการประมาณค่าของ μ และ σ^2 จึงได้ชักสิ่งตัวอย่างเชิงสุ่มออกมา 4 คน พบว่าคะแนนของนักศึกษา 4 คนเป็นดังนี้คือ $x_1 = 25$, $x_2 = 31$, $x_3 = 28$, $x_4 = 30$ (คะแนนเต็ม 100 คะแนน) ดังนั้น sample mean และ sample variance จะคำนวณได้ดังนี้คือ

$$\text{sample mean} = \hat{\mu} = \frac{25+31+28+30}{4} = 28.5$$

$$\text{sample variance} = \hat{\sigma}^2 = \frac{(25-28.5)^2+(31-28.5)^2+(28-28.5)^2+(30-28.5)^2}{4}$$

$$= 0.07$$

โดยการใช้ $\hat{\mu}$ ประมาณค่าของ μ และ $\hat{\sigma}^2$ ประมาณค่าของ σ^2 ดังนั้นจะสรุปได้ว่า μ และ σ^2 ของคะแนนสอบของนักเรียน 500 คน มีค่าเท่ากับ 28.5 และ 0.07 ตามลำดับ

ปัญหาในการประมาณค่ามักจะเกิดขึ้นบ่อยในงานเชิงวิศวกรรม มีบ่อยครั้งที่เรามีความจำเป็นจะต้องประมาณค่าของ

(1) ค่าเฉลี่ย σ^2 ของประชากรกลุ่มเดียว

(2) ความแปรปรวน σ^2 (หรือส่วนเบี่ยงเบนมาตราฐาน σ) ของประชากรกลุ่มเดียว

(3) อัตราส่วน p ของสมาชิกในประชากรซึ่งจะอยู่ในกลุ่มที่เราสนใจ

(4) ผลต่างของค่าเฉลี่ยสำหรับประชากรสองกลุ่ม ; $\mu_1 - \mu_2$

(5) ผลต่างของอัตราส่วนสำหรับประชากรสองกลุ่ม ; $p_1 - p_2$

ตัวประมาณค่าแบบจุดที่สมเหตุสมผลกับพารามิเตอร์ดังกล่าวข้างต้นเป็นดังนี้คือ

(1') สำหรับ μ, ตัวประมาณค่าคือ $\hat{\mu} = \bar{X}$ (ซึ่งก็คือ sample means)

(2') สำหรับ σ^2, ตัวประมาณค่าคือ $\hat{\sigma}^2 = S^2$ (ซึ่งก็คือ sample variance)

(3') สำหรับ p, ตัวประมาณค่าคือ $\hat{p} = X/n$ (ซึ่งก็คือ sample proportion)
(X คือจำนวนของ success ในสิ่งตัวอย่างขนาด n)

(4') สำหรับ $\mu_1 - \mu_2$, ตัวประมาณค่าคือ $\hat{\mu}_1 - \hat{\mu}_2 = \bar{X}_1 - \bar{X}_2$ (ซึ่งก็คือผลต่างระหว่าง sample means ของตัวแปรสุ่มอิสระสองตัว)

(5') สำหรับ $p_1 - p_2$, ตัวประมาณค่าคือ $\hat{p}_1 - \hat{p}_2$ (ซึ่งก็คือผลต่างระหว่าง sample proportions ของตัวแปรสุ่มอิสระสองตัว)

2.2 คุณสมบัติของตัวประมาณค่า (Properties of Estimators)

คุณสมบัติที่ตัวประมาณค่าควรจะมีก็คือ มันควรจะอยู่ "ใกล้" กับค่าจริงของตัวพารามิเตอร์ (ซึ่งเราไม่ทราบค่า) พอควร นักศึกษาก็คงจะมีคำถามต่อไปว่า คำว่า อยู่ "ใกล้" นี้จะให้มีความหมายว่าอย่างไร คำตอบที่เป็นไปได้ทางหนึ่งก็คือเราอยากจะกล่าวว่า "โดยเฉลี่ย" แล้วค่าของมันควรจะเท่ากับตัวพารามิเตอร์ที่เราต้องการประมาณค่า นั่นคือ เราจะมาถึงบทนิยามต่อไปนี้

บทนิยาม จะเรียกตัวประมาณค่า $\hat{\theta}$ ว่าเป็น<u>ตัวประมาณค่าที่ไม่เอนเอียง</u> (unbiased estimator) ของพารามิเตอร์ θ ถ้า

$$E(\hat{\theta}) = \theta$$

ขอให้สังเกตว่าเงื่อนไขนี้สมมูลกับการกำหนดให้ mean ของ sampling distribution ของ $\hat{\theta}$ เท่ากับ θ

ตัวอย่าง สมมุติว่า X เป็นตัวแปรสุ่มที่มี mean = μ และมี variance = σ^2 ให้ $X_1, X_2, \ldots, X_n$ เป็น random sample ขนาด n ที่เลือกมาจาก X จงแสดงว่า

(ก) Sample mean $\bar{X}$ เป็น unbiased estimator ของ μ

(ข) Sample variance S^2 เป็น biased estimator ของ σ^2

วิธีทำ

(ก) พิจารณา

$$E(\bar{X}) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)$$

$$= \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} E(X_i)$$

และเนื่องจาก $E(X_i) = \mu$ สำหรับทุก $i = 1, 2, \ldots, n$ ดังนั้น

$$E(\bar{X}) = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu$$

เพราะฉะนั้น sample mean $\bar{X}$ จึงเป็น unbiased estimator ของ population mean μ

(ข) จากการที่

(1) $$S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 / n$$

ดังนั้น

$$E(S^2) = E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2 / n\right]$$

$$= \frac{1}{n} E\left[\sum_{i=1}^{n}\left(X_i^2 + \bar{X}^2 - 2X_i\bar{X}\right)\right]$$

$$= \frac{1}{n} E\left[\sum_{i=1}^{n}\left(X_i^2 - n\bar{X}^2\right)\right]$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n} E\left(X_i^2\right) - nE\left(\bar{X}^2\right)\right]$$

อย่างไรก็ดีเนื่องจาก $E(X_i^2) = \mu^2 + \sigma^2$ และ $E(\bar{X}^2) = \mu^2 + \sigma^2/n$

เราจะได้ว่า

$$E(S^2) = \frac{1}{n}\left[\sum_{i=1}^{n}\left(\mu^2 + \sigma^2\right) - n\left(\mu^2 + \sigma^2/n\right)\right]$$

$$= \frac{1}{n}\left(n\mu^2 + n\sigma^2 - n\mu^2 - \sigma^2\right)$$

$$= \frac{n-1}{n}\sigma^2$$

จึงสรุปได้ว่า S^2 เป็น biased estimator สำหรับ σ^2 ทั้งนี้เพราะว่า

$$E(S^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

จากสมการสุดท้ายข้างต้น ผู้อ่านคงพบแล้วว่าถึงแม้ว่า S^2 จะไม่เป็น unbiased estimator ของ σ^2 ก็จริง แต่ก็เกือบที่จะเป็น unbiased estimator ของ σ^2 อยู่แล้ว ทั้งนี้เพราะว่า

$$\frac{n-1}{n}\sigma^2 \approx \sigma^2$$

เพื่อให้ได้ตัวประมาณค่าที่เป็น unbiased estimator สำหรับ σ^2 เราจึงขอนิยาม sample variance เสียใหม่ให้แตกต่างไปจากสูตร (1) ในข้อ (ข) เล็กน้อยดังนี้ คือ

(2) $$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

ด้วยนิยามของ sample variance ที่ได้ให้ไว้ใหม่ตามสมการ (2) นี้จะทำให้

$$E(S^2) = \frac{(n-1)}{(n-1)}\sigma^2 = \sigma^2$$

ซึ่งก็คือ S^2 จะเป็น unbiased estimator ของ σ^2 ไปให้ทันที เพราะฉะนั้นเพื่อความสะดวกการอภิปรายตั้งแต่นี้เป็นต้นไป เราจะใช้นิยามใหม่ของ sample variance ตามที่ได้นิยามไว้ในสมการ (2)

2.3 Precision of Estimation : The standard error

เมื่อเรานึกถึง statistics ขึ้นมา 1 ตัว (อาจจะ sample mean, sample variance, หรืออื่น ๆ) สิ่งที่เราจะคิดต่อไปก็คือ sampling distribution ของ statistics ตัวนั้น เช่น สมมุติว่าขณะนี้เรากำลังพิจารณา sample mean เราก็จะสามารถคำนวณหา

- mean ของ sampling distribution ของ mean

และ - standard deviation ของ sampling distribution ของ mean ได้

ดังนั้นคำว่า standard error ในกรณีของ mean นี้ก็จะหมายถึง standard deviation ของ sampling distribution ของ mean (มีนิยามเช่นเดียวกับกรณีของ variance หรืออื่น ๆ)

คำถามต่อไปก็คือว่า standard error นี้มีประโยชน์อย่างไร คำตอบก็คือเราสามารถใช้ standard error นี้วัดความเที่ยงตรงของการประมาณค่าแบบจุดได้ ผู้อ่านคงจะสามารถนึกถึงตัวอย่างเกี่ยวกับคะแนนสอบของนักศึกษาในหัวข้อ 2.1 ผ่านมา สำหรับคำถามแรกที่ว่า mean ของคะแนนสอบวิชา Prob.& Stat. ควรจะมีค่าเท่าไร? เรานำเอา sample mean มาให้คำตอบโดยประมาณว่า

mean ของประชากรควรจะมีค่าเท่ากับ 28.5 ถ้าเกิดมีคำถามต่อไปอีกว่าค่าโดยประมาณที่ตอบมานี้จะมีความน่าเชื่อถือมากน้อยเพียงใด เราจะใช้เครื่องมืออะไรมาวัดดี คำตอบก็คือเราสามารถใช้ standard error วัดความเที่ยงตรงการประมาณค่าแบบจุดได้ดังตัวอย่างต่อไปนี้

ตัวอย่าง สมมุติว่าเราต้องการหาสภาพการนำความร้อนของเหล็กชนิดหนึ่ง (สมมุติว่าชื่อ Armco iron) ไม่มีใครรู้ว่าสภาพการนำความร้อนของเหล็กชนิดนี้จริง ๆ แล้วมีค่าเท่าใดกันแน่ เพราะว่ามันเป็น unknown parameter แต่เราก็สามารถที่จะประมาณค่าสภาพการนำความร้อนของเหล็กชิ้นนี้ได้ โดยการไปตัดเหล็กชนิดนี้ให้มีขนาดและความยาวเท่ากันให้ได้จำนวนพอควร สมมุติว่าเป็น 10 ชิ้น (ค่า 10 นี้คือขนาดของตัวอย่าง) ต่อจากนั้นก็ให้นำเหล็กแต่ละชิ้นควบคุมอุณหภูมิไว้ที่ 100° F และปล่อยกระแสไฟฟ้าขนาด 550 W ผ่านเข้าไป แล้ววัดสภาพการนำความร้อนของเหล็กแต่ละชิ้น (หน่วยเป็น Btu / hr - ft - F°)
ได้ผลลัพธ์ดังนี้

41.60	41.48	42.34	41.95	41.86
42.18	41.72	42.26	41.81	42.04

ดังนั้นค่าเฉลี่ยของสภาพการนำความร้อน (ที่อุณหภูมิ 100° F และกำลังไฟ 500 W) ที่คำนวณได้จากสิ่งตัวอย่างก็คือ

$$\bar{x} = 41.924 \quad \text{Btu/hr - ft - °F}$$

สำหรับ standard error ของ sample mean นี้คือ

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

แต่ว่า σ เป็นค่าของประชากรเราไม่ทราบค่า ดังนั้นจึงจะแทน σ ในสมการสุดท้ายด้วย sample standard deviation s ผลก็คือ

$$\hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{0.284}{\sqrt{10}} = 0.0898$$

ขอให้สังเกตว่า standard error มีค่าราว 0.2% ของ sample mean ซึ่งพอจะบอกได้ว่าการคาดคะเนแบบจุดของเราใกล้เคียงค่าของประชากรพอสมควร ※

2.4 การประมาณค่าแบบช่วง (Interval estimation)

ในหลายสถานะการณ์การประมาณค่าแบบจุดอาจจะไม่ให้ข้อมูลที่เพียงพอเกี่ยวกับตัวพารามิเตอร์ที่เราสนใจ ตัวอย่างเช่นถ้าเราสนใจที่จะประมาณค่าเฉลี่ยกำลังอัดของคอนกรีต การประมาณด้วยค่าเพียงจำนวนเดียวอาจจะไม่มีประโยชน์มากนัก ดังนั้นในบางครั้งเราจึงชอบที่จะประมาณด้วยช่วงของจำนวนจริง และเรายังมีวิธีการจัดด้วยว่า ค่าของพารามิเตอร์ที่เราสนใจ นั้นจะตกอยู่ในช่วงดังกล่าวด้วยระดับความเชื่อมั่นเท่าใด เพื่อที่จะเข้าใจถึงวิธีการสร้างช่วงดังกล่าง สมมุติว่าเราเลือกสิ่งตัวอย่างเชิงสุ่ม $x_1, x_2, \ldots, x_n$ $(n \geq 30)$ มาจากประชากรที่เรายังไม่ทราบค่าเฉลี่ย μ แต่ของสมมุติว่าทราบค่าของความแปรปรวน σ^2 แล้ว การสร้างช่วงเพื่อประมาณค่าเฉลี่ย μ นั้นทำได้ดังนี้ : ให้ $\bar{x}$ เป็นค่าเฉลี่ยที่ได้สิ่งตัวอย่างดังกล่าวจากการที่เราได้เคยเรียนมาจากหัวข้อที่ 1 แล้วว่า sampling distribution ของ $\bar{x}$ นั้นจะเป็น $N(\mu, \sigma^2/n)$ ดังนั้นเมื่อมีการกำหนดค่าความน่าจะเป็น $1 - \alpha$ มาให้ (α เป็นค่าน้อยๆ เช่น 0.01, 0.05, 0.10) เราสามารถที่จะหาค่าของ $z(\alpha/2)$ จากตารางการแจกแจงปรกติได้เพื่อว่า

$$(1) \qquad P\left[-z(\alpha/2) \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z(\alpha/2)\right] = 1-\alpha$$

[ตัวอย่างเช่นถ้า $1 - \alpha = .095$ แล้ว $z(\alpha/2) = z(0.025) = 1.96$ และถ้า $1 - \alpha = 0.90$ แล้ว $z(\alpha/2) = z(0.05) = 1.645$]

แต่ว่านิพจน์ในวงเล็บของสมการ (1) นั้นสมมูลกับ

$$-z(\alpha/2) \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z(\alpha/2)$$

$$-z(\alpha/2)\frac{\sigma}{\sqrt{n}} \leq \bar{X}-\mu \leq z(\alpha/2)\frac{\sigma}{\sqrt{n}}$$

$$-\bar{X}-z(\alpha/2)\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X}+z(\alpha/2)\frac{\sigma}{\sqrt{n}}$$

$$\bar{X}-z(\alpha/2)\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}+z(\alpha/2)\frac{\sigma}{\sqrt{n}}$$

ดังนั้นความน่าจะเป็นในสมการ (1) จึงเขียนได้ใหม่ดังนี้

$$(2) \qquad P\left[\bar{X}-z(\alpha/2)\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}+z(\alpha/2)\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

ต่อไปสมมุติว่าได้ทำการทดลองชักสิ่งตัวอย่างขนาด n ออกมาประชากรกลุ่มดังกล่าวสมมุติว่าสิ่งตัวอย่างดังกล่าวคือ $x_1, x_2, \ldots, x_n$ และคำนวณหา sample mean ได้เท่ากับ $\bar{x}$ เมื่อแทนลงในสมการ (2) แล้วจะทำให้เราได้ช่วงเพื่อการประมาณค่า μ ดังนี้

$$(3) \qquad \left[\bar{x} - z(\alpha/2)\frac{\sigma}{\sqrt{n}}, \quad \bar{x} + z(\alpha/2)\frac{\sigma}{\sqrt{n}}\right]$$

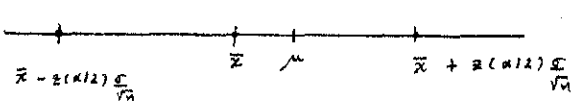
ตามความเป็นจริงแล้วเราก็ไม่ทราบว่าช่วง (3) นี้จะบรรจุค่า μ อยู่ด้วยหรือเปล่าแต่เรา สามารถพูดได้โดยอ้าง (2) ว่า "เรามีความเชื่อมั่นถึง $100(1-\alpha)$ เปอร์เซนต์ ว่าช่วง (3) จะบรรจุอยู่ μ ด้วย" มีบ่อยครั้งที่เราจะพูดอย่างง่าย ๆ ว่า $\bar{x} \pm z(\alpha/2)\sigma/\sqrt{n}$ คือ cenfidence interval สำหรับการประมาณค่า μ ด้วยระดับความเชื่อมั่น $100(1-\alpha)\%$

อยากจะให้นักศึกษาสังเกตว่าในการคำนวณหา cenfidence interval จาก (3) นั้น เราจะต้องทราบค่าอะไรบ้าง อันดับแรกก็คือ $\bar{x}$ ค่านี้หาได้จากการชักสิ่งตัวอย่าง, อันดับที่สองก็คือ $z(\alpha/2)$ ค่านี้หาได้จากตารางการแจกแจงแบบปรกติ (ส่วนค่าของ α เป็นค่าที่เรากำหนดขึ้นเอง ถ้าต้องการให้มีความเชื่อมั่นสูงก็กำหนด α ให้มีค่าน้อยเพื่อ $(1-\alpha)100\%$ จะได้มีค่ามาก, อันดับต่อไปที่จะต้องรู้ก็คือค่าของ σ ซึ่งเป็นค่าส่วนเบี่ยงเบนมาตราฐานของประชากร ค่าตัวนี้ลำบากหน่อยเพราะเป็นค่าของประชากร แต่ในหลายกรณีเราก็สามารถรู้มาก่อนแล้ว, ส่วนค่าของ n เป็นขนาดของตัวอย่างซึ่งจะทราบได้เสมอเมื่อทำการชักสิ่งตัวอย่างออกมา

มีนักศึกษาบางคนอาจจะตั้งข้อสังเกตว่าขณะนี้เราไม่ทราบค่าของ population mean μ จึงได้ดำเนินกิจกรรมเพื่อหาช่วงประมาณค่าของ μ แต่เมื่อไม่ทราบค่าของ μ ก็มักพลอยไม่ทราบค่าของ σ ตามไปด้วย ถ้าเป็นกรณีนี้จะหา cenfidence interval ได้อย่างไร นักสถิติก็พยายามหาคำตอบโดยให้นำเอาค่าของ sample standard deviation s ซึ่งจะมีค่าใกล้เคียงกับ σ เมื่อ$(n \geq 30)$ เข้าไปแทน μ ใน (3) จะทำให้เราได้ $(1-\alpha)100\%$ cenfidence interval for μ ใหม่เป็น

$$(4) \qquad \left[\bar{x} - z(\alpha/2)\frac{s}{\sqrt{n}}, \quad \bar{x} + z(\alpha/2)\frac{s}{\sqrt{n}}\right]$$

แต่เนื่องจากสูตร (4) เกิดจากการประมาณค่า σ ด้วย s ไปครั้งหนึ่งแล้ว ดังนั้นเมื่อพูดถึงระดับความเชื่อมั่น 95% เราอาจจะเลือกค่าที่ใกล้เคียงกับ 95% เช่น 93.5%หรือ 96.5% แทน



รูปแสดง confidence interval

ตัวอย่างที่ 1 จงเปิดตารางค่าของการแจกแจงแบบปรกติ N(0, 1) เพื่อหาค่าของ $z(\alpha/2)$ เมื่อ $\alpha = 0.1$, $\alpha = 0.05$

วิธีทำ : กรณีที่ 1 $\alpha = 0.1$ ดังนั้น $z(\alpha/2) = z(0.1/2)$

$= z(0.05)$

จากบทนิยามของ $z(0.1/2)$ เราทราบว่า $z(0.1/2)$ จะมีค่าเท่ากับค่าบนแกน x ที่ทำให้

$$P\left[-z(0.1/2) \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le z(0.1/2)\right] = 1 - 0.1 = 0.9$$

จากการที่ พ.ท. ใต้เส้นโค้งปรกติ N(0, 1) มีค่าเท่ากับ 1 ดังนั้นเราจะควรจะเลือก $z(0.1/2) = z(0.05) = 1.645$

กรณีที่ 2 $\alpha = 0.05$ เราจะได้ $z(0.05/2) = z(0.025) = 1.96$ สร้างเป็นตารางได้ดังนี้

Conifdence level	90%	95%
$z(\alpha/2)$	1.645	1.96

#

ตัวอย่างที่ 2 โรงงานผลิตสายเบรคแห่งหนึ่งต้องการทราบว่าสายเบรคที่ผลิตได้จะสามารถรับแรงดึงได้โดยเฉลี่ยเท่าใด เพื่อที่จะประมาณค่าเฉลี่ยของแรงดึงนี้ ได้มีการชักสิ่งตัวอย่าง เชิงสุ่มมา 32 เส้น แล้วทำการทดสอบแรงดึงของแต่ละเส้นต่อจากนั้นจึงหา sample mean ออกมาได้ $\bar{x} = 42{,}196$ ปอนด์ จากประสบการณ์ทางโรงงานทราบว่าส่วนเบี่ยงเบนมาตราฐานของการจัดแรงดึงมีค่าเท่ากับ $\sigma = 500$ ปอนด์ จงประมาณค่าแรงงานดึงเฉลี่ย μ ของสายเบรคที่ผลิตจากโรงงานนี้ กำหนดให้ $\alpha = 0.1$

วิธีทำ : จะใช้วิธีการประมาณด้วยวิธีการหา confidence interval. ในที่นี้โจทย์กำหนด $\sigma = 500$ ปอนด์มาให้ เราจึงจะใช้สูตร (3) เพื่อสร้าง confidence interval. ขอให้สังเกตว่า

(ก) $\alpha = 0.1$ ดังนั้น $z(0.1/2) = z(0.05) = 1.645$

(ข) $\sigma/\sqrt{n} = 500/\sqrt{32}$

ดังนั้นโดยสูตร (3) confidence interval สำหรับ μ ด้วยระดับความเชื่อมั่น 90% คือ

$$\left[\bar{x} - 1.645\,\sigma/\sqrt{32},\ \bar{x} + 1.645 * \sigma\sqrt{32}\right]$$

$$= \left[42{,}196 - 1.645(500/\sqrt{32}),\ 42{,}196 + 1.645(500/\sqrt{32})\right]$$

$$= [42{,}051,\ 42{,}341]$$

กล่าวโดยสรุปก็คือแรงดึงเฉลี่ยของสายเบรคที่ผลิตจากโรงงานนี้จะมีค่าอยู่ในช่วงดังกล่าว ด้วยระดับความเชื่อมั่นเท่ากับ 90%

<u>ข้อสังเกตจากตัวอย่างที่ 2</u> : ขอให้นักศึกษาทำตัวอย่างที่ 2 ใหม่โดยใช้ $\alpha = 0.05$ นั่นคือ เราจะเพิ่มระดับความเชื่อมั่นเป็น 95% ผลของการคำนวณจะพบว่า confidence interval จะขยายเพิ่มขึ้นไปมากกว่ากรณีของ $\alpha = 0.1$ กล่าวโดยสรุปแล้วคือ ถ้าเรายิ่งต้องการความเชื่อมั่นมากขึ้นช่วงของการประมาณก็จะเพิ่มมากขึ้นด้วย ซึ่งก็ดูสมเหตุสมผลกับปรากฏการณ์ตามความเป็นจริง

<u>ตัวอย่างที่ 8</u> : ทำตัวอย่างที่ 2 ใหม่แต่คราวนี้สมมุติว่าเราไม่ทราบค่าของ standard deviation σ และเปลี่ยนค่าของ α เป็น 0.05

<u>วิธีทำ</u> : ในกรณีที่ไม่ทราบค่าของ σ เราจะต้องใช้ค่า s ซึ่งเป็น standard deviation ที่หาได้จากสิ่งตัวอย่างเชิงประมาณค่าของ σ แล้วใช้สูตรที่ (4) ขอให้สังเกตว่า

$$z(\alpha/2) = z(0.05/2) = z(0.025) = 1.96 \approx 2$$

ดังนั้น 95% confidence interval for μ คือ

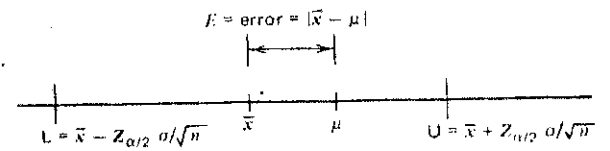
$$\bar{x} \pm (1.96)\,s/\sqrt{n} = 42{,}196 \pm 2(614)/\sqrt{32}$$

$$= [41{,}937,\ 42{,}413]$$

นั่นคือค่าเฉลี่ยของแรงดึงจะอยู่ในช่วง [41,937, 42,413] ด้วยระดับความเชื่อมั่น 95% #

2.5 การหาขนาดของสิ่งตัวอย่าง

ในการวางแผนการชักสิ่งตัวอย่างจากประชากรเพื่อนำมาประมาณค่าของ μ นั้น มีบ่อยครั้งที่เราต้องการทราบล่วงหน้าว่าควรจะเลือกสิ่งตัวอย่างเท่าไรดี เพื่อให้ค่าผิดพลาดในการประมาณซึ่งจะเขียนแทนด้วย $E = |\bar{x} - \mu|$ มีค่าไม่เกินค่าที่กำหนดให้ ดูรูปข้างล่างนี้



Error in estimating μ with $\bar{x}$.

วิธีการหาค่าของ n นั้นขอให้เราสังเกตสิ่งที่เคยเรียนมาแล้วว่า

$$P\left[\bar{X} - z(\alpha/2)\frac{\sigma}{\sqrt{n}} \le \bar{X} + z(\alpha/2)\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

ดังนั้นเราจึงควรจะเลือก n ที่สอดคล้องกับสมการ

$$E = z(\alpha/2)\frac{\sigma}{\sqrt{n}}$$

หรือ

$$n = \left[\frac{z(\alpha/2)\sigma}{E}\right]^2 \quad (5)$$

<u>ตัวอย่างที่ 4</u> : โรงงานผลิตสีแห่งหนึ่งสามารถผลิตสีได้โดยเฉลี่ยวันละ 70 ตัน ปัจจัยที่ทำให้ปริมาณการผลิตสีเปลี่ยนแปลงไปทุกวันก็คือมีการเปลี่ยนแปลงปริมาณของวัตถุดิบและสภาพของโรงงานเป็นต้น

สมมุติว่าปริมาณการผลิตในแต่ละวันมีการแจกแจงแบบปรกติด้วย mean $\mu = 70$ ตัน และ standard deviation $\sigma = 3$ ตัน

จากการที่ตลาดมีการขยายตัว ดังนั้นจึงมีความต้องการสีมากขึ้น ทางโรงงานจึงได้ปรับปรุงระบบการผลิตใหม่ เพื่อเพิ่มผลผลิต ขณะนี้ทางโรงงานจึงต้องการทราบว่าหลังจากเริ่มต้นเครื่องจักรด้วยระบบใหม่นี้แล้วจะสามารถผลิตสีได้โดยเฉลี่ยวันละเท่าใด เพื่อที่จะตอบคำถามนี้ ทางโรงงานได้ทำการชักสิ่งตัวอย่างเพื่อทดสอบ มีคำถามว่าควรจะเลือกขนาดของตัวอย่างเท่าไรดี จึงจะแน่ใจได้ว่าค่าผิดพลาดที่เกิดจากการประมาณมีค่าไม่เกิน 1 ตัน ด้วยระดับความเชื่อมั่น 95%

<u>วิธีทำ</u> : ให้ μ เป็นค่าเฉลี่ยของปริมาณการผลิตสีในแต่ละวันของโรงงานหลังจากที่ได้ปรับปรุงแล้ว ดังนั้น μ จึงเป็นตัวพารามิเตอร์ที่เราไม่ทราบค่า

ให้ $\bar{x}$ เป็นค่าเฉลี่ยของการผลิตสีในแต่ละวันที่ได้จากการชักสิ่งตัวอย่าง ดังนั้น $\bar{x}$ จึงเป็น sample mean

ให้ $E = |\bar{x} - \mu|$ เป็นค่าผิดพลาดที่เกิดจากการประมาณค่า

ในที่นี้โจทย์ต้องการให้ $E \leq 1$ โดยใช้ $\alpha = 0.05$

ใช้สูตร (5)

$$n = \left[\frac{z(0.05/2)(3)}{1}\right]^2 = \left[\frac{(1.96)(3)}{1}\right]^2 = 34.6$$

ดังนั้นเพื่อให้ค่าผิดพลาดไม่เกิน (1) เราต้องชักสิ่งตัวอย่างอย่างน้อย 35 วัน

2.6 <u>Confidence Interval on a Proportion</u>

ต่อไปเราจะเสนอการอภิปรายเกี่ยวกับวิธีการหา confidence interval สำหรับ <u>อัตราส่วน</u> (proportion) p ซึ่งเป็นค่าของประชากรที่เรายังไม่ทราบค่ามาก่อน ตัวอย่างเช่นผลิตภัณฑ์จากโรงงานแห่งหนึ่งที่สามารถยอมรับได้ (หรือมีตำหนิ) อัตราส่วน p ของโรงงานที่ใช้วิธีการทางสถิติในกระบวนการควบคุมคุณภาพเป็นต้น เนื่องจากผลิตภัณฑ์ที่ผลิตออกมาจากโรงงานแต่ละชิ้นจะสามารถจัดให้อยู่ในกลุ่มใดกลุ่มหนึ่งในสองกลุ่มดังต่อไปนี้คือ :

<u>กลุ่มแรก</u> (เรียกชื่อกลุ่มนี้ว่า "success") เป็นกลุ่มของผลิตภัณฑ์ที่ยอมรับได้ นั่นคือเป็นกลุ่มของสินค้าที่ได้มาตราฐานและสามารถที่จะส่งไปขายได้

<u>กลุ่มที่สอง</u> (เรียกชื่อกลุ่มนี้ว่า "failure") เป็นกลุ่มของผลิตภัณฑ์ที่มีตำหนิ นั่นคือเป็นกลุ่มของสินค้าที่ไม่ได้มาตราฐานและจะต้องถูกคัดออก

ดังนั้นถ้ามีการชักสิ่งตัวอย่างเชิงสุ่มขนาด n มาจากโรงงานนี้ เหตุการณ์ต่าง ๆ ที่เกิดขึ้นในตัวอย่างนี้ก็จะสมนัยกับการทดลองแบบแบร์นูลลี จำนวน n ครั้ง ด้วยค่าความน่าจะเป็นเท่ากับ p (ซึ่งเป็นค่าที่เรายังไม่ทราบ) ต่อไปให้ X เป็นจำนวนของ success ในสิ่งตัวอย่างเชิงสุ่มของเราและให้ X/n เป็นอัตราส่วนของ success ที่ได้จากตัวอย่าง (sample proportion) ดังนั้น X จึงมีการแจกแจงเป็นแบบทวินามที่มี mean เท่ากับ np และ variance เท่ากับ np(1 - p) และ sample proportion

$$\hat{p} = \frac{X}{n}$$

ก็เป็น unbiased estimater สำหรับ p ด้วย เพื่อที่จะเห็นความจริงข้อนี้ขอให้สังเกตว่า

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n} \cdot np = p$$

นอกจากนี้เมื่อ n มีค่ามากพอ (นั่นคือ $n \geq 30$) โดย Central Limit Theorem เราจะได้ว่าการแจกแจงของตัวแปรสุ่ม

$$\frac{X - np}{\sqrt{np(1-p)}} = \frac{(X/n) - p}{\sqrt{p(1-p)/n}}$$

จะเข้าใกล้การแจกแจงแบบปรกติ N(0, 1) ซึ่งหมายความว่า

$$P\left[-z(\alpha/2) \leq \frac{(X/n) - p}{\sqrt{p(1-p)/n}} \leq z(\alpha/2)\right] \approx 1 - \alpha$$

อสมการในวงเล็บข้างต้นสามารถที่จะเขียนได้ใหม่เป็น

(1) $$\frac{X}{n} - z(\alpha/2)\sqrt{\frac{p(1-p)}{n}} \leq p \leq \frac{X}{n} + z(\alpha/2)\sqrt{\frac{p(1-p)}{n}}$$

เราอยากจะใช้อสมการ (1) สร้าง confidence interval สำหรับ p ขอให้เรามาสำรวจดูว่าจะแทนค่าตัวแปรต่าง ๆ ในอสมการ (1) ได้อย่างไร ตัวแปรตัวแรกคือ $z(\alpha/2)$ ตัวแปรตัวนี้เราจะทราบค่าได้ทันทีเมื่อมีการกำหนดระดับของความเชื่อมั่น α มาให้จึงหมดปัญหาไป ตัวแปรตัวต่อไปก็คือ X/n ตัวแปรตัวนี้เราจะทราบค่าได้เมื่อมีการชักสิ่งตัวอย่างขนาด n ขึ้นมาและนับจำนวนของ success เทียบกับขนาดของตัวอย่าง สมมุติว่าได้ค่าเป็น x/n ก็นำค่านี้ไปแทน X/n ตัวแปรตัวสุดท้ายคือ p ค่านี้เป็นค่าของประชากรซึ่งเรายังไม่รู้ค่า อย่างไรก็คือเราก็สามารถประมาณค่า p ด้วยค่าของ x/n ซึ่งเป็นค่าที่ได้จากการชักสิ่งตัวอย่าง ด้วยวิธีการเช่นนี้เราก็สามารถสร้าง confidence interval สำหรับ p ด้วยระบบความเชื่อมั่น $100(1 - \alpha)$ เปอร์เซนต์ได้ดังนี้

(2) $$\frac{x}{n} - z(\alpha/2)\sqrt{\frac{(x/n)(1-x/n)}{n}} \leq p \leq \frac{x}{n} + z(\alpha/2)\sqrt{\frac{(x/n)(1-x/n)}{n}}$$

ซึ่งมีความหมายว่า "ความน่าจะเป็นที่ค่าพารามิเตอร์ p จะตกอยู่ในช่วงที่กำหนดโดย (2) นั้นมีค่าเท่ากับ $1-\alpha$" หรือ "เรามีความเชื่อมั่นนั่นคือ $100(1-\alpha)$ เปอร์เซนต์ว่าช่วงที่กำหนดโดย (2) จะบรรจุค่าพารามิเตอร์ p อยู่ด้วย"

เพื่อให้จำได้ง่ายจะให้ $\hat{p} = x/n$ และ $\hat{q} = 1 - p$ ดังนั้น confidence interval ตามนิยาม (2) จะอยู่ในรูป

$$\hat{p} - z(\alpha/2)\sqrt{\frac{\hat{p}\hat{q}}{n}} \le p \le \hat{p} + z(\alpha/2)\sqrt{\frac{\hat{p}\hat{q}}{n}} \tag{3}$$

หรือเขียนให้อยู่ในรูปกระทัดรัดยิ่งขึ้นอีกจะเป็น

$$\hat{p} \pm z(\alpha/2)\sqrt{\frac{\hat{p}\hat{q}}{n}} \tag{4}$$

ตัวอย่างที่ 1 ผู้จัดการโรงงานผลิตเพลารถยนต์แห่งหนึ่งต้องการทราบว่าอัตราส่วน p ของเพลาที่ผลิตใหม่ไม่ได้มาตราฐานเมื่อเทียบกับปริมาณการผลิตทั้งหมดจะมีค่าเท่าใด นักสถิติของโรงงานแห่งนี้จึงได้ชักสิ่งตัวอย่างมา 75 ชิ้นแล้วทำการทดสอบ พบว่ามีเพลาที่ไม่ได้มาตรฐานรวมอยู่ด้วย 12 ชิ้น ดังนั้นค่าของ sample proportion p จึงเท่ากับ $x/n = 12/75 = 0.16$

ด้วยวิธีการ point estimate เขาจึงสามารถประมาณค่าของ p โดยใช้ค่าของ $\hat{p}$ ดังนี้

$$p \approx \hat{p} = 0.16$$

ความจริงค่า 0.16 นี้ก็คือคำตอบที่ต้องการแล้ว แต่ผู้จัดการยังสนใจที่จะทราบ confidence interval สำหรับ p ด้วยระดับความเชื่อมั่น 95% นักสถิติจึงใช้สูตรที่ (3) สร้าง confidence interval สำหรับ p ได้ดังนี้

$$\hat{p} - z(0.25)\sqrt{\frac{pq}{n}} \le p \le \hat{p} + z(.025)\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

หรือ

$$.16 - 1.96\sqrt{\frac{(.16)(.84)}{75}} \le p \le .16 + 1.96\sqrt{\frac{(.16)(.84)}{75}}$$

หลังจากการคำนวณแล้วจะได้

$$.08 \leq p \leq .24$$

ซึ่งก็จะได้ข้อสรุปว่าเรามีความเชื่อมั่นถึง 95% ว่าค่าของประชากร p จะอยู่ในช่วง [0.08, 0.24] ด้วยการสร้าง confidence interval เช่นนี้ทำให้เรามั่นใจว่าค่า p คงจะอยู่ห่างจาก 0.16 ไม่มากนัก #

ขอให้สังเกตว่าถ้าให้ $E = |p - \hat{p}|$ แล้วสมการ (3) จะบอกว่าเราสามารถเชื่อได้ถึง $100(1 - \alpha)$ เปอร์เซนต์ว่าค่าผิดพลาด E นี้จะมีค่าน้อยกว่า $z(\alpha/2)\sqrt{\hat{p}\hat{q}/n}$ ดังนั้นการที่จะเลือกค่าของ n เพื่อให้เราสามารถเชื่อได้ถึง $100(1 - \alpha)$ เปอร์เซนต์ว่าค่าผิดพลาดในการประมาณไม่เกิน E จะทำได้โดยการให้

$$E = z(\alpha/2)\sqrt{\hat{p}\hat{q}/n}$$

และแก้สมการเพื่อหาค่าของ n จะได้ว่า

(5) $$n = \left(\frac{z(\alpha/2)}{E}\right)^2 \hat{p}\hat{q}$$

ตัวอย่างที่ 2 อ้างถึงตัวอย่างที่ 1 นักสถิติต้องการทราบว่าควรจะชักสิ่งตัวอย่างขนาด n เท่าไรดีเพื่อให้ค่าผิดพลาดที่เกิดการใช้ $\hat{p}$ ประมาณ p นั้นมีค่าน้อยกว่า 0.05 ด้วยระดับความเชื่อมั่น 95%

วิธีทำ ในที่นี้เรามี $\alpha = 0.05$ $\hat{p} = 0.16$ $\hat{q} = 0.84$ ต่อไปอ้างสมการ (5) จะได้ว่า

$$n = \left(\frac{z(0.025)}{E}\right)^2 \hat{p}\hat{q}$$
$$= \left(\frac{1.96}{0.05}\right)^2 (0.16)(0.84) = 207$$

ดังนั้นถ้าเลือกขนาดของสิ่งตัวอย่างเท่ากับ 207 เรามีความมั่นใจว่าถึง 95% ว่าการใช้ $\hat{p}$ ประมาณค่าของ p นั้นจะมีความผิดพลาดน้อยกว่า 0.05 เปอร์เซนต์ #

แบบฝึกหัด

การประมาณค่าแบบจุด :

1. Suppose we have a random sample of size 2n from a population denoted by X, and $E(X) = \mu$ and $V(X) = \sigma^2$. let

$$\overline{X}_1 = \frac{1}{2n}\sum_{i=1}^{2n} X_i \quad \text{and} \quad \overline{X}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

be two estimators of μ. Which is the better estimator of μ ? Explain your choice.

Ans : Both estimators are unbiased. Now,

$V(\overline{X}_1) = \sigma^2/2n$ while $V(\overline{X}_2) = \sigma^2/n$. Since

$V(\overline{X}_1) < V(\overline{X}_2)$, $\overline{X}_1$ is an more efficient than $\overline{X}_2$.

2. Let three random sample of sizes $n_1 = 10$, $n_2 = 8$, and $n_3 = 6$ be taken from a population with mean μ and variance σ^2. Let S_1^2, S_2^2 and S_3^2 be the sample variances. Show that

$$S^2 = \frac{10S_1^2 + 8S_2^2 + 6S_3^2}{24}$$

is an unbiased estimator of σ^2.

3. A sample of five measurement of the diameter of a sphere were recorded by a scientist as 6.33, 6.37, 6.36, 6.32, 6.37 inches. Determine unbiased and efficient estimates of (a) the true mean. (b) the true variance.

Ans : (a) $\mu = \dfrac{\sum_{i=1}^{5} x_i}{5} = 6.35$ in.

(b) $\hat{\sigma}^2 = s^2 = \dfrac{\sum_{i=1}^{5} (x_i - \hat{\mu})^2}{4} = 0.00055$ in.^{2}

Note that $s = \sqrt{0.00055}$ is an estimate of the true standard deviation but this estimate is neither unbiased nor efficient

4. A civil engineer is analyzing the compressive strength of concrete. Compressive strength is approximately normally distributed with variance $\sigma^2 = 1000$ (psi). A random sample of 12 specimens has a mean compressive strength of $\bar{x} = 3250$ psi.

(a) Construct a 95 percent two-sided confidence interval on mean compressive strength.

(b) Construct a 99 percent two-sided confidence on mean compressive strength. Compare the width of this confidence interval with the width of the one found in part (a)

Ans : (a) ✓ $3232.11 \le \mu \le 3267.89$

(b) ✓ $3226.49 \le \mu \le 3273.51$

5. Suppose that in Exercise 4 it is desired to estimate the compressive strength with an error that is less that 15 psi. What sample size is required?

6. A sample poll of 100 voters chosen at random from all voters in a given district indicated that 55% of them were in favor of a particular candidate. Find (a) 95%, (b) 99% confidence limits for the proportion of all the voters in favor of this candidate.

Ans : (a) ✓ $0.55 - 0.10 \le p \le 0.55 + 0.10$

(b) $0.55 - 0.15 \le p \le 0.55 + 0.15$

7. How large a sample of voters should we take in Exercise 6 in order to be (a) 95% and (b) 99.73% confident that the condidate will be elected?

Ans : (a) $n = 384.2$ or at least 385

(b) $n = 900$

3. การทดสอบสมมุติฐาน (Tests of Hypotheses)

ในทางปฏิบัติมีบ่อยครั้งที่เราต้องตัดสินใจ เพื่อให้ได้ข้อสรุปเกี่ยวกับค่าของประชากร โดยอาศัยข้อมูลที่ได้มาจากการชักสิ่งตัวอย่าง การตัดสินใจบนพื้นฐานของการชักสิ่งตัวอย่างนี้ จะเรียกว่า "การตัดสินใจเชิงสถิติ" (statistical decision theory) ตัวอย่างเช่น เราอาจจะต้องการตัดสินว่าตัวยาชนิดใหม่ที่ได้ผลิตขึ้นมานี้ ให้ผลในการรักษาดีกว่าตัวยาที่ใช้อยู่ในขณะนี้หรือไม่? หรือระบบการศึกษาแผนใหม่จะดีกว่าระบบเดิมหรือไม่ หรือต้องการตัดสินว่าเหรียญอันหนึ่งเป็นเหรียญเที่ยงตรงหรือไม่? เป็นต้น

3.1 สมมุติฐานเชิงสถิติ (statistical hypotheses)

ในกระบวนการตัดสินใจนั้นนักสถิติต้องสร้างข้อสมมุติฐานบางประการเกี่ยวกับค่าของประชากรที่กำลังสนใจ ข้อสมมุติดังกล่าวอาจจะถูกต้องหรือไม่ก็ได้ซึ่งเราจะเรียกว่า สมมุติฐานเชิงสถิติ ตัวอย่างเช่น สมมุติว่าเราสนใจค่าเฉลี่ยของกำลังอัด (compressive strength) ของคอนกรีตชนิดหนึ่ง และเราต้องการที่จะตัดสินว่าค่าเฉลี่ยของกำลังอัด (μ) ของคอนกรีตชนิดนี้จะเท่ากับ 2,500 psi หรือไม่? ในกรณีเช่นนี้เราอาจจะตั้งสมมุติฐานไว้ว่า

(1)
$$H_0 : \mu = 2{,}500 \text{ psi}$$
$$H_1 : \mu \neq 2{,}500 \text{ psi}$$

จะเรียกสมมติฐาน $H_0 : \mu = 2{,}500$ psi ว่า "null hypotheses" และเรียก $H_1 : \mu \neq 2{,}500$ psi ว่า "Alternative hypotheses" เนื่องจากสมมุติฐาน H_1 นั้นได้ระบุว่า μ อาจจะมีค่ามากกว่า 2,500 psi หรือน้อยกว่า 2,500 psi ก็ได้ เราจึงเรียกสมมุติฐาน H_1 ได้ในอีกชื่อหนึ่งว่า "two-sided alternative hypotheses" ในบางกรณีเราอาจจะตั้งข้อสมมุติฐานเป็นทำนอง "one-sided alternative hypotheses" ดังนี้

(2)
$$H_0 : \mu = 2{,}500 \text{ psi}$$
$$H_1 : \mu > 2{,}500 \text{ psi}$$

เป็นเรื่องสำคัญสำหรับผู้อ่านที่จะต้องจำไว้เสมอว่า ข้อสมมุติฐานที่ได้ตั้งขึ้นนั้นเป็นข้อความเกี่ยวกับค่าของประชากรที่เรากำลังศึกษาอยู่ ไม่ใช่ข้อความเกี่ยวกับค่าของสิ่งตัวอย่าง ค่าของประชากร (ค่าพารามิเตอร์) ตามที่เรานำมาใช้ในการตั้ง null hypotheses (ดังตัวอย่างข้างต้น $\mu = 2{,}500$ psi) ได้มาตามวิธีการใดวิธีการหนึ่งในสามวิธีดังนี้คือ

(ก) ได้มาจากประสบการณ์ที่ผ่านมา, ได้มาจากการที่เรารู้เรื่องกระบวนการผลิต, หรือได้มาจากการทดลองเป็นต้น ถ้าเราได้ค่าของพารามิเตอร์มาจากวิธีการดังกล่าวข้างต้นแล้ว วัตถุประสงค์ของการทดสอบสมมุติฐานก็คือการตัดสินว่าได้มีการเปลี่ยนแปลงในวิธีการผลิตบ้างหรือไม่

(ข) ได้มาจากทฤษฎีบางทฤษฎี หรือได้มาจากแบบจำลองที่สร้างขึ้นเพื่ออธิบายกระบวนการผลิตนั้น สำหรับกรณีนี้วัตถุประสงค์ของการทดสอบสมมุติฐานก็จะเป็นการแสดงว่าตัวทฤษฎีหรือแบบจำลองที่ได้ตั้งไว้เป็นจริง

(ค) ได้มาจากข้อกำหนดที่ตกลงกันไว้ล่วงหน้า เช่น ข้อกำหนดในงานออกแบบเชิงวิศวกรรม หรือข้อกำหนดตามพันธะสัญญาเป็นต้น สำหรับกรณีนี้วัตถุประสงค์ของการทดสอบก็จะเป็นการตัดสินว่าได้มีการปฏิบัติตามพันธะสัญญา (conformal testing) หรือไม่?

เมื่อเราสนใจในการตัดสินว่าข้อสมมุติฐานอันหนึ่งควรเป็นจริงหรือเท็จ กระบวนการที่จะนำไปสู่การตัดสินดังกล่าวจะเรียกว่า "การทดสอบสมมุติฐาน" (test of a hypothesis) กระบวนการนี้จะใช้ข้อมูลในสิ่งตัวอย่างเชิงสุ่ม ซึ่งได้ถูกชักออกมาจากกลุ่มของประชากรที่เราสนใจ ถ้าข้อมูลที่ได้จากสิ่งตัวอย่างนี้ลงรอยกับข้อสมมุติฐานที่ได้ตั้งไว้แล้ว เราจะสรุปว่าข้อสมมุติฐานดังกล่าวเป็นจริง อย่างไรก็ดีถ้าข้อมูลดังกล่าวไม่ลงรอยกันกับข้อสมมุติฐานก็จะสรุปว่าข้อสมมุติฐานเป็นเท็จ

เพื่อที่จะทำการทดสอบสมมุติฐาน เราจะเลือกสิ่งตัวอย่างเชิงสุ่มนี้ขึ้นมาหนึ่งชุด และคำนวณหาค่า statistic (ซึ่งต่อไปจะเรียกว่า test statistic) จากสิ่งตัวอย่างนั้น ต่อจากนั้นก็นำข้อมูลที่ได้จากสิ่งตัวอย่างมาช่วยในการตัดสินใจ ตัวอย่างเช่น การทดสอบ null hypothesis เกี่ยวกับกำลังอัดของคอนกรีตตามสมการ (1) สมมุติว่าได้มีการชักสิ่งตัวอย่างเชิงสุ่มจำนวน 10 ชิ้น นำมาทดสอบกำลังอัดและคำนวณหาค่า sample mean $\bar{x}$ ในที่นี้จะขอตกลงกันว่าถ้าค่าของ $\bar{x} \geq 2{,}550$ psi หรือ $\bar{x} \leq 2{,}450$ psi แล้วเราจะตัดสินว่าค่าเฉลี่ยของกำลังอัดของคอนกรีตชนิดนี้ไม่ใช่ 2,500 psi นั่นคือจะปฏิเสธ $H_0 : \mu = 2{,}500$ psi การปฏิเสธ H_0 จะหมายถึงการยอมรับว่า H_1 เป็นจริง เซตของค่าของ $\bar{x}$ ทั้งหมดที่มากกว่า 2,550 psi หรือน้อยกว่า 2,450 psi นั้นจะเรียกว่าเป็น critical region หรือ rejection region ของการทดสอบ ถ้า $2{,}450 \text{ psi} \leq \bar{x} \leq 2{,}550$ psi แล้วเราจะยอมรับ null hypothesis $H_0 : \mu = 2{,}500$ psi ดังนั้นจะเรียกช่วง [2,450, 2,550] นี้ว่า acceptance region สำหรับการทดสอบ ขอให้สังเกตว่าค่าที่ขอบของ critical region (จะเรียกว่า critical value ของ test statistic) ซึ่งก็คือ 2,450 psi หรือ 2,550 psi ได้ถูกกำหนดขึ้นมาตามใจชอบในขณะนี้

แต่ว่าในลำดับต่อไปนั้นจะได้มีการแสดงให้เห็นถึงวิธีการสร้าง test statistic ที่เหมาะสมเพื่อหา critical region สำหรับการทดสอบสมมุติฐานในสถานการณ์ต่าง ๆ กัน

รูป 1 แสดงบริเวณ critical region และ acceptance region

3.2 ความผิดพลาดประเภทที่ I และ II (Type I and Type II error)

การที่จะตัดสินใจว่าจะยอมรับหรือปฏิเสธ null hypothesis นั้นเราจะใช้ test statistic ซึ่งคำนวณได้จากข้อมูลในสิ่งตัวอย่างเชิงสุ่มเป็นตัวช่วยในการตัดสินใจ ดังนั้นเมื่อมีการใช้ข้อมูลจากสิ่งตัวอย่างเราจะต้องจดจำไว้เสมอว่าอาจจะมีความผิดพลาดเกิดขึ้นในการตัดสินใจดังกล่าวด้วย ความผิดพลาดที่เกิดขึ้นในการทดสอบสมมุติฐานมีได้สองชนิดดังนี้คือ

(ก) Type I error เป็นความผิดพลาดที่เกิดขึ้นจากการปฏิเสธ null hypothesis เมื่อ null hypothesis ดังกล่าวถูกต้อง

(ข) Type II error เป็นความผิดพลาดที่เกิดขึ้นจากการยอมรับ null hypothesis เมื่อ null hypothesis ดังกล่าวไม่ถูกต้อง

สถานการณ์ดังกล่างข้างต้นสามารถอธิบายได้เป็นตารางดังต่อไปนี้

Decision in Hypothesis Testing

Decision	H_0 is True	H_0 is fabe
Accept H_0	No error	Type II error
Reject H_0	Type I error	No error

ความน่าจะเป็นที่จะเกิด type I และ type II error นั้นจะกำหนดไว้เป็นสัญญลักษณ์พิเศษดังนี้

$$\alpha = P\{\text{type I error}\} = P\{\text{reject } H_0 \mid H_0 \text{ is true}\}$$

$$\beta = P\{\text{type II error}\} = P\{\text{accept } H_0 \mid H_0 \text{ is false}\}$$

ต่อไปนิยาม "<u>กำลังของการทดสอบ</u>" (power of test) ดังนี้

$$\text{Power} = 1 - \beta = P\{\text{reject } H_0 \mid H_0 \text{ is false}\}$$

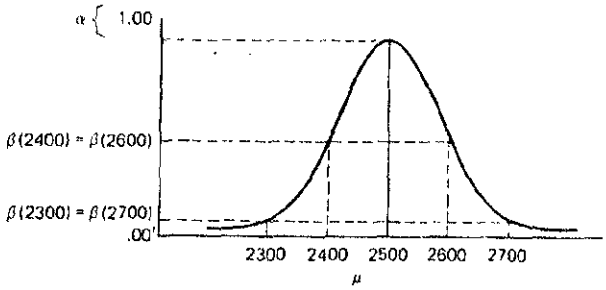
ขอให้สังเกตว่ากำลังของการทดสอบ คือ ความน่าจะเป็นที่จะปฏิเสธ null hypothesis เมื่อ null hypothesis นั้นผิด จากการที่ผลของการทดสอบสมมุติฐานขึ้นอยู่กับความผิดพลาดด้วย และเราก็ไม่สามารถที่จะ "พิสูจน์" หรือ "หาข้อขัดแย้ง" ได้ว่าสมมุติฐานอันไหนถูกต้องกันแน่ ดังนั้นสิ่งที่จะทำได้อย่างดีที่สุดก็คือพยายามออกแบบวิธีการทดสอบเพื่อที่ควบคุมค่าความน่าจะเป็นของความผิดพลาดให้มีค่าน้อยลงอยู่ในระดับที่น่าพอใจ

ความน่าจะเป็น α ของการเกิด type I error นี้ มีบางครั้งจะเรียกว่า <u>ระดับของความมีนัยสำคัญ</u> (level of significance) ของการทดสอบ ในตัวอย่างการทดสอบกำลังอัดของคอนกรีตตามที่กล่าวถึงข้างต้น สมมุติว่าค่าเฉลี่ยของกำลังอัดที่แท้จริงคือ $\mu = 2{,}500$ psi ดังนั้น type I error จะเกิดขึ้นเมื่อค่า sample mean $\bar{x} > 2{,}550$ psi หรือ $\bar{x} < 2{,}450$ psi โดยทั่วไปแล้วค่าความน่าจะเป็นที่จะยอมให้เกิด type I error นั้นจะเป็นค่าคงตัวซึ่งผู้ทำการทดสอบสามารถกำหนดได้เองตามความเหมาะสม ดังนั้นความน่าจะเป็นที่จะปฏิเสธ H_0 เมื่อ H_0 นั้นถูกต้องจึงสามารถจะควบคุมได้โดย ผู้ที่ทำการทดสอบเพราะฉะนั้นข้อสรุปที่เป็นการปฏิเสธ H_0 จึงถือได้ว่าเป็น "ข้อสรุปที่แรง" (strong conclusion)

ต่อไปสมมุติว่า null hypothesis $H_0 : \mu = 2{,}500$ psi นั้นไม่ถูกต้อง ดังนั้นค่าเฉลี่ยของกำลังอัด μ ที่ถูกต้องก็จะเป็นค่าอื่น ๆ ที่ไม่เท่ากับ 2,500 psi ดังนั้นความน่าจะเป็นที่จะเกิด type II error, ซึ่งก็คือ

$$\begin{aligned}\beta = P\{\text{type II error}\} &= P\{\text{accept } H_0 \mid H_0 \text{ is false}\} \\ &= P\{\text{accept } H_0 \mid H_1 \text{ is true}\},\end{aligned}$$

จึงไม่เป็นค่าคงตัวแต่มันจะขึ้นอยู่กับค่าเฉลี่ย μ ที่ทำให้ H_1 เป็นจริง ถ้าให้ μ แทนกำลังอัดที่แท้จริงของคอนกรีต (ซึ่งเป็นตัวแปรเพราะเราไม่ทราบว่ากำลังอัดที่แท้จริงของคอนกรีตมีค่าเท่าใด) $\beta(\mu)$ ก็จะเป็นค่าของความน่าจะเป็นที่จะเกิด type II error ที่สมนัยกับ μ เรานิยาม operating characteristic curve (หรือเรียกสั้น ๆ ว่า OC curve) ของการทดสอบหนึ่งให้หมายถึงกราฟของฟังก์ชัน $\beta(\mu)$ ตัวอย่างของ OC curve สำหรับปัญหาการทดสอบคอนกรีตได้ถูกแสดงไว้ในรูป 2 ข้างล่างนี้



รูป 2 แสดง OC curve สำหรับปัญหาการทดสอบคอนกรีต

จากเส้นโค้ง OC ในรูป 2 ข้างต้น นักศึกษาจะพบว่าค่าของ $\beta(\mu)$ ที่จุด $\mu = 2{,}500$ นั้นมีค่าสูงสุดและมีค่าเกือบเท่ากับ 1 แต่ขณะที่ μ เคลื่อนที่ห่างออกจาก 2,500 (ทั้งทางซ้ายและขวา) ค่าของ $\beta(\mu)$ จะค่อย ๆ ลดลงตามลำดับ คำอธิบายเหตุผลในเรื่องนี้นั้นจะใช้บทนิยามของ $\beta(\mu)$ เป็นหลัก ตัวอย่างเช่น $\beta(2{,}500)$ มีค่ามากเกือบเท่ากับ 1 เพราะว่าโดยบทนิยามแล้ว

$$\beta(2{,}500) = P\{\text{accept } H_0 : \mu = 2{,}500 \mid H_1 : \mu = 2{,}500 \text{ is true}\}$$

จึงพบว่าด้วยบทนิยามของ $\beta(2{,}500)$ เช่นนี้ จึงพบว่าได้มีการสมมุติไว้ในเงื่อนไขของความน่าจะเป็นแล้วว่า H_1 เป็นจริง นั่นคือค่าเฉลี่ยของกำลังอัดที่แท้จริงมีค่าเท่ากับ 2,500 psi ดังนั้นความน่าจะเป็นที่เราจะยอมรับ H_0 ว่า $\mu = 2{,}500$ จึงควรมีค่ามากและเข้าใกล้ 1 เป็นธรรมดา อีกตัวอย่างหนึ่งก็คือ ค่าของ $\beta(2{,}600)$ จากรูป 2 จะพบว่า $\beta(2{,}700) < \beta(2{,}600) < \beta(2{,}500)$ คำอธิบายอสมการนี้ก็มาจากบทนิยามอีกเช่นกัน เนื่องจาก

$$\beta(2{,}600) = P\{\text{accept } H_0 : \mu = 2{,}500 \mid H_1 : \mu = 2{,}600 \text{ true}\}$$

ตามที่ได้สมมุติไว้ในเงื่อนไขของความน่าจะเป็นว่า H_1 เป็นจริง ดังนั้นสำหรับกรณีนี้ค่าเฉลี่ยกำลังอัดที่แท้จริงมีค่าเท่ากับ 2,600 psi ความน่าจะเป็นที่จะยอมรับว่า $\mu = 2,500$ โดยที่ค่าเฉลี่ยจริงเท่ากับ 2,600 จึงควรมีค่าน้อยลง นั่นคือเราจะได้ความสัมพันธ์ที่ว่า

$$\beta(2,600) < \beta(2,500)$$

ส่วนอสมการที่เหลือมีการพิจารณาเช่นเดียวกัน อย่างไรก็ดีในกรณีพิเศษเมื่อ $\mu = 2,500$ ความน่าจะเป็นของ type II error จะมีค่าเท่ากับ $\beta = 1 - \alpha$ เท่าที่ได้ศึกษากันมาพอที่จะสรุปได้ว่า ความน่าจะเป็นที่จะเกิด type II error เป็นฟังก์ชันของปริมาณที่ทำให้ null hypothesis H_0 เป็นเท็จ

สำหรับในตัวอย่างเกี่ยวกับกำลังอัดของคอนกรีตนี้ปริมาณดังกล่าวคือ ค่าเฉลี่ยที่แท้จริง μ เพราะฉะนั้น $\beta = \beta(\mu)$ แต่ยังเป็นเรื่องที่สามารถพิสูจน์ได้ต่อไปอีกว่า β เป็นฟังก์ชันของ sample size n ด้วย เพราะฉะนั้นเราจึงสามารถเขียน $\beta = (\mu, n)$ ถ้าจัดให้ μ มีค่าคงที่ และ n เพิ่มมากขึ้น β จะมีค่าลดลง ข้อสังเกตนี้สอดคล้องกับความจริงที่ว่ายิ่งขนาดของ sample size เพิ่มขึ้นมากเท่าไร ความน่าจะเป็นที่จะเกิดความผิดพลาดในการตัดสินค่าประชากรก็จะลดลงด้วย

เนื่องจากความน่าจะเป็นของ type II error เป็นฟังก์ชันของตัวแปร 2 ตัวดังกล่าวข้างต้น ดังนั้นในการคำนวณค่าของ β เราต้องการข้อมูลถึง 2 ประการ (ในตัวอย่างข้างต้นก็คือ μ และ n) แต่ β เป็นค่าความน่าจะเป็นที่จะยอมรับ H_0 ดังนั้นการตัดสินใจในการยอมรับ H_0 จึงถือว่าเป็น <u>ข้อสรุปที่อ่อน</u> (weak conclusion) เพราะฉะนั้นแทนที่เราจะกล่าวว่า "ยอมรับ H_0 " เราจึงนิยมที่จะกล่าวว่า "ไม่สามารถปฏิเสธ H_0 " แทน คำว่า "ไม่สามารถปฏิเสธ H_0 " นั้นมีความหมายว่าเราไม่สามารถหาหลักฐานที่พอเพียงในการปฏิเสธ H_0 ดังนั้นการที่ไม่สามารถปฏิเสธ H_0 จึงไม่ได้หมายความว่ามีความน่าจะเป็นสูงที่ว่า H_0 เป็นจริง แต่อาจจะมีความหมายแต่เพียงว่าต้องการข้อมูลมากกว่านี้จึงจะสามารถให้ข้อสรุปที่แรงได้

3.3 <u>สมมุติฐานข้างเดียวและสองข้าง (One-sided and Two-sided Hypothesis)</u>

เนื่องจากการปฏิเสธ H_0 จะถือว่าเป็นข้อสรุปที่แรง และการที่ไม่สามารถปฏิเสธ H_0 จะถือว่าเป็นข้อสรุปที่อ่อน (เว้นแต่เรารู้ว่า β มีค่าน้อย) เพราะฉะนั้นเทคนิคในการตั้งสมมุติฐานจึงนิยมที่จะเอาข้อสรุปที่แรงมาไว้เป็น alternative hypothesis H_1 ตัวอย่างของปัญหาที่เราควรจะตั้งสมมุติฐานแบบสองข้างเป็นดังนี้คือ:

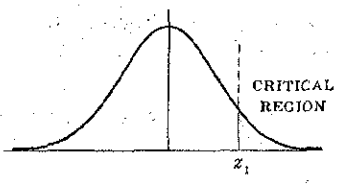
สมมุติว่าเราต้องการตัดสินว่าค่าเฉลี่ย μ ของการแจกแจงชนิดหนึ่งจะเท่ากับค่าคงตัวค่าใดค่าหนึ่งเช่น μ หรือไม่? และในโอกาสเดียวกันเราก็สนใจที่จะตรวจสอบค่าที่ถูกต้องของ μ ว่าสามารถที่จะมากกว่า μ หรือน้อยกว่า μ หรือไม่? ด้วยในกรณีเช่นนี้ควรจะตั้งสมมุติฐานเป็นประเภทสองข้างดังนี้คือ

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

แต่ก็ยังคงมีปัญหาในการทดสอบหลายปัญหาที่เหมาะสมสำหรับการตั้งสมมุติฐานเพียงข้างเดียว ตัวอย่างเช่นสมมุติว่าเราต้องการที่จะปฏิเสธ H_0 เฉพาะกรณีค่าที่ถูกต้องของ mean มีค่ามากเกิน μ_0 ในกรณีนี้จึงควรตั้งสมมุติฐานเป็นดังนี้คือ

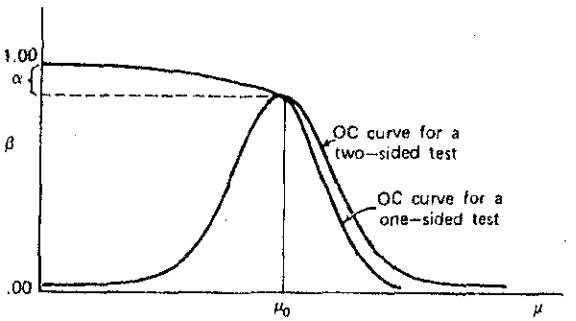
(3)
$$H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0$$

ในกรณี critical region จะอยู่ทางด้านขวามือ ดังรูป 3 ข้างล่างนี้



รูป 3

ดังนั้นถ้าการตัดสินใจอยู่บนพื้นฐานของ sample mean $\bar{x}$ แล้ว เราควรจะปฏิเสธ H_0 ในสมการ (4) ถ้า $\bar{x}$ มีค่ามากเกินไป รูป 4 ข้างล่างนี้ได้แสดงเส้นโค้ง OC สำหรับการทดสอบข้างเดียวและการทดสอบสองข้างไปพร้อม ๆ กันเลย



รูป 4 แสดง OC curve สำหรับการทดสอบสองข้างและข้างเดียว

ขอให้สังเกตจากรูป 4 ข้างต้นว่า เมื่อค่าเฉลี่ยที่แท้จริง μ มากกว่า μ_0 (นั่นคือ alternate hypothesis $H_1 : \mu > \mu_0$ เป็นจริง) การทดสอบข้างเดียวจะให้ผลลัพธ์ดีกว่าการทดสอบสองข้างในความหมายที่ว่า การทดสอบข้างเดียวมี OC curve ที่มีความชันมากกว่า แต่เมื่อค่าเฉลี่ยที่แท้จริง $\mu = \mu_0$ การทดสอบข้างเดียวและสองข้างจะให้ผลลัพธ์ที่สมมูลกัน อย่างไรก็ดีเมื่อ OC curve ของการทดสอบทั้งสองจะแตกต่างกันมากและในกรณีนี้การทดสอบสองข้างจะให้ค่าความน่าจะเป็นในการตรวจสอบความแตกต่างไปจาก μ_0 ได้ดีกว่าการทดสอบข้างเดียว กล่าวโดยสรุปแล้วเราจะใช้การทดสอบข้างเดียวเมื่อเรามั่นใจว่า μ มีค่าไม่น้อยกว่า μ_0 หรือถ้า $\mu < \mu_0$ เราจะยอมรับ null hypothesis แทน

3.4 ตัวอย่างเกี่ยวกับค่าผิดพลาดทั้งสอง

ผู้จัดการโรงงานผลิตแผงวงจรไฟฟ้า (printed eircuit) แห่งหนึ่งได้สังเกตว่าโดยเฉลี่ยแล้วความน่าจะเป็นที่จะพบแผงวงจรไฟฟ้าที่ไม่ได้มาตรฐาน ซึ่งถูกผลิตออกมาจากโรงงานนี้มีค่าเท่ากับ $p = 0.03$ มีวิศวกรท่านหนึ่งของโรงงานนี้ ได้แนะนำผู้จัดการว่าด้วยวิธีการปรับปรุงกระบวนการผลิตตามที่เขาได้คิดขึ้นใหม่นี้จะทำให้จำนวนแผงวงจรไฟฟ้าที่ไม่ได้มาตรฐานลดลง โดยเขาอ้างว่า p ตัวใหม่จะมีค่าเท่ากับ 0.01 เท่านั้น ท่านผู้จัดการโรงงานเห็นด้วยที่จะทดลองปรับปรุงกระบวนการผลิตตามข้อเสนอใหม่นี้ เพื่อที่ตรวจสอบว่ากระบวนการผลิตใหม่จะให้ผลลัพธ์ตามคำกล่าวอ้างหรือไม่ จึงได้มีการชักสิ่งตัวอย่างเชิงสุ่มมา 100 แผง มาทำการทดสอบ และได้ตกลงกันว่าถ้าพบแผงวงจรไฟฟ้าไม่ได้มาตรฐานไม่เกิน 1 แผงทางโรงงานก็จะยอมรับกระบวนการผลิตใหม่นี้ แต่ถ้าไม่เป็นเช่นนั้นทางโรงงานก็จะหวนกลับไปใช้กระบวนการผลิตแบบเดิม

แต่ก่อนที่การทดสอบจะเริ่มขึ้นมีวิศวกรอีกท่านหนึ่งซึ่งได้เคยศึกษาสถิติมาบ้างได้เกิดข้อสงสัยว่าถึงแม้ว่าการทดสอบแผงวงจรไฟฟ้าจำนวน 100 แผง และพบแผงวงจรไม่ได้มาตรฐานเพียงไม่เกิน 1 แผง ก็ตามแต่ว่าค่า p ตัวใหม่นี้ก็อาจจะยังคงมีค่าเท่ากับ 0.03 เหมือนเดิมก็ได้ ซึ่งจะทำให้ทางโรงงานยอมรับกระบวนการผลิตใหม่ด้วยความเข้าใจผิดก็ได้ ความผิดพลาดเช่นนี้เราได้เคยเรียนรู้มาแล้วโดยเรียกชื่อว่า type I error ในอีกด้านหนึ่งกระบวนการผลิตใหม่นี้อาจจะให้ผลในการปรับปรุงคุณภาพของการผลิตจริง ๆ แต่เกิดไปตรวจพบแผงวงจร ในกลุ่มตัวอย่างมากกว่า 2 แผงที่ไม่ได้ มาตรฐานก็จะเป็นผลให้ทางโรงงานหันไปยอมรับกระบวนการผลิตเดิม และปฏิเสธกระบวนการผลิตใหม่นี้ทั้งๆ ที่กระบวนการผลิตใหม่ดีกว่า ความผิดพลาดประเภทนี้เราได้เคยเรียนรู้มาแล้วเช่นกัน และได้เรียกชื่อว่า type II error

เพื่อที่จะทดสอบว่ากระบวนการผลิตใหม่จะให้ผลดีกว่ากระบวนการผลิตเดิมหรือไม่ เขาจึงตั้งสมมุติฐานโดยเอาความเชื่อเดิมเป็น null hypiothesis ดังนี้

$$H_0 : p = 0.03$$
$$H_1 : p < 0.03$$

สำหรับวิธีการทดสอบจะกล่าวถึงในหัวข้อต่อไป

3.5 การทดสอบสมมุติฐานเกี่ยวกับ Mean โดยที่รู้ค่าของ Variance

สำหรับการทดสอบสมมุติฐานเกี่ยวกับ mean นี้เราจะเสนอการอภิปรายออกเป็น 2 กรณี คือ กรณีแรกจะขอสมมุติว่าเรารู้ค่า variance σ^2 ของประชากรแล้ว ส่วนกรณีที่ไม่ทราบค่าของ σ^2 จะมีทำการทดสอบโดยใช้ค่าของ sample variance s^2 แทน ซึ่งจะมีการอภิปรายในลำดับต่อไป

กรณีที่ 1 ขั้นตอนในการทดสอบสมมุติฐานเกี่ยวกับ mean เมื่อทราบค่า population variance σ^2 แล้ว มีดังนี้คือ

1. จัดตั้ง null hypothesis และ alternative hypothesis ประเภทข้างเดียวหรือสองข้างตามความเหมาะสม

   ปัญหาที่ยากสำหรับนักศึกษาก็คือว่าควรจะนำข้อความใดมาตั้งเป็น null hypothesis แนวคิดก็คือเราจะใช้หลักการเหมือนกับการสอบสวนในศาลที่ว่า

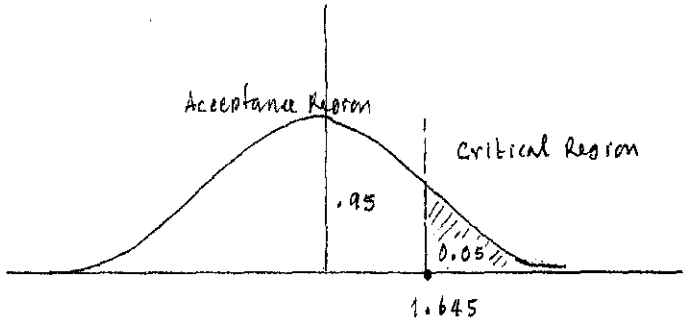
   "ผู้ต้องหาที่มาศาลทุกคนศาลจะถือว่าเป็นผู้บริสุทธิ์ เว้นแต่จะสามารถหาหลักฐานมายืนยันได้ว่าเขากระทำความผิดจริง"

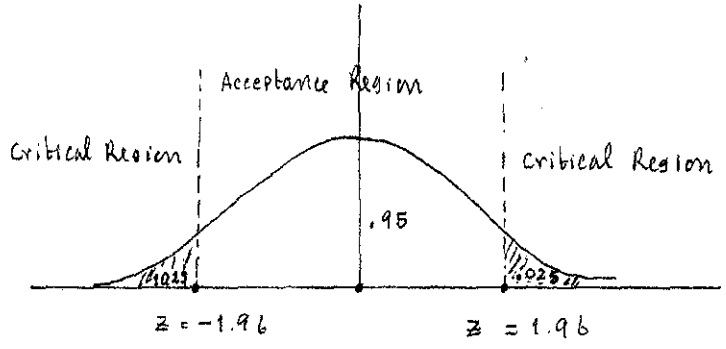
ด้วยหลักการเช่นนี้ข้อความที่จะนำมาตั้งเป็น null hypothesis ก็จะเป็นข้อความตามความเชื่อหรือประสบการณ์เดิมที่เคยมีมา ส่วนข้อความที่จะนำมาตั้งเป็น alternative hypothesis นั้นจะเป็นข้อความตามคำกล่าวอ้างใหม่ซึ่งต้องการการพิสูจน์ว่าเป็นจริง ตัวอย่างเช่น ผลลัพธ์ต่าง ๆ ที่เกิดจากการวิจัยหรือการทดลองใหม่ ๆ เป็นต้น

2. <u>การกำหนดระดับของความมีนัยสำคัญ (ความน่าจะเป็นของ type I error) α และการคำนวณหา $z(\alpha)$</u>

ระดับของความมีนัยสำคัญนี้ผู้ทดสอบสามารถกำหนดได้เองตามความเหมาะสมโดยส่วนใหญ่แล้วจะกำหนดให้ $\alpha = 0.05$ หรือ $\alpha = 0.01$ ต่อจากนั้นให้คำนวณหาค่าของ $z(\alpha)$ หรือ $z(\alpha/2)$ จากตารางแบบปรกติ ตัวอย่างผลของการคำนวณเป็นดังนี้คือ

ระดับความมีนัยสำคัญ α =	0.05	0.01
$z(\alpha)$	1.645	2.33
$z(\alpha/2)$	1.96	2.58





3. จัดสร้างเงื่อนไขสำหรับการปฏิเสธ null hypothesis $H_0 : \mu = \mu_0$

ให้ $$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

โดยที่ตัวแปรต่าง ๆข้างต้นมีความหมายดังนี้

$\bar{x}$ - เป็นค่าของ sample mean ซึ่งหาได้จากการชักสิ่งตัวอย่าง

μ_0 - เป็นค่าของ population mean ตามความเชื่อที่เคยมีมา (อาจจะไม่ใช่ true population mean ก็ได้) และเป็นตัวที่เราต้องการทดสอบ

σ - เป็นค่าของ true population variance ซึ่งสำหรับกรณีแรกนี้สมมุติว่าทราบค่าแล้ว

n - เป็นขนาดของสิ่งตัวอย่าง

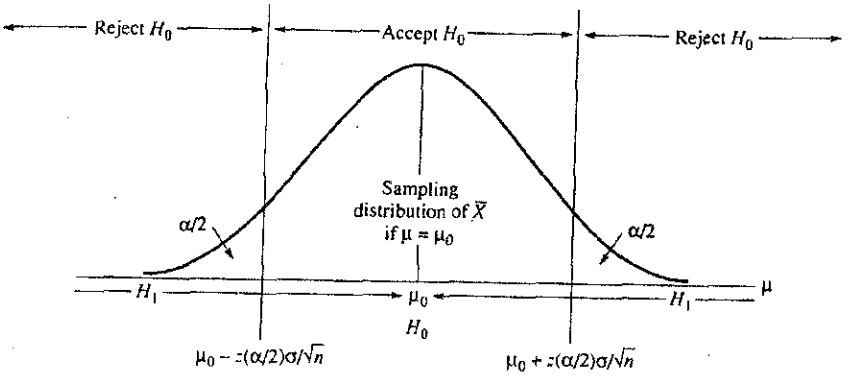
เงื่อนไขสำหรับการปฏิเสธ null hypothesis $H_0 : \mu = \mu_0$ จะขึ้นอยู่กับ alternative hypothesis ด้วย ตามรายละเอียดดังนี้คือ

Alternative hypothesis	Reject null hypohtesis if:
$H_1 : \mu \neq \mu_0$	$z < -z(\alpha/2)$ or $z > z(\alpha/2)$
$H_1 : \mu < \mu_0$	$z < -z(\alpha/2)$
$H_1 : \mu > \mu_0$	$z > z(\alpha/2)$

แนวคิดในการสร้างเงื่อนไขสำหรับการปฏิเสธ null hypothesis $H_0 : \mu = \mu_0$ ก็คือว่าเราจะปฏิเสธ H_0 เมื่อค่า test statistic z ตกอยู่ใน critical region ค่า observed test statistic z ตัวนี้ได้มาจากการชักสิ่งตัวอย่างเชิงสุ่มขนาด n สมมุติว่าเป็น $X_1, X_2, \ldots, X_n$ จากกลุ่มประชากรที่มี mean เท่ากับ μ ซึ่งยังไม่ทราบค่าและมี variance σ^2 ซึ่งเราทราบค่าแล้ว ภายใต้ข้อสมมุติ $H_0 : \mu = \mu_0$ และโดย central limit theorem เราพบว่าถ้า n มีค่ามากพอแล้ว ตัวแปรสุ่ม

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

จะมีการแจกแจงแบบปรกติ N(0, 1) เพราะฉะนั้นเราจะปฏิเสธ H_0 เมื่อค่า observed test statistic $z = (\bar{x} - \mu_0) / (\sigma/\sqrt{n})$ ไปตกอยู่ใน critical region ตามรูปข้างล่างนี้



รูป 5 แสดง critical region สำหรับการทดสอบแบบสองข้าง

4. คำนวณหาค่าของ testy statistic z จากข้อมูลที่กำหนดให้
5. ทำการตัดสินว่าควรจะปฏิเสธ null hypothesis หรือยอมรับ (หรือเก็บไว้ก่อนเพื่อการพิจารณาภายหลัง)

<u>ตัวอย่างที่ 1</u> โรงงานผลิตซีเมนต์บลอคได้แจ้งต่อลูกค้าว่าค่าเฉลี่ยสภาพการนำความร้อนของซีเมนต์บลอคที่ผลิตจากโรงงานนี้มีค่าเท่ากับ 0.340 หน่วย เพื่อที่ตรวจสอบความเชื่อดังกล่าวได้มีการชักสิ่งตัวอย่างมาจำนวน 35 ก้อน และทำการทดสอบ พบว่าได้ค่าเฉลี่ยของตัวอย่างเท่ากับ 0.343 หน่วย จงทำการตัดสินว่าจะเชื่อคำกล่าวอ้างของโรงงานได้หรือไม่? ใช้ระบบความเชื่อมั่นเท่ากับ 95% กำหนดให้ population variance $\sigma^2 = 0.01(0)^2$

<u>วิธีทำ</u> เราจะทำไปตามขั้นตอนที่ได้ให้ไว้ข้างต้นดังนี้

1. <u>จัดตั้งสมมุติฐาน</u>

   null hypohtesis $H_0 : \mu = 0.340$

   Alternative hypothesis $H_0 : \mu \neq 0.340$

2. <u>กำหนดระดับของความมีนัยสำคัญ</u> $\alpha = 0.05$ และดังนั้น

   $z(\alpha/2) = z(0.05/2) = z(0.025) = 1.96$

3. <u>จัดสร้างเงื่อนสำหรับปฏิเสธ</u> H_0

   จะปฏิเสธ H_0 เมื่อ $z < -1.96$ หรือ $z > -1.96$

4. คำนวณหาค่า test statistic z

ในที่นี้ sample mean $\bar{x} = 0.343$

sample size n = 35 $n = 35$

population variance $\sigma^2 = (0.010)^2$

mean ของประชากรที่ต้องการทดสอบ $\mu_0 = 0.340$

เพราะฉะนั้น

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{0343 - 0.340}{0.010/\sqrt{35}} = 1.77$$

5. การตัดสิน

เนื่องจากค่า z = 1.77 ตกอยู่ acceptance region [-1.96, 1.96] ดังนั้นเราจึงจะไม่ปฏิเสธ H_0 ด้วยระดับความเชื่อมั่น 95% [สำหรับคำถามที่ว่าเราจะยอมรับกันอย่างจริงจังหรือไม่ว่าเป็นจริงคำตอบก็คงจะขึ้นอยู่กับสถานการณ์ในขณะนั้นว่าผู้ถามต้องการการตัดสินอย่างไร] #

กรณีที่ 2: การทดสอบเกี่ยวกับ mean เมื่อไม่ทราบค่าของ variance σ^2

สำหรับกรณีนี้ถ้า $n \geq 30$ แล้วเราจะสามารถนำ sample variance s^2 เอาไปแทนที่ σ^2 ในสูตรสำหรับการคำนวณ test statistic ตามขั้นตอนที่ 4 ได้ แต่ถ้าเกิดไม่รู้ค่าของ σ^2 และ n ก็มีค่าน้อยกว่า 30 ด้วย จะใช้การแจกแจงแบบ t เข้าช่วยในการทดสอบ ดังตัวอย่างที่ 3 ที่จะเสนอต่อไป

<u>ตัวอย่างที่ 2</u> บริษัทขนส่งแห่งหนึ่งได้เกิดความสงสัยว่าอายุการใช้งานเฉลี่ยของยางรถยนต์ยี่ห้อหนึ่งจะมีค่าอย่างน้อยเท่ากับ 28,000 ไมล์ จริงหรือไม่? เพื่อที่จะตรวจสอบข้อสงสัยนี้ ทางบริษัทจึงได้สุ่มยางรถยนต์มา 40 เส้น แล้วใส่ยางดังกล่าวเข้ากับรถบรรทุกของบริษัทและหาอายุการใช้งานเฉลี่ยได้เท่ากับ 27,463 ไมล์ และมีค่า standard deviation เท่ากับ 1,348 ไมล์ เราจะสรุปอะไรได้จากการทดสอบนี้กำหนดให้ค่า $\alpha \leq 0.01$

วิธีทำ

1. ตั้งสมมุติฐาน

   null hypothesis $H_0 : \mu \geq 28{,}000$ ไมล์

   alternative hypothesis $H_0 : \mu < 28{,}000$ ไมล์

2. กำหนดระดับของความมีนัยสำคัญ : $\alpha \leq 0.01$, $z(0.01) = 2.33$

3. กำหนดเงื่อนไขของการปฏิเสธ $H_0 : \mu \geq 28{,}000$ ไมล์ :
   ใช้เงื่อนไขเดียวกันกับการทดสอบ

   $H_0' : \mu = 28{,}000$ ไมล์

   $H_1' : \mu < 28{,}000$ ไมล์

   นั่นคือเงื่อนไขสำหรับการปฏิเสธ $H_0 : \mu \geq 28{,}000$ ไมล์ ก็คือ $z < -2.33$ เมื่อ

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

   [ในที่นี้เราไม่รู้ค่าของ population standard deviation σ จึงใช้ sample standard deviation S แทน]

4. คำนวณค่าของ test statisitc z จากข้อมูลที่โจทย์กำหนด

$$z = \frac{27{,}463 - 28{,}000}{1{,}348/\sqrt{40}} = -2.52$$

5. การตัดสิน

   เนื่องจาก $z = -2.52 < -2.33$ จึงทำให้ค่า z ตกอยู่ใน critical region เราจึงจะปฏิเสธ H_0 ไปด้วยความเชื่อมั่นเท่ากับ 99% พูดอีกอย่างหนึ่งก็คือบริษัทขนส่งเชื่อว่า $H_1 : \mu < 28{,}000$ ไมล์

สำหรับกรณีที่เราไม่ทราบค่าของ σ และ sample size ก็มีขนาดน้อยกว่า 30 ด้วย ต้องมีข้อสมมุติเพิ่มเติมนิดหน่อยว่าเราได้ชักสิ่งตัวอย่างมาจากประชากรซึ่งมีการแจกแจงแบบปรกติ ในกรณีนี้การทดสอบ null hypothesis $H_0 : \mu = \mu_0$ จะวางอยู่บนพื้นฐานของตัว statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

ซึ่งมาจาก ตัวแปรสุ่มที่มีการแจกแจงแบบ t พร้อมด้วย (n - 1) degree of freedom. ส่วนการกำหนดเงื่อนไขสำหรับการปฏิเสธ H_0 เป็นเช่นเดียวกันกับ z เพียงแต่เปลี่ยนตัวแปร z เป็น t และเปลี่ยน $z(\alpha)$ และ $z(\alpha/2)$ เป็น $t(\alpha; n-1)$ และ $t(\alpha/2; n-1)$ ตามลำดับ

<u>ตัวอย่างที่ 3</u> ข้อกำหนดของ ribbon ชนิดหนึ่งก็คือจะต้องมี mean breaking strength เท่ากับ 180 ปอนด์ มีการหยิบ ribbon อย่างอย่างสุ่มมา 5 ชิ้น เพื่อทำการทดสอบ พบว่ามี mean breaking strength เท่ากับ 169.5 ปอนด์ และมี standard deviation เท่ากับ 5.7 ปอนด์ จงทดสอบสมมุติฐาน $\mu = 180$ เทียบกับสมมุติฐาน $\mu < 180$ โดยใช้ระดับของความมีนัยสำคัญ เท่ากับ 0.01 สมมุติว่าการแจกแจงของประชากรเป็นแบบปรกติ

<u>วิธีทำ</u>

1. <u>ตั้งสมมุติฐาน</u>

   null hypothesis $H_0 : \mu = 180$ ปอนด์

   Alternative hypohtesis $H_1 : \mu < 180$ ปอนด์

2. <u>ระดับของความมีนัยสำคัญ</u> : $\alpha = 0.01$; $t(0.01, 5 - 1) = 3.747$

3. <u>เงื่อนไขในการปฏิเสธ</u> H_0 ก็คือ $t < -3.747$ เมื่อ test statistic t มีค่าเท่ากับ

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

4. <u>คำนวณหาค่าของ t จากสูตรในข้อ 3</u>

$$t = \frac{169.5 - 180}{5.71\sqrt{5}} = -4.12$$

5. <u>การตัดสิน</u>

   เนื่องจาก $t = -4.12 < -3.747$ ดังนั้นเราจะปฏิเสธ H_0 ด้วยระดับของความมีนัยสำคัญเท่ากับ 0.01 พูดอีกอย่างหนึ่งก็คือ เราจะยอมรับ H_1 ที่ว่า mean breaking strength ของ ribbon มีค่าน้อยกว่า 180 ปอนด์ #

แบบฝึกหัด

1. According to the norms established for a mechanical aptitude test, persons who are 18 years old should average 73.2 with a standard deviation of 8.6. If 45 randomly selected persons of that age averaged 76.7, test null hypothesis $\mu = 73.2$ against the alternative hypothesis $\mu > 73.2$ at the 0.01 level of significance. **Ans** $z = 2.73$; reject H_0.

2. In 64 randomly selected hours of production the mean and the standard deviation of the number of acceptable pieces produced by an automatic stamping machine are $\bar{x} = 1{,}038$ and $s = 146$. At 0.05 level of significance does this enable us to reject the null hypothesis $\mu = 1{,}000$ against the alternature hypothesis $\mu = 1{,}000$? **Ans** $z = 2.08$; reject H_0.

3. In a labor - managemant discussion it was brought up that workers at a certain large plant take on the average 32.6 minutes to get work. If a random sample of 60 workers took on the average 33.8 minutes with a standard deviation of 6.1 minutes, can we reject the null hypothesis $\mu = 32.6$ against the alternative hypothesis $\mu > 32.6$ at 0.05 level of significance? **Ans** $z = 1.52$; cannot reject H_0.

4. Given a random sample of 5 pints from different production lots, we want to test whether the fat content of a certain kind of ice cream exceeds 14%. What can we conclude at the 0.01 level of significance about the null hypothasis $\mu = 14\%$ if the sample has the mean $\bar{x} = 14.9\%$ and the standard deviation $s = 0.42\%$. **Ans** $t = 4.79$; reject H_0.

5. A random sample from a company's very extensive files show that orders for a certain piece of machinery were filled, respectively in 10, 12, 19, 14, 15, 18, 11 and 13 days. Use the level of significance $\alpha = 0.01$ to test the claim that on the average such orders are filled in 10.5 days. Choose the alternative hypothesis so that rejection of the null hypothesis $\mu = 10.5$ implies that it takes longer than indicated. Assume normality. **Ans** $t = 3.087$; reject H_0.

8 Inferences Concerning Variances

In Chapter 7 we learned how to judge the size of the error in estimating a population mean, how to construct confidence intervals for means, and how to perform tests of hypotheses about the means of one and of two populations. As we shall see in this and in subsequent chapters, very similar methods apply to inferences about other population parameters.

In this chapter we shall concentrate on population variances, or standard deviations, which are not only important in their own right, but which must sometimes be estimated before inferences about other parameters can be made. Section 8.1 is devoted to the estimation of σ^2 and σ, and Sections 8.2 and 8.3 deal with tests of hypotheses about these parameters.

8.1 THE ESTIMATION OF VARIANCES

In the preceding chapter, there were several instances where we estimated a population standard deviation by means of a sample standard deviation—we substituted s for σ in the large sample confidence interval for μ on page 222, in the large sample test concerning μ on page 240, and in the large sample test concerning the difference between two means on page 253. There are many statistical procedures in which s is thus substituted for σ, or s^2 for σ^2. There are also situations where σ is the primary parameter of interest.

Let $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ be the sample variance based on a random sample from any population, discrete or continuous, having variance σ^2. It follows from the Example on page 179 that the mean of the sampling distribution of S^2 is given by σ^2.

Unbiased estimation of population variance

The sample variance

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \text{ is an unbiased estimator of } \sigma^2$$

Although the sample variance is an unbiased estimator of σ^2, it does not follow that the sample standard deviation is also an unbiased estimator of σ; in fact, it is not. However, for large samples the bias is small and it is common practice to estimate σ with s.

Besides s, population standard deviations are sometimes estimated in terms of the **sample range** R, which we defined in Section 2.6 as the largest value of a sample minus the smallest. Given a random sample of size n from a normal population, it can be shown that the sampling distribution of the range has the mean $d_2\sigma$ and the standard deviation $d_3\sigma$, where d_2 and d_3 are constants which depend on the size of the sample. For $n = 1, 2, \ldots$, and 10, their values are as shown in the following table:

n	2	3	4	5	6	7	8	9	10
d_2	1.128	1.693	2.059	2.326	2.534	2.704	2.847	2.970	3.078
d_3	0.853	0.888	0.880	0.864	0.848	0.833	0.820	0.808	0.797

Thus, R/d_2 is an unbiased estimate of σ, and for very small samples, $n \leq 5$, it provides nearly as good an estimate of σ as does s; as the sample size increases, it becomes more efficient to use s instead of R/d_2. Nowadays, the range is used to estimate σ primarily in problems of industrial quality control, where sample sizes are usually small and computational ease is of prime concern. This application will be discussed in Chapter 14, where we shall need the above values of the constant d_3.

EXAMPLE With reference to the example on page 256, use the range of the first sample to estimate σ for the heat-producing capacity of coal from the first mine.

Solution Since the smallest value is 8,070, the largest value is 8,350, and $n = 5$ so that $d_2 = 2.326$, we get

$$\frac{R}{d_2} = \frac{8{,}350 - 8{,}070}{2.326} = 120.4$$

Note that this is fairly close to the sample standard deviation $s = 125.5$.

■

In most practical applications, interval estimates of σ or σ^2 are based on the sample standard deviation or the sample variance. For random samples from normal populations, we make use of Theorem 6.4, according to which

$$\frac{(n-1)S^2}{\sigma^2}$$

is a random variable having the chi-square distribution with $n-1$ degrees of freedom. Thus, with χ^2_α defined as on page 211 for a chi-square distribution with $n-1$ degrees of freedom, we can assert with probability $1-\alpha$ that the inequality

$$\chi^2_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}$$

will be satisfied; once the data have been obtained, we make the same assertion with $(1-\alpha)100\%$ confidence. Solving this inequality for σ^2, we obtain the following result:

Confidence interval for σ^2

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

If we take the square root of each member of this inequality, we obtain a corresponding $(1-\alpha)100\%$ confidence interval for σ.

Note that confidence intervals for σ or σ^2 obtained by taking "equal tails," as in the above formula, do not actually give the narrowest confidence intervals, because the chi-square distribution is not symmetrical (see Exercise 7.16). Nevertheless, they are used in most applications in order to avoid fairly complicated calculations.

EXAMPLE Returning to the example on page 211, suppose that the refractive indices of 20 pieces of glass (randomly selected from a large shipment purchased by the optical firm) have a variance of $1.20 \cdot 10^{-4}$. Construct a 95% confidence interval for σ, the standard deviation of the population sampled.

Solution For $20 - 1 = 19$ degrees of freedom, $\chi^2_{0.975} = 8.907$ and $\chi^2_{0.025} = 32.852$ according to Table 5, so that substitution into the formula yields

$$\frac{(19)(1.20 \cdot 10^{-4})}{32.852} < \sigma^2 < \frac{(19)(1.20 \cdot 10^{-4})}{8.907}$$

$$0.000069 < \sigma^2 < 0.000256$$

and, hence,

$$0.0083 < \sigma < 0.0160$$

This means we are 95% confident that the interval from 0.0083 to 0.0160 contains σ, the true standard deviation of the refractive index.

■

The method which we have discussed applies only to random samples from normal populations (or at least to random samples from populations which can be approximated closely with normal distributions).

EXERCISES

8.1 Use the data of Exercise 7.46 on page 250 to estimate σ for the length of time the experimental engine will operate with the given fuel in terms of

(a) the sample standard deviation; ($s = 4.145$)

(b) the sample range. (4.341)

Compare the two estimates by expressing their difference as a percentage of the first.

8.2 With reference to the example on page 256, use the range of the second sample to estimate σ for the heat-producing capacity of coal from the second mine, and compare the result with the standard deviation of the second sample.

8.3 Use the data of part (a) of Exercise 7.70 to estimate σ for the Brinell hardness of Alloy 1 in terms of

(a) the sample standard deviation;

(b) the sample range.

Compare the two estimates by expressing their difference as a percentage of the first.

8.4 With reference to Exercise 7.47, construct a 99% confidence interval for the variance of the amount of time it takes the company to fill an order for a piece of the given kind of machinery.

8.5 With reference to Exercise 7.48, construct a 99% confidence interval for the variance of the population sampled.

8.6 Use the value of s obtained in Exercise 8.3 to construct a 98% confidence interval for σ, measuring the actual variability in the hardness of Alloy 1.

8.2 HYPOTHESES CONCERNING ONE VARIANCE

In this section we shall consider the problem of testing the null hypothesis that a population variance equals a specified constant against a suitable one-sided or two-sided alternative; that is, we shall test the null hypothesis $\sigma^2 = \sigma_0^2$ against one of the alternatives $\sigma^2 < \sigma_0^2$, $\sigma^2 > \sigma_0^2$, or $\sigma^2 \neq \sigma_0^2$. Tests like these are important whenever it is desired to control the uniformity of a product or an operation. For example, suppose that a silicon disc, or "wafer," is to be cut into small squares, or "dice," to be used

in the manufacture of a semiconductor device. Since certain electrical characteristics of the finished device may depend on the thickness of the die, it is important that all dice cut from a wafer have approximately the same thickness. Thus, not only must the mean thickness of a wafer be kept within specifications, but also the variation in thickness from location to location on the wafer.

Using the same sampling theory as on page 270, we base such tests on the fact that for random samples from a normal population with the variance σ_0^2

Statistic for test concerning variance

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

is a random variable having the chi-square distribution with $n-1$ degrees of freedom. The critical regions for such tests are as shown in the following table:

Critical Regions for Testing $\sigma^2 = \sigma_0^2$
(Normal population)

Alternative hypothesis	*Reject null hypothesis if:*
$\sigma^2 < \sigma_0^2$	$\chi^2 < \chi^2_{1-\alpha}$
$\sigma^2 > \sigma_0^2$	$\chi^2 > \chi^2_{\alpha}$
$\sigma^2 \neq \sigma_0^2$	$\chi^2 < \chi^2_{1-\alpha/2}$ or $\chi^2 > \chi^2_{\alpha/2}$

In this table χ^2_α is as defined on page 211. Note that "equal tails" are used for the two-sided alternative, and this is actually not the best procedure since the chi-square distribution is not symmetrical.

EXAMPLE The lapping process which is used to grind certain silicon wafers to the proper thickness is acceptable only if σ, the population standard deviation of the thickness of dice cut from the wafers, is at most 0.50 mil. Use the 0.05 level of significance to test the null hypothesis $\sigma = 0.50$ against the alternative hypothesis $\sigma > 0.50$, if the thicknesses of 15 dice cut from such wafers have a standard deviation of 0.64 mil.

Solution

1. *Null hypothesis*: $\sigma = 0.50$
   Alternative hypothesis: $\sigma > 0.50$
2. *Level of significance*: $\alpha = 0.05$

3. *Criterion*: Reject the null hypothesis if $\chi^2 > 23.685$, the value of $\chi^2_{0.05}$ for 14 degrees of freedom, where

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

4. *Calculations*:

$$\chi^2 = \frac{(15-1)(0.64)^2}{(0.50)^2} = 22.94$$

5. *Decision*: Since $\chi^2 = 22.94$ does not exceed 23.685, the null hypothesis cannot be rejected; even though the sample standard deviation exceeds 0.50, there is not sufficient evidence to conclude that the lapping process is unsatisfactory. ■

There exist tables similar to Table 8, which enable us to read the probabilities of Type II errors connected with this kind of test. As given in the *National Bureau of Standards Handbook 91* (see the bibliography), they contain the *OC* curves for the different one-sided and two-sided alternatives, for $\alpha = 0.05$ and $\alpha = 0.01$, and for various values of n.

8.3 HYPOTHESES CONCERNING TWO VARIANCES

The two-sample t test, described in Section 7.9, requires that the variances of the two populations sampled are equal. In this section we describe a test of the null hypothesis $\sigma_1^2 = \sigma_2^2$, which applies to independent random samples from two normal populations; it must be used with some discretion as it is very sensitive to departures from this assumption.

If independent random samples of size n_1 and n_2 are taken from normal populations having the same variance, it follows from Theorem 6.5 that

Statistic for test of equality of two variances

$$F = \frac{S_1^2}{S_2^2}$$

is a random variable having the F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. Thus, if the null hypothesis $\sigma_1^2 = \sigma_2^2$ is true, the ratio of the sample variances S_1^2 and S_2^2 provides a statistic on which tests of the null hypothesis can be based.

The critical region for testing the null hypothesis $\sigma_1^2 = \sigma_2^2$ against the alternative hypothesis $\sigma_1^2 > \sigma_2^2$ is $F > F_\alpha$, where F_α is as defined on page 212. Similarly,

the critical region for testing the null hypothesis against the alternative hypothesis $\sigma_1^2 < \sigma_2^2$ is $F < F_{1-\alpha}$, and this causes some difficulties since Table 6 only contains values corresponding to right-hand tails of $\alpha = 0.05$ and $\alpha = 0.01$. As a result, we use the reciprocal of the original test statistic and make use of the relation

$$F_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{F_\alpha(\nu_2, \nu_1)}$$

first given on page 213. Thus, we base the test on the statistic $F = S_2^2/S_1^2$ and the critical region for testing the null hypothesis $\sigma_1^2 = \sigma_2^2$ against the alternative hypothesis $\sigma_1^2 < \sigma_2^2$ becomes $F > F_\alpha$, where F_α is the appropriate critical value of F for $n_2 - 1$ and $n_1 - 1$ degrees of freedom.

For the two-sided alternative $\sigma_1^2 \neq \sigma_2^2$ the critical region is $F < F_{1-\alpha/2}$ or $F > F_{\alpha/2}$, where $F = S_1^2/S_2^2$ and the degrees of freedom are $n_1 - 1$ and $n_2 - 1$. In practice, we modify this test as in the preceding paragraph, so that we can again use the table of F values corresponding to right-hand tails of $\alpha = 0.05$ and $\alpha = 0.01$. To this end we let S_M^2 represent the larger of the two sample variances, S_m^2 the smaller, and we write the corresponding sample sizes as n_M and n_m. Thus, the test statistic becomes $F = S_M^2/S_m^2$ and the critical region is as shown in the following table:

Critical Regions for Testing $\sigma_1^2 = \sigma_2^2$
(Normal populations)

Alternative hypothesis	*Test statistic*	*Reject null hypothesis if:*
$\sigma_1^2 < \sigma_2^2$	$F = \dfrac{S_2^2}{S_1^2}$	$F > F_\alpha(n_2 - 1, n_1 - 1)$
$\sigma_1^2 > \sigma_2^2$	$F = \dfrac{S_1^2}{S_2^2}$	$F > F_\alpha(n_1 - 1, n_2 - 1)$
$\sigma_1^2 \neq \sigma_2^2$	$F = \dfrac{S_M^2}{S_m^2}$	$F > F_{\alpha/2}(n_M - 1, n_m - 1)$

The level of significance of these tests is α and the figures indicated in parentheses are the respective degrees of freedom. Note that, as in the chi-square test, "equal tails" are used in the two-tailed test as a matter of mathematical convenience, even though the F distribution is not symmetrical.

EXAMPLE It is desired to determine whether there is less variability in the silver plating done by Company 1 than in that done by Company 2. If independent random samples of size 12 of the two companies' work yield $s_1 = 0.035$ mil and $s_2 = 0.062$ mil, test the

null hypothesis $\sigma_1^2 = \sigma_2^2$ against the alternative hypothesis $\sigma_1^2 < \sigma_2^2$ at the 0.05 level of significance.

Solution

1. *Null hypothesis*: $\sigma_1^2 = \sigma_2^2$
   Alternative hypothesis: $\sigma_1^2 < \sigma_2^2$
2. *Level of significance*: $\alpha = 0.05$
3. *Criterion*: Reject the null hypothesis if $F > 2.82$, and the value of $F_{0.05}$ for 11 and 11 degrees of freedom, where

$$F = \frac{S_2^2}{S_1^2}$$

4. *Calculations*:

$$F = \frac{(0.062)^2}{(0.035)^2} = 3.14$$

5. *Decision*: Since $F = 3.14$ exceeds 2.82, the null hypothesis must be rejected; in other words, the data support the contention that the plating done by Company 1 is less variable than that done by Company 2.

EXAMPLE With reference to the example dealing with the heat-producing capacity of coal from two mines on page 256, use the 0.02 level of significance to test whether it is reasonable to assume that the variances of the two populations sampled are equal.

Solution

1. *Null hypothesis*: $\sigma_1^2 = \sigma_2^2$
   Alternative hypothesis: $\sigma_1^2 \neq \sigma_2^2$
2. *Level of significance*: $\alpha = 0.02$
3. *Criterion*: Reject the null hypothesis if $F > 11.4$, the value of $F_{0.01}$ for 4 and 5 degrees of freedom, where the value of F is

$$F = \frac{S_1^2}{S_2^2}$$

since $s_1^2 = 15{,}750$ is greater than $s_2^2 = 10{,}920$.

4. *Calculations*:

$$F = \frac{15{,}750}{10{,}920} = 1.44$$

5. *Decision*: Since $F = 1.44$ does not exceed 11.4, the null hypothesis cannot be rejected; there is no real reason to doubt the equality of the variances of the two populations.

■

Had we wanted to use the level of significance $\alpha = 0.05$ or $\alpha = 0.01$ in this example, we would have required tables of the values of $F_{0.025}(\nu_1, \nu_2)$ or $F_{0.005}(\nu_1, \nu_2)$; such tables may be found in the *Biometrika Tables for Statisticians* listed in the Bibliography under Pearson and Hartley. Also, *OC* curves for the one-tailed F tests may be found in the *National Bureau of Standards Handbook 91*.

Caution: In marked contrast to the procedures for making inferences about μ, the validity of the procedures in this chapter depend rather strongly on the assumption that the underlying population is normal. The sampling variance of S^2 can change when the population departs from normality by having, for instance, a single long tail. It can be shown that, when the underlying population is normal, the sampling variance of S^2 is $2\sigma^4/(n-1)$. However, for nonnormal distributions, the sampling variance of S^2 depends not only on σ^2 but also on the population third and fourth moments, μ_3 and μ_4 (see page 143). Consequently, it could be much larger than $2\sigma^4/(n-1)$. This behavior completely invalidates any tests of hypothesis or confidence intervals for σ^2. We say that these procedures for making inferences about σ^2 are not **robust** with respect to deviations from normality.

EXERCISES

8.7 With reference to Exercise 7.44 on page 250, test the null hypothesis $\sigma = 600$ psi for the compressive strength of the given kind of steel against the alternative hypothesis $\sigma > 600$ psi. Use the 0.05 level of significance. ($\chi^2 = 5.832$; cannot reject H_0)

8.8 If 12 determinations of the specific heat of iron have a standard deviation of 0.0086, test the null hypothesis that $\sigma = 0.010$ for such determinations. Use the alternative hypothesis $\sigma \neq 0.010$ and the level of significance $\alpha = 0.01$.

8.9 With reference to Exercise 7.64, test the null hypothesis that $\sigma = 15.0$ minutes for the time that is required for repairs of the first kind of photocopying equipment against the alternative hypothesis that $\sigma > 15.0$ minutes. Use the 0.05 level of significance and assume normality. ($\chi^2 = 115.88$; reject H_0)

8.10 Use the 0.01 level of significance to test the null hypothesis that $\sigma = 0.015$ inch for the diameters of certain bolts against the alternative hypothesis that $\sigma \neq 0.015$ inch, given that a random sample of size 15 yielded $s^2 = 0.00011$.

8.11 Playing 10 rounds of golf on his home course, a golf professional averaged 71.3 with a standard deviation of 1.32. Test the null hypothesis that the consistency of his game on his home course is actually measured by $\sigma = 1.20$, against the alternative hypothesis that he is less consistent. Use the level of significance $\alpha = 0.05$. ($\chi^2 = 10.89$; not reject H_0)

8.12 The security department of a large office building wants to test the null hypothesis that $\sigma = 2.0$ minutes for the time it takes a guard to walk his round against the alternative hypothesis that $\sigma \neq 2.0$ minutes. What can it conclude at the 0.01 level of significance if a random sample of size $n = 31$ yields $s = 1.8$ minutes?

8.13 Justify the use of the two-sample t test in Exercise 7.67 on page 261 by testing the null hypothesis that the two populations have equal variances. Use the 0.02 level of significance. ($F = 1.496$; we cannot reject H_0)

8.14 With reference to Exercise 7.68 on page 261, use the 0.02 level of significance to test the assumption that the two populations have equal variances.

8.15 Two different lighting techniques are compared by measuring the intensity of light at selected locations in areas lighted by the two methods. If 15 measurements in the first area had a standard deviation of 2.7 foot-candles and 21 measurements in the second area had a standard deviation of 4.2 foot-candles, can it be concluded that the lighting in the second area is less uniform? Use a 0.01 level of significance. What assumptions must be made as to how the two samples are obtained? ($F = 2.42$ we cannot reject H_0)

8.16 With reference to Exercise 7.66, where we had $n_1 = 40$, $n_2 = 30$, $s_1 = 15.2$, and $s_2 = 18.7$, use the 0.05 level of significance to test the claim that there is a greater variability in the number of cars which make left turns approaching from the south between 4 P.M. and 6 P.M. at the second intersection. Assume the distributions are normal.

8.17 Random samples of size n_1 and n_2, respectively, are taken from two log-normal populations, and the resulting sample means are $\bar{x}_1 = 3.74$ and $\bar{x}_2 = 13.91$ and the sample variances are 1.2 and 9.5. You wish to test whether the second population has a mean value four times as large as the first.

(a) Can you directly use a two-sample to test? Why?

(b) Is there a transformation that can be made on the data that could conceivably allow the use of a two-sample test?

8.4 REVIEW EXERCISES

8.18 With reference to the example on page 243, construct a 95% confidence interval for the true standard deviation of the breaking strength of the given kind of ribbon.

8.19 With reference to the example on page 257, find separate 95% confidence intervals for the standard deviations of the two aluminum alloys.

8.20 While performing a strenuous task, the pulse rate of 25 workers increased on the average by 18.4 beats per minute with a standard deviation of 4.9 beats per minute. Find a 95% confidence interval for the corresponding population standard deviation. What assumption did you make about the population?

8.21 With reference to Exercise 8.20, use the 0.05 level of significance to test the null hypothesis that $\sigma^2 = 30.0$ for such increases in the pulse rate (while performing the given task) against the alternative hypothesis that $\sigma^2 < 30.0$.

8.22 If 31 measurements of the boiling point of sulfur have a standard deviation of 0.83 degree Celsius, construct a 98% confidence interval for the true standard deviation of such measurements. What assumption did you make about the population?

8.23 Past data indicate that the variance of measurements made on sheet metal stampings by experienced quality control inspectors is 0.18 square inch. Such measurements made by an inexperienced inspector could have too large a variance (perhaps because of inability to read instruments properly) or too small a variance (perhaps because unusually high or low measurements are discarded). If a new inspector measures 101 stampings with a variance of 0.13 square inch, test at the 0.05 level of significance whether the inspector is making satisfactory measurements. Assume normality.

8.24 With reference to Exercise 7.69 on page 261, use the 0.02 level of significance to test the assumption that the two populations have equal variances.

8.25 Pull-strength tests on 10 soldered leads for a semiconductor device yield the following results in pounds force required to rupture the bond:

15.8, 12.7, 13.2, 16.9, 10.6, 18.8, 11.1, 14.3, 17.0, 12.5

Another set of eight leads was tested after encapsulation to determine whether the pull strength has been increased by encapsulation of the device, with the following results:

24.9, 23.6, 19.8, 22.1, 20.4, 21.6, 21.8, 22.5

As a preliminary to the two-sample t test, use the 0.02 level of significance to test whether it is reasonable to assume that the two samples come from populations with equal variances.

8.26 With reference to the example on page 257, test the equality of the variances for the two aluminum alloys. Use the 0.02 level of significance.

8.5
CHECKLIST OF KEY TERMS (with page references)

Robust 276 *Sample range 269*

9 Inferences Concerning Proportions

Many engineering problems deal with proportions, percentages, or probabilities. In acceptance sampling we are concerned with the proportion of defectives in a lot, and in life testing we are concerned with the percentage of certain components which will perform satisfactorily during a stated period of time, or the probability that a given component will last at least a given number of hours. It should be clear from these examples that problems concerning proportions, percentages, or probabilities are really equivalent; a percentage is merely a proportion multiplied by 100, and a probability may be interpreted as a proportion in a long series of trials.

Sections 9.1 and 9.2 deal with the estimation of proportions; Section 9.3 deals with tests concerning proportions; Section 9.4 deals with tests concerning two or more proportions; in Section 9.5 we shall learn how to analyze data tallied into a two-way classification; and in Section 9.6 we shall learn how to judge whether differences between an observed frequency distribution and corresponding expectations can be attributed to chance.

9.1 ESTIMATION OF PROPORTIONS

The information that is usually available for the estimation of a proportion is the number of times, X, that an appropriate event occurs in n trials, occasions, or ob-

servations. The point estimator, itself, is usually the **sample proportion** $\frac{X}{n}$, namely, the proportion of the time that the event actually occurs. If the n trials satisfy the assumptions underlying the binomial distribution listed on page 95, we know that the mean and the standard deviation of the number of successes are given by np and $\sqrt{np(1-p)}$. If we divide both of these quantities by n, we find that the mean and the standard deviation of the proportion of successes (namely, of the sample proportion) are given by

$$\frac{np}{n} = p \quad \text{and} \quad \frac{\sqrt{np(1-p)}}{n} = \sqrt{\frac{p(1-p)}{n}}$$

The first of these results shows that the sample proportion is an unbiased estimator of the binomial parameter p, namely, of the true proportion we are trying to estimate on the basis of a sample.

In the construction of confidence intervals for the binomial parameter p, we meet several obstacles. First, since x and $\frac{x}{n}$ are values of discrete random variables, it may be impossible to get an interval for which the degree of confidence is exactly $(1-\alpha)100\%$. Second, the standard deviation of the sampling distribution of the number of successes, as well as that of the proportion of successes, involves the parameter p that we are trying to estimate.

To construct a confidence interval for p having approximately the degree of confidence $(1-\alpha)100\%$, we first determine for a given set of values of p the corresponding quantities x_0 and x_1, where x_0 is the largest integer for which the binomial probabilities $b(k; n, p) = P[X = k]$ satisfy

$$\sum_{k=0}^{x_0} b(k; n, p) \leq \frac{\alpha}{2}$$

while x_1 is the smallest integer for which

$$\sum_{k=x_1}^{n} b(k; n, p) \leq \frac{\alpha}{2}$$

To emphasize the point that x_0 and x_1 depend on the value of p, we shall write these quantities as $x_0(p)$ and $x_1(p)$. Thus, we can assert with a probability of approximately $1-\alpha$, and at least $1-\alpha$, that the inequality

$$x_0(p) < x < x_1(p)$$

will be satisfied; here x is a value of a random variable and p is a fixed constant. To change inequalities like these into confidence intervals for p, we can use a simple graphical method which is illustrated by the following example: Suppose, for instance,

that we want to find approximate 95% confidence intervals for p for samples of size $n = 20$. Using Table I at the end of the book, we first determine x_0 and x_1 for selected values of p such that x_0 is the largest integer for which

$$B(x_0; 20, p) \leq 0.025$$

while x_1 is the smallest integer for which

$$1 - B(x_1 - 1; 20, p) \leq 0.025$$

Letting p equal 0.1, 0.2, . . . , and 0.9, we thus obtain the values shown in the following table:

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
x_0	—	0	1	3	5	7	9	11	14
x_1	6	9	11	13	15	17	19	20	—

Plotting the points with coordinates p and $x(p)$ as in Figure 9.1, and drawing smooth curves, one through the x_0 points and one through the x_1 points, we can now "solve" for p. For any given value of x we can obtain approximate 95% confidence limits for p by going horizontally to the two curves and marking off the corresponding values of p (see Figure 9.1). Thus, for $x = 4$ we obtain the approximate 95% confidence interval

$$0.06 < p < 0.45$$

Graphs similar to the one shown in Figure 9.1 are given in Tables 9(a) and 9(b) at the end of the book for various values of n and for the 95% and 99% degrees of confidence. These tables differ from the one of Figure 9.1 in that the sample proportion $\frac{x}{n}$ is used instead of x, thus making it possible to graph curves corresponding to various values of n on the same diagram. Also, for increased accuracy, Tables 9(a) and 9(b) are arranged so that values of $\frac{x}{n}$ from 0.00 to 0.50 are marked on the bottom scale while those from 0.50 to 1.00 are marked on the top scale of the diagram. For values of $\frac{x}{n}$ from 0.00 to 0.50 the confidence limits for p are read off the left-hand scale of the diagram, while for values of $\frac{x}{n}$ from 0.50 to 1.00 they are read off the right-hand scale. Note that for $n = 20$ and $x = 4$, Table 9(a) yields the 95% confidence interval $0.06 < p < 0.44$, which is very close, indeed, to the results obtained with Figure 9.1.

On page 151 we gave the general rule of thumb that the normal distribution provides a good approximation to the binomial distribution when np and $n(1 - p)$ are

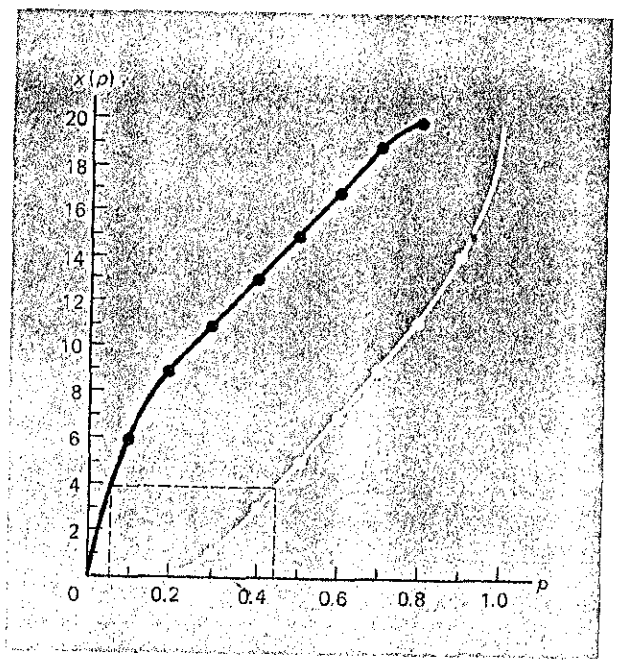


FIGURE 9.1
95% confidence intervals for proportions ($n = 20$).

both greater than 15. Thus, for $n = 50$ the normal curve approximation may be used if it can be assumed that p lies between 0.30 and 0.70; for $n = 100$ it may be used if it can be assumed that p lies between 0.15 and 0.85; for $n = 200$ it may be used if it can be assumed that p lies between 0.075 and 0.925; and so forth. This is what we shall mean here, and later in this chapter, by "n being large."

When n is large, we can construct approximate confidence intervals for the binomial parameter p by using the normal approximation to the binomial distribution. Accordingly, we can assert with probability $1 - \alpha$ that the inequality

$$-z_{\alpha/2} < \frac{X - np}{\sqrt{np(1-p)}} < z_{\alpha/2}$$

will be satisfied. Solving this quadratic inequality for p, we can obtain a corresponding set of approximate confidence limits for p in terms of the observed value x (see Exercise 9.14 on page 289), but since the necessary calculations are involved, we shall make the further approximation of substituting $\frac{x}{n}$ for p in $\sqrt{np(1-p)}$. This yields

Large sample confidence interval for p

$$\frac{x}{n} - z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}} < p < \frac{x}{n} + z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}}$$

where the degree of confidence is $(1 - \alpha)100\%$.

EXAMPLE If $x = 36$ of $n = 100$ persons interviewed are familiar with the tax incentives for installing certain energy-saving devices, construct a 95% confidence interval for the corresponding true proportion.

Solution Substituting $\frac{x}{n} = \frac{36}{100} = 0.36$ and $z_{\alpha/2} = 1.96$ into the above formula, we get

$$0.36 - 1.96\sqrt{\frac{(0.36)(0.64)}{100}} < p < 0.36 + 1.96\sqrt{\frac{(0.36)(0.64)}{100}}$$

or

$$0.266 < p < 0.454$$

We are 95% confident that the population proportion of persons familiar with the tax incentives, p, is contained in the interval from 0.266 to 0.454. Note that if we had used Table 9(a), we would have obtained

$$0.27 < p < 0.46$$

■

The magnitude of the error we make when we use $\frac{X}{n}$ as an estimator of p is given by $\left|\frac{X}{n} - p\right|$. Again using the normal approximation, we can thus assert with probability $1 - \alpha$ that the inequality

$$\left|\frac{X}{n} - p\right| \leq z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

will be satisfied, namely, that the error will be at most $z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$.

Maximum error estimate

$$E = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

With the observed value $\frac{x}{n}$ substituted for p we obtain an estimate of E.

EXAMPLE In a sample survey conducted in a large city, 136 of 400 persons answered Yes to the question whether their city's public transportation is adequate. With 99% confidence,

what can we say about the maximum error, if $\frac{x}{n} = \frac{136}{400} = 0.34$ is used as an estimate of the corresponding true proportion?

Solution Substituting $\frac{x}{n} = 0.34$ and $z_{\alpha/2} = 2.575$ into the above formula, we find that the error is at most

$$E = 2.575\sqrt{\frac{(0.34)(0.66)}{400}} = 0.061$$

■

The preceding formula for E can also be used to determine the sample size that is needed to attain a desired degree of precision. Solving for n, we get

Sample size

$$n = p(1-p)\left[\frac{z_{\alpha/2}}{E}\right]^2$$

but this formula cannot be used as it stands unless we have some information about the possible size of p (on the basis of collateral data, say, a pilot sample). If no such information is available, we can make use of the fact that $p(1-p)$ is at most $\frac{1}{4}$, corresponding to $p = \frac{1}{2}$, as can be shown by the methods of elementary calculus. Thus, if

Sample size

$$n = \frac{1}{4}\left[\frac{z_{\alpha/2}}{E}\right]^2$$

we can assert with a probability of at least $1-\alpha$ that the error in using $\frac{X}{n}$ as an estimate of p will not exceed E; once the data have been obtained, we will be able to assert with at least $(1-\alpha)100\%$ confidence that the error does not exceed E.

EXAMPLE Suppose that we want to estimate the true proportion of defectives in a very large shipment of adobe bricks, and that we want to be at least 95% confident that the error is at most 0.04. How large a sample will we need if

(a) we have no idea what the true proportion might be;
(b) we know that the true proportion does not exceed 0.12?

Solution (a) Using the second of the two formulas for the sample size, we get

$$n = \frac{1}{4}\left[\frac{1.96}{0.04}\right]^2 = 600.25$$

or $n = 601$ rounded up to the nearest integer.

(b) Using the first of the two formulas for the sample size with $p = 0.12$ (the possible value closest to $p = \frac{1}{2}$), we get

$$n = (0.12)(0.88)\left[\frac{1.96}{0.04}\right]^2 = 253.55$$

or $n = 254$ rounded up to the nearest integer. This serves to illustrate how some collateral information about the possible size of p can substantially reduce the size of the required sample. ■

When p is very close to 0, as is the case in problems of high reliability and p is the probability of failure, none of the confidence intervals we have discussed provides a satisfactory solution. What we really need here are **one-sided confidence intervals** of the form $p < C$, where C is a constant depending on the degree of confidence and the size of the sample. As we already pointed out on page 117, the binomial distribution is best approximated with a Poisson distribution with $\lambda = np$ when p is small and n is large. Based on this approximation, it can be shown that

One-sided confidence interval for p

$$p < \frac{1}{2n} \cdot \chi^2_\alpha$$

is a one-sided confidence interval for p, where χ^2_α is as defined on page 211 and the number of degrees of freedom equals $2(x + 1)$. A discussion of this result may be found in the book by A. Hald mentioned in the bibliography.

EXAMPLE If there are $x = 4$ failures among $n = 2{,}000$ parts used continuously for a month, construct a one-sided 99% confidence interval for the probability that one such part will fail under the stated conditions.

Solution Since $\chi^2_{0.01} = 23.209$ for $2(4+1) = 10$ degrees of freedom, substitution into the formula yields

$$p < \frac{1}{2(2{,}000)} \cdot 23.209$$

and, hence,

$$p < 0.0058$$

That is, 0.0058 is an approximate 95% upper confidence bound for p.

9.2 BAYESIAN ESTIMATION†

In the preceding section we looked upon the true proportions we tried to estimate as unknown constants; in Bayesian estimation these parameters are looked upon as random variables having prior distributions which reflect the strength of one's belief about the possible values they can take on, or other indirect information. As in Section 7.3, we are thus faced with the problem of combining prior information with direct sample evidence.

To illustrate how this might be done, suppose that a manufacturer, who regularly receives large shipments of electronic components from a vendor, knows that about 25% of the time 0.005 (half of 1%) of the components are defective, about 25% of the time 0.01 of the components are defective, and about 50% of the time 0.02 of the components are defective. Thus, before a shipment from this vendor is inspected, we have the following prior distribution for the proportion of defectives:

Value of p	*Prior probability*
0.005	0.25
0.01	0.25
0.02	0.50

Now suppose that 200 of these components, randomly selected from the shipment, are inspected, and only one of them is found to be defective. The probability of this happening when $p = 0.005$, $p = 0.01$, or $p = 0.02$ are, respectively,

$$\binom{200}{1}(0.005)^1(0.995)^{199} = 0.37$$

$$\binom{200}{1}(0.01)^1(0.99)^{199} = 0.27, \quad \text{and} \quad \binom{200}{1}(0.02)^1(0.98)^{199} = 0.07$$

where we used the formula for the binomial distribution on page 96 and logarithms to simplify the calculations. Combining these probabilities by means of the formula for Bayes' theorem. (Theorem 3.11 on page 79), we find that the posterior probability for $p = 0.005$ is

$$\frac{(0.25)(0.37)}{(0.25)(0.37) + (0.25)(0.27) + (0.50)(0.07)} = 0.47$$

† This section may be omitted without loss of continuity.

and that the corresponding posterior probabilities for $p = 0.01$ and $p = 0.02$ are 0.35 and 0.18. We have thus arrived at the following posterior distribution for the proportion of defective components:

Value of p	*Posterior probability*
0.005	0.47
0.01	0.35
0.02	0.18

Note that whereas the odds were originally 3 to 1 against $p = 0.005$, it is now almost an even bet; of course, this shift is accounted for by the fact that in the sample only $\frac{1}{200} = 0.005$ of the components inspected were defective.

In the preceding example we assumed that p had to be 0.005, 0.01, or 0.02, and this restriction was imposed mainly to simplify the calculations; the method would have been the same if we had considered 10 different values of p, or even 100. It would be more logical, perhaps, to let p take on any value on the continuous interval from 0 to 1, and in that case it is customary to use as the prior distribution the beta distribution of Section 5.8. The parameters of this distribution are α and β, and its mean and variance can be expressed in terms of α and β in accordance with the formulas on page 162. It can then be shown that the posterior distribution of p, namely, the conditional distribution of p for a given (observed) value of x, is also a beta distribution, and that its parameters are $x + \alpha$ and $n - x + \beta$ instead of α and β.† Thus, the mean and the variance of the posterior distribution may be obtained by substituting $x + \alpha$ for α and $n - x + \beta$ for β in the formulas on page 162.

EXAMPLE A person doing research for a large oil company feels that the proportion of persons requiring oil as well as gasoline at one of the oil company's service stations is a random variable having the beta distribution with $\alpha = 10$ and $\beta = 400$. In a random sample of size $n = 800$, she finds that only $x = 3$ persons required oil as well as gasoline. Find the mean and the variance of

(a) the prior distribution of p;

(b) the posterior distribution of p.

Solution (a) For the prior distribution we get

$$\mu_0 = \frac{10}{10 + 400} = 0.024$$

and

† Proofs of these results may be found in the book by John E. Freund listed in the bibliography.

$$\sigma_0^2 = \frac{10 \cdot 400}{410^2 \cdot 411} = 0.000058$$

(b) For the posterior distribution we get

$$\mu_1 = \frac{3 + 10}{10 + 400 + 800} = 0.011$$

and

$$\sigma_1^2 = \frac{(3 + 10)(800 - 3 + 400)}{(10 + 400 + 800)^2 (10 + 400 + 800 + 1)}$$
$$= 0.0000088$$

If we have mathematical tables giving the values of beta integrals, we can continue with an example like this and calculate prior as well as posterior probabilities associated with various intervals of the values of p.

EXERCISES

9.1 In a random sample of 200 claims filed against an insurance company writing collision insurance on cars, 84 exceeded \$1,200. Construct a 95% confidence interval for the true proportion of claims filed against this insurance company that exceed \$1,200, using
(a) Table 9; $(0.35 < p < 0.49)$
(b) the large sample confidence interval formula. $(0.352 < p < 0.488)$

9.2 With reference to Exercise 9.1, what can we say with 99% confidence about the maximum error, if we use the sample proportion as an estimate of the true proportion of claims filed against this insurance company that exceed \$1,200?

9.3 In a random sample of 400 industrial accidents, it was found that 231 were due at least partially to unsafe working conditions. Construct a 99% confidence interval for the corresponding true proportion using
(a) Table 9; $(0.52 < p < 0.64)$
(b) the large sample confidence interval formula. $(0.514 < p < 0.642)$

9.4 With reference to Exercise 9.3, what can we say with 95% confidence about the maximum error if we use the sample proportion to estimate the corresponding true proportion?

9.5 In a sample survey of the "safety explosives" used in certain mining operations, explosives containing potassium nitrate were found to be used in 95 of 250 cases.
(a) Use Table 9 to construct a 95% confidence interval for the corresponding true proportion. $(0.319 < p < 0.445)$
(b) If $\frac{95}{250} = 0.38$ is used as an estimate of the corresponding true proportion, what can we say with 95% confidence about the maximum error? $E = 0.06$

9.6 In a random sample of 60 sections of pipe in a chemical plant, 8 showed signs of serious corrosion. Construct a 95% confidence interval for the true proportion of pipe sections showing signs of serious corrosion, using
(a) Table 9;
(b) the large sample confidence interval formula.

9.7 In a recent study, 69 of 120 meteorites were observed to enter the earth's atmosphere with a velocity of less than 26 miles per second. If we estimate the corresponding true

proportion as $\frac{69}{120} = 0.575$, what can we say with 95% confidence about the maximum error? (E = 0.0885)

9.8 Among 100 fish caught in a large lake, 18 were inedible due to the pollution of the environment. If we use $\frac{18}{100} = 0.18$ as an estimate of the corresponding true proportion, with what confidence can we assert that the error of this estimate is at most 0.065?

9.9 A random sample of 300 shoppers at a supermarket includes 204 who regularly use cents-off coupons. Construct a 98% confidence interval for the probability that any one shopper at the supermarket, selected at random, will regularly use cents-off coupons.

9.10 What is the size of the smallest sample required to estimate an unknown proportion to within a maximum error of 0.06 with at least 95% confidence?

9.11 With reference to Exercise 9.10, how would the required sample size be affected if it is known that the proportion to be estimated is at least 0.75? (n = 201)

9.12 Suppose that we want to estimate what percentage of all drivers exceed the 55-mile per hour speed limit on a certain stretch of road. How large a sample will we need to be at least 99% confident that the error of our estimate, the sample percentage, is at most 3.5%?

9.13 With reference to Exercise 9.12, how would the required sample size be affected if it is known that the percentage to be estimated is at most 40%? (n = 1,300)

9.14 Show that the inequality on page 282 leads to the following $(1 - \alpha)100\%$ confidence limits:

$$\frac{x + \frac{1}{2}z^2_{\alpha/2} \pm z_{\alpha/2}\sqrt{\frac{x(n-x)}{n} + \frac{1}{4}z^2_{\alpha/2}}}{n + z^2_{\alpha/2}}$$

9.15 Use the formula of Exercise 9.14 to rework Exercise 9.3.

9.16 Use the formula of Exercise 9.14 to rework Exercise 9.6.

9.17 In a random sample of 500 remote controls for home entertainment centers, 7 failed during the 90-day warranty period. Construct an upper 95% confidence limit for the true probability of failure during warranty. ($p < 0.021$)

9.18 Observing the amount of pollutants in the air in a western city on 500 days, it was found that it exceeded 200 micrograms per cubic meter only four times. Construct an upper 99% confidence limit for the probability that the air pollution in this city will exceed 200 micrograms per cubic meter on any one day.

9.19 The head of a highway department feels that four out of five road building jobs stay within cost estimates, while his assistant feels that it should be only three out of five.

(a) If the head of the highway department is regarded to be "three times as good" as his assistant in determining figures like these, what prior probabilities should we assign to their claims? (0.75 and 0.25)

(b) What posterior probabilities should we assign to their claims if it is found that among 12 road building jobs (randomly selected from the department's files) only two stayed within cost estimates? (0.0052 and 0.9948)

9.20 The purchasing agent of a firm feels that the probability is 0.80 that any one of several shipments of steel recently received will meet specifications. The head of the firm's quality control department feels that this probability is 0.90, and the chief engineer feels (somewhat more pessimistically) that it is 0.60.

(a) If the managing director of the firm feels that in this matter the purchasing agent is 10 times as reliable as the chief engineer while the head of the quality

control department is 14 times as reliable as the chief engineer, what pri probabilities would she assign to their claims?

(b) If five of the shipments are inspected and only two meet specifications, wh posterior probabilities should the managing director of the firm assign to th respective claims?

9.21 The output of a certain transistor production line is checked daily by inspecting a sampl of 200 units. Over a long period of time, the process has maintained a yield of 80%, that is a proportion defective of 0.20, and the variation of the proportion defective (from day t day) is measured by a standard deviation of 0.0125. If on a certain day the sample contain 86 defectives, find the mean of the posterior distribution of the proportion defective a an estimate of that day's proportion defective. Assume that the prior distribution of the proportion defective can be approximated closely with a beta distribution.

9.22 Records of the dean of an engineering school (collected over many years) show that on the average 75% of all applicants have an IQ of at least 115. Of course, the percentage varies somewhat from year to year and this variation is measured by a standard deviation of 2.15%.

(a) Verify that if the prior distribution of the proportion of applicants with an IQ of at least 115 can be approximated closely with a beta distribution, we can use the beta distribution with $\alpha = 300$ and $\beta = 100$.

(b) If a sample check of 25 of this year's applicants shows that only 16 of them have an IQ of at least 115, use the results and the assumptions of part (a) to find the mean and the standard deviation of the posterior distribution of the proportion of this year's applicants who have an IQ of at least 115.

9.3 HYPOTHESES CONCERNING ONE PROPORTION

Many of the methods used in sampling inspection, quality control, and reliability verification are based on tests of the null hypothesis that a proportion (percentage, or probability) equals some specified constant. The details of the application of such tests to quality control will be discussed in Chapter 14, where we shall also go into some problems of sampling inspection; applications to reliability and life testing will be taken up in Chapter 15.

Although there are exact tests based on the binomial distribution that can be performed with the use of Table I, we shall consider here only approximate large-sample tests based on the normal approximation to the binomial distribution. In other words, we shall test the null hypothesis $p = p_0$ against one of the alternatives $p < p_0$, $p > p_0$, or $p \neq p_0$ with the use of the statistic

Statistic for large-sample test concerning p

$$Z = \frac{X - np_0}{\sqrt{np_0(1 - p_0)}}$$

which is a random variable having approximately the standard normal distribution.† The critical regions are like those shown in the table on page 240 with p and p_0 substituted for μ and μ_0.

EXAMPLE In a study designed to investigate whether certain detonators used with explosives in coal mining meet the requirement that at least 90% will ignite the explosive when charged, it is found that 174 of 200 detonators function properly. Test the null hypothesis $p = 0.90$ against the alternative hypothesis $p < 0.90$ at the 0.05 level of significance.

Solution

1. *Null hypothesis*: $p = 0.90$
   Alternative hypothesis: $p < 0.90$
2. *Level of significance*: $\alpha = 0.05$
3. *Criterion*: Reject the null hypothesis if $Z < -1.645$, where

$$Z = \frac{X - np_0}{\sqrt{np_0(1 - p_0)}}$$

4. *Calculations*: Substituting $x = 174, n = 200$, and $p_0 = 0.90$ into the formula above, we get

$$z = \frac{174 - 200(0.90)}{\sqrt{200(0.90)(0.10)}} = -1.41$$

5. *Decision*: Since $z = -1.41$ is not less than -1.645, the null hypothesis cannot be rejected; in other words, there is not sufficient evidence to say that the given kind of detonator fails to meet the required standard.

■

9.4 HYPOTHESES CONCERNING SEVERAL PROPORTIONS

When we compare the consumer response (percentage favorable and percentage unfavorable) to two different products, when we decide whether the proportion of defectives of a given process remains constant from day to day, when we judge whether there is a difference in political persuasion among several nationality groups, and in many similar situations, we are interested in testing whether two or more binomial

† Some authors write the numerator of this formula for Z as $X \pm \frac{1}{2} - np_0$, whichever is numerically smaller, but there is generally no need for this continuity correction so long as n is large.

populations have the same parameter p. Referring to these parameters as $p_1, p_2, \ldots,$ and p_k, we are, in fact, interested in testing the null hypothesis

$$p_1 = p_2 = \cdots = p_k = p$$

against the alternative hypothesis that these population proportions are not all equal. To perform a suitable large sample test of this hypothesis, we require independent random samples of size $n_1, n_2, \ldots,$ and n_k from the k populations; then, if the corresponding numbers of "successes" are $X_1, X_2, \ldots,$ and X_k, the test we shall use is based on the fact that (1) for large samples the sampling distribution of

$$Z_i = \frac{X_i - n_i p_i}{\sqrt{n_i p_i (1 - p_i)}}$$

is approximately the standard normal distribution, (2) the square of a random variable having the standard normal distribution is a random variable having the chi-square distribution with 1 degree of freedom, and (3) the sum of k independent random variables having chi-square distributions with 1 degree of freedom is a random variable having the chi-square distribution with k degrees of freedom. (Proofs of these last two results may be found in the book by John E. Freund mentioned in the bibliography.) Thus,

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - n_i p_i)^2}{n_i p_i (1 - p_i)}$$

is a value of a random variable having approximately the chi-square distribution with k degrees of freedom, and in practice we substitute for the p_i, which under the null hypothesis are all equal, the pooled estimate

$$\hat{p} = \frac{x_1 + x_2 + \cdots + x_k}{n_1 + n_2 + \cdots + n_k}$$

Since the null hypothesis should be rejected if the differences between the x_i and the $n_i\hat{p}$ are large, the critical region is $\chi^2 > \chi^2_\alpha$, where χ^2_α is as defined on page 211 and the number of degrees of freedom is $k - 1$. The loss of one degree of freedom results from substituting for p the estimate $\hat{p}$.

In actual practice, when we compare two or more sample proportions it is convenient to determine the value of the χ^2 statistic by looking at the data as arranged in the following way:

	Sample 1	*Sample 2*	⋯	*Sample k*	*Total*
Successes	x_1	x_2	⋯	x_k	x
Failures	$n_1 - x_1$	$n_2 - x_2$	⋯	$n_k - x_k$	$n - x$
Total	n_1	n_2	⋯	n_k	n

The notation is the same as before, except for x and n, which represent, respectively, the total number of successes and the total number of trials for all samples combined. With reference to this table, the entry in the cell belonging to the ith row and jth column is called the **observed cell frequency** o_{ij} with $i = 1, 2$ and $j = 1, 2, \ldots, k$.

Under the null hypothesis $p_1 = p_2 = \cdots = p_k = p$, we estimate p, as before, as the total number of successes divided by the total number of trials, which we now write as $\hat{p} = \frac{x}{n}$. Hence, the expected number of successes and failures for the jth sample are estimated by

$$e_{1j} = n_j \cdot \hat{p} = \frac{n_j \cdot x}{n}$$

and

$$e_{2j} = n_j(1 - \hat{p}) = \frac{n_j \cdot (n - x)}{n}$$

The quantities e_{1j} and e_{2j} are called the **expected cell frequencies** for $j = 1, 2, \ldots, k$. Note that **the expected frequency for any given cell may be obtained by multiplying the totals of the column and the row to which it belongs and then dividing by the grand total n.**

In this notation, the χ^2 statistic on page 292, with $\hat{p}$ substituted for the p_i, can be written in the form

Statistic for test concerning difference among proportions

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

as the reader will be asked to verify in Exercise 9.40 on page 300. This formula has the advantage that it can easily be extended to the more general case, to be treated in Section 9.5, where each trial permits more than two possible outcomes, and there are, thus, more than two rows in the tabular presentation of the various frequencies.

EXAMPLE Samples of three kinds of materials, subjected to extreme temperature changes, produced the results shown in the following table:

	Material A	*Material B*	*Material C*	*Total*
Crumbled	41	27	22	90
Remained intact	79	53	78	210
Total	120	80	100	300

Use the 0.05 level of significance to test whether, under the stated conditions, the probability of crumbling is the same for the three kinds of materials.

Solution

1. *Null hypothesis*: $p_1 = p_2 = p_3$
   Alternative hypothesis: p_1, p_2, and p_3 are not all equal.
2. *Level of significance*: $\alpha = 0.05$
3. *Criterion*: Reject the null hypothesis if $\chi^2 > 5.991$, the value of $\chi^2_{0.05}$ for $3 - 1 = 2$ degrees of freedom, where χ^2 is given by the formula above.
4. *Calculations*: The expected frequencies for the first two cells of the first row are

$$e_{11} = \frac{90 \cdot 120}{300} = 36 \quad \text{and} \quad e_{12} = \frac{90 \cdot 80}{300} = 24$$

and, as it can be shown that **the sum of the expected frequencies for any row or column equals that of the corresponding observed frequencies** (see Exercise 9.41 on page 300), we find by subtraction that $e_{13} = 90 - (36+24) = 30$, and that the expected frequencies for the second row are $e_{21} = 120 - 36 = 84$, $e_{22} = 80 - 24 = 56$, and $e_{23} = 100 - 30 = 70$. Then, substituting these values together with the observed frequencies into the formula for χ^2, we get

$$\chi^2 = \frac{(41-36)^2}{36} + \frac{(27-24)^2}{24} + \frac{(22-30)^2}{30} + \frac{(79-84)^2}{84} + \frac{(53-56)^2}{56} + \frac{(78-70)^2}{70}$$

$$= 4.575$$

5. *Decision*: Since $\chi^2 = 4.575$ does not exceed 5.991, the null hypothesis cannot be rejected; in other words, the data do not refute the hypothesis that, under the stated conditions, the probability of crumbling is the same for the three kinds of material.

■

Most of the entries of Table 5 are given to three decimal places, but, since rounding errors tend to average out, there is seldom any need to give more than two decimal places in the final value of the χ^2 statistic. The test we have been discussing here is only an approximate test since the sampling distribution of the χ^2 statistic is only approximately the chi-square distribution, and it should not be used when one or more of the expected frequencies is less than 5. If this is the case, we can sometimes combine two or more of the samples in such a way that none of the e's is less than 5.

If the null hypothesis of equal proportions is rejected, it is a good practice to graph the confidence intervals (see page 282) for the individual proportions p_i. The graph helps illuminate differences between the proportions.

EXAMPLE Four methods are under development for making discs of a super conducting material. Fifty discs are made by each method and they are checked for superconductivity when cooled with liquid nitrogen.

	Method 1	*Method 2*	*Method 3*	*Method 4*	*Total*
Super conductors	31	42	22	25	120
Failures	19	8	28	25	80
Total	50	50	50	50	200

Perform a chi-square test with $\alpha = 0.05$. If there is a significant difference between the proportions of super conductors produced, plot the individual confidence intervals.

Solution

1. *Null hypothesis*: $p_1 = p_2 = p_3 = p_4$
   Alternative hypothesis: p_1, p_2, p_3, and p_4 are not all equal.
2. *Level of significance*: $\alpha = 0.05$
3. *Criterion*: Reject the null hypothesis if $\chi^2 > 7.815$, the value of $\chi^2_{0.05}$ for $4 - 1 = 3$ degrees of freedom.
4. *Calculations*: Each cell in the first row has expected frequency $120 \cdot \frac{50}{200} = 30$, and each cell in the second row has expected frequency $80 \cdot \frac{50}{200} = 20$.
   The chi-square statistic is

$$\chi^2 = \frac{1}{30} + \frac{144}{30} + \frac{64}{30} + \frac{25}{30} + \frac{1}{20} + \frac{144}{20} + \frac{64}{20} + \frac{25}{20} = 19.50$$

5. *Decision*: Since 19.50 greatly exceeds 7.815, we reject the null hypothesis of equal proportions at the 5% level of significance.

The confidence intervals, obtained from the large-sample formula on page 282 are

$$0.62 \pm 0.13,\ 0.84 \pm 0.14,\ 0.44 \pm 0.14,\ 0.50 \pm 0.14$$

These are plotted in Figure 9.2. Note how Method 2 stands out as being better. ■

Although there has been no mention of randomization in the development of the χ^2 statistic, wherever possible the experimental units should be randomly assigned to methods. In the example above, the discs could be numbered from 1 to 200 and random numbers selected from 1 to 200 without replacement. The discs corresponding to the first fifty numbers drawn would be assigned to method 1 and so on. This will prevent uncontrolled sources of variation from systematically influencing the test concerning the four methods.

So far, the alternative hypothesis has been that $p_1, p_2, \ldots$, and p_k are not all equal, and for $k = 2$ this reduces to the alternative hypothesis $p_1 \neq p_2$. In problems where

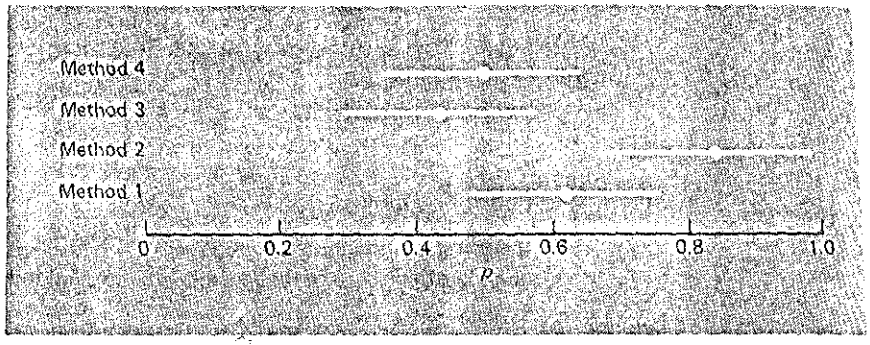


FIGURE 9.2
Confidence intervals for several proportions.

the alternative hypothesis may also be $p_1 < p_2$ or $p_1 > p_2$, we can base the test on the statistic

Statistic for test concerning difference between two proportions

$$Z = \frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{with} \quad \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

which, for large samples, is a random variable having approximately the standard normal distribution. The test based on this statistic is equivalent to the one based on the χ^2 statistic on page 293 with $k = 2$, in the sense that the square of this Z statistic actually equals the χ^2 statistic (see Exercise 9.42 on page 300). The critical regions for this alternative test of the null hypothesis $p_1 = p_2$ are like those shown in the table on page 240 with p_1 and p_2 substituted for μ and μ_0.

EXAMPLE A study shows that 16 of 200 tractors produced on one assembly line required extensive adjustments before they could be shipped, while the same was true for 14 of 400 tractors produced on another assembly line. At the 0.01 level of significance, does this support the claim that the second production line does superior work?

Solution

1. *Null hypothesis*: $p_1 = p_2$
   Alternative hypothesis: $p_1 > p_2$
2. *Level of significance*: $\alpha = 0.01$
3. *Criterion*: Reject the null hypothesis if $Z > 2.33$, where Z is given by the above formula.
4. *Calculations*: Substituting $x_1 = 16, n_1 = 200, x_2 = 14, n_2 = 400$, and

$$\hat{p} = \frac{16 + 14}{200 + 400} = 0.05$$

into the formula for Z, we get

$$Z = \frac{\frac{16}{200} - \frac{14}{400}}{\sqrt{(0.05)(0.95)\left(\frac{1}{200} + \frac{1}{400}\right)}} = 2.38$$

5. *Decision*: Since $Z = 2.38$ exceeds 2.33, the null hypothesis must be rejected; we conclude that the true proportion of tractors requiring extensive adjustments is greater for the first assembly line than for the second.

■

The test we have described here applies to the null hypothesis $p_1 = p_2$, but it can easily be modified (see Exercise 9.38 on page 299) so that it applies also to the null hypothesis $p_1 - p_2 = \delta$.

The statistic for testing $p_1 = p_2$ leads to a confidence interval which provides the set of plausible values for $p_1 - p_2$.

Large-sample confidence interval for the difference of two proportions

$$\frac{x_1}{n_1} - \frac{x_2}{n_2} \pm z_{\alpha/2}\sqrt{\frac{\frac{x_1}{n_1}\left(1-\frac{x_1}{n_1}\right)}{n_1} + \frac{\frac{x_2}{n_2}\left(1-\frac{x_2}{n_2}\right)}{n_2}}$$

EXAMPLE With reference to the preceding example, find the large-sample 95% confidence interval for $p_1 - p_2$.

Solution Since $x_1/n_1 = \hat{p}_1 = \frac{16}{200} = 0.08$ and $x_2/n_2 = \hat{p}_2 = \frac{14}{400} = 0.035$

$$\frac{x_1}{n_1} - \frac{x_2}{n_2} \pm z_{\alpha/2}\sqrt{\frac{\frac{x_1}{n_1}\left(1-\frac{x_1}{n_1}\right)}{n_1} + \frac{\frac{x_2}{n_2}\left(1-\frac{x_2}{n_2}\right)}{n_2}}$$

$$= 0.08 - 0.035 \pm 1.96\sqrt{\frac{(0.08)(0.92)}{200} + \frac{(0.035)(0.965)}{400}},$$

or $0.003 < p_1 - p_2 < 0.087$

The first shift has a rate of extensive adjustment between 3 out of 1,000 and 87 out of 1,000, higher than the rate for the second shift.

■

EXERCISES

9.23 A manufacturer of submersible pumps claims that at most 30% of the pumps require repairs within the first 5 years of operation. If a random sample of 120 of these pumps includes 47 which required repairs within the first 5 years, test the null hypothesis $p = 0.30$ against the alternative hypothesis $p > 0.30$ at the 0.05 level of significance.

(z = 2.19 ; reject H_0)

9.24 The performance of a computer is observed over a period of 2 years to check the claim that the probability is 0.20 that its down time will exceed 5 hours in any given week. Testing the null hypothesis $p = 0.20$ against the alternative hypothesis $p \neq 0.20$, what can we conclude at the level of significance $\alpha = 0.05$, if there were only 11 weeks in which the downtime of the computer exceeded 5 hours?

9.25 To check on an ambulance service's claim that at least 40% of its calls are life-threatening emergencies, a random sample was taken from its files, and it was found that only 49 of 150 calls were life-threatening emergencies. Can the null hypothesis $p \geq 0.40$ be rejected against the alternative hypothesis $p < 0.40$ if the probability of a Type I error is to be at most 0.01? ($z = -1.83$; cannot reject H_0)

9.26 In a random sample of 600 cars making a right turn at a certain intersection, 157 pulled into the wrong lane. Test the null hypothesis that actually 30% of all drivers make this mistake at the given intersection, using the alternative hypothesis $p \neq 0.30$ and the level of significance

(a) $\alpha = 0.05$; (b) $\alpha = 0.01$.

9.27 An airline claims that only 6% of all lost luggage is never found. If, in a random sample, 17 of 200 pieces of lost luggage are not found, test the null hypothesis $p = 0.06$ against the alternative hypothesis $p > 0.06$ at the 0.05 level of significance. ($z = 1.484$; not sig.)

9.28 Suppose that 4 of 13 undergraduate engineering students state that they will go on to graduate school. Test the dean's claim that 60% of the undergraduate students will go on to graduate school, using the alternative hypothesis $p < 0.60$ and the level of significance $\alpha = 0.05$. [*Hint*: Use Table 1 to determine the probability of getting "at most 4 successes in 13 trials" when $p = 0.60$.]

At most 3 or at least 12 Heads; 0.0352

9.29 Suppose that we want to test the "honesty" of a coin on the basis of the number of heads we will get in 15 flips. Using Table 1, determine how few or how many heads we would have to get so that we could reject the null hypothesis $p = 0.50$ against the alternative hypothesis $p \neq 0.50$ at the level of significance no larger than 0.05. what is the *actual* level of significance we would be using with this criterion?

9.30 It costs more to test a certain type of ammunition than to manufacture it, and, hence, only three rounds are tested from each large lot. If the lot is rejected unless all three rounds function according to specifications,

(a) sketch the *OC* curve for this test;

(b) find the actual proportion of defectives for which the test procedure will cause a lot to be rejected with a probability of 0.10.

9.31 Tests are made on the proportion of defective castings produced by five different molds. If there were 14 defectives among 100 castings made with Mold I, 33 defectives among 200 castings made with Mold II, 21 defectives among 180 castings made with Mold III, 17 defectives among 120 castings made with Mold IV, and 25 defectives among 150 castings made with Mold V, use the 0.01 level of significance to test whether the true proportion of defectives is the same for each mold. ($\chi^2 = 2.37$; cannot reject H_0)

9.32 A study showed that 64 of 180 persons who saw a photocopying machine advertised during the telecast of a baseball game and 75 of 180 other persons who saw it advertised on a variety show remembered the brand name 2 hours later. Use the χ^2 statistic to test at the 0.05 level of significance whether the difference between the corresponding sample proportions is significant.

9.33 The following data come from a study in which random samples of the employees of three government agencies were asked questions about their pension plan:

($\chi^2 = 9.39$; reject H_0)

	Agency 1	Agency 2	Agency 3	
For the pension plan	67	84	109	260
Against the pension plan	33	66	41	140
	100	150	150	400

Use the 0.01 level of significance to test the null hypothesis that the actual proportions of employees favoring the pension plan are the same.

9.34 The owner of a machine shop must decide which of two snack-vending machines to install in his shop. If each machine is tested 250 times, the first machine fails to work (neither delivers the snack nor returns the money) 13 times, and the second machine fails to work 7 times, test at the 0.05 level of significance whether the difference between the corresponding sample proportions is significant, using

(a) the χ^2 statistic on page 293;

(b) the Z statistic on page 296.

9.35 With reference to the preceding exercise, verify that the square of the value obtained for Z in part (b) equals the value obtained for χ^2 in part (a).

9.36 Photolithography plays a central role in manufacturing integrated circuits made on thin disks of silicon. Prior to a quality-improvement program, too many rework operations were required. In a sample of 200 units, 20 required reworking of the photolithographic step. Following training in the use of Pareto charts and other approaches to identify significant problems, improvements were made. A new sample of size 200 had only 12 that needed rework.

Is this sufficient evidence to conclude at the 0.01 level of significance that the improvements have been effective in reducing the rework?

9.37 With reference to Exercise 9.36, find a large sample 99% confidence interval for the true difference of the proportions.

9.38 To test the null hypothesis that the difference between two population proportions equals some constant δ, not necessarily 0, we can use the statistic

$$Z = \frac{\frac{X_1}{n_1} - \frac{X_2}{n_2} - \delta}{\sqrt{\frac{\frac{X_1}{n_1}\left(1 - \frac{X_1}{n_1}\right)}{n_1} + \frac{\frac{X_2}{n_2}\left(1 - \frac{X_2}{n_2}\right)}{n_2}}}$$

which, for large samples, is a random variable having the standard normal distribution.

(a) With reference to Exercise 9.36, use this statistic to test at the 0.05 level of significance whether the true proportion of units requiring rework is now at least 4% less than before the improvements were made.

(b) In a true-false test, a test item is considered to be good if it discriminates between well-prepared students and poorly prepared students. If 205 of 250 well-prepared students and 137 of 250 poorly prepared students answer a certain item correctly, test at the 0.01 level of significance whether for the given item the proportion of correct answers can be expected to be at least 15% higher among well-prepared students than among poorly prepared students.

9.39 With reference to part (b) of Exercise 9.38, find a large-sample 99% confidence interval for the true difference of the proportions.

9.40 Verify that the formulas for the χ^2 statistic on page 292 (with $\hat{p}$ substituted for the p_i) and on page 293 are equivalent.

9.41 Verify that if the expected frequencies are determined in accordance with the rule on page 293, the sum of the expected frequencies for each row and column equals the sum of the corresponding observed frequencies.

9.42 Verify that the square of the Z statistic on page 296 equals the χ^2 statistic on page 293 for $k = 2$.

9.5 THE ANALYSIS OF $r \times c$ TABLES

As we suggested earlier, the method by which we analyzed the example on page 293 lends itself also to the analysis of **$r \times c$ tables**, or r-by-c tables, that is, tables in which data are tallied into a two-way classification having r rows and c columns. Such tables arise in essentially two kinds of problems. First, we might again have samples from several populations, with the distinction that now each trial permits more than two possible outcomes. This might happen, for example, if persons belonging to different income groups are asked whether they favor a certain political candidate, whether they are against him, or whether they are indifferent or undecided. The other situation giving rise to an $r \times c$ table is one in which we sample from one population but classify each item with respect to two (usually qualitative) categories. This might happen, for example, if a consumer testing service rates cars as excellent, superior, average, or poor with regard to performance and also with regard to appearance. Each car tested would then fall into one of the 16 cells of a 4×4 table, and it is mainly in connection with problems of this kind that $r \times c$ tables are referred to as **contingency tables**.

The essential difference between the two kinds of situations giving rise to $r \times c$ tables is that in the first case the column totals (the sample sizes) are fixed, while in the second case only the **grand total** (the total for the entire table) is fixed. As a result, there are also differences in the null hypotheses we shall want to test. In the first case we want to test whether the probability of obtaining an observation in the ith row is the same for each column; symbolically, we shall want to test the null hypothesis

$$p_{i1} = p_{i2} = \cdots = p_{ic} \quad \text{for } i = 1, 2, \ldots, r$$

where p_{ij} is the probability of obtaining an observation belonging to the ith row and the jth column, and $\sum_{i=1}^{r} p_{ij} = 1$ for each column. The alternative hypothesis is that the p's are not all equal for at least one row. In the second case we shall want to test the null hypothesis that the random variables represented by the two classifications are

independent, so that p_{ij} is the product of the probability of getting a value belonging to the ith row and the probability of getting a value belonging to the jth column. The alternative hypothesis is that the two random variables are not independent.

In spite of the differences we have described, the analysis of an $r \times c$ table is the same for both cases. First we calculate the expected cell frequencies e_{ij} as on page 293, namely, by multiplying the totals of the respective rows and columns and then dividing by the grand total. In practice, we make use of the fact that the observed frequencies and the expected frequencies total the same for each row and column, so that only $(r-1)(c-1)$ of the e_{ij} have to be calculated directly, while the others can be obtained by subtraction from appropriate row or column totals. We then substitute into the formula

Statistic for analysis of $r \times c$ table

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$

and we reject the null hypothesis if the value of this statistic exceeds χ^2_α for $(r-1)(c-1)$ degrees of freedom. This expression for the number of degrees of freedom is justified by the above observation that after we determine $(r-1)(c-1)$ of the expected cell frequencies, the others are automatically determined, that is, they may be obtained by subtraction from appropriate row or column totals.

EXAMPLE To determine whether there really is a relationship between an employee's performance in the company's training program and his or her ultimate success in the job, it takes a sample of 400 cases from its very extensive files and obtains the results shown in the following table:

		Performance in training program			
		Below average	*Average*	*Above average*	*Total*
Success in job (employer's rating)	*Poor*	23	60	29	112
	Average	28	79	60	167
	Very good	9	49	63	121
	Total	60	188	152	400

Use the 0.01 level of significance to test the null hypothesis that performance in the training program and success in the job are independent.

Solution

1. *Null hypothesis*: Performance in training program and success in job are independent.
   Alternative hypothesis: Performance in training program and success in job are not independent.

2. *Level of significance*: $\alpha = 0.01$
3. *Criterion*: Reject the null hypothesis if $\chi^2 > 13.277$, the value of $\chi^2_{0.01}$ for $(3-1)(3-1) = 4$ degrees of freedom, where χ^2 is given by the formula on page 301.
4. *Calculations*: Calculating first the expected cell frequencies for the first two cells of the first two rows, we get

$$e_{11} = \frac{112 \cdot 60}{400} = 16.80, \quad e_{12} = \frac{112 \cdot 188}{400} = 52.64,$$

$$e_{21} = \frac{167 \cdot 60}{400} = 25.05, \quad e_{22} = \frac{167 \cdot 188}{400} = 78.49$$

Then, by subtraction, we find that the expected frequencies for the third cell of the first two rows are 42.56 and 63.46, and those for the third row are 18.15, 56.87, and 45.98. Thus,

$$\begin{aligned}\chi^2 = {} & \frac{(23-16.80)^2}{16.80} + \frac{(60-52.64)^2}{52.64} + \frac{(29-42.56)^2}{42.56} \\ & + \frac{(28-25.05)^2}{25.05} + \frac{(79-78.49)^2}{78.49} + \frac{(60-63.46)^2}{63.46} \\ & + \frac{(9-18.15)^2}{18.15} + \frac{(49-56.87)^2}{56.87} + \frac{(63-45.98)^2}{45.98} \\ = {} & 20.179\end{aligned}$$

5. *Decision*: Since $\chi^2 = 20.179$ exceeds 13.277, the null hypothesis must be rejected; we conclude that there is a dependence between an employee's performance in the training program and his or her success in the job.

■

We pursue this example further in order to determine the form of the dependence.

EXAMPLE With reference to the preceding example, find the individual contributions to the chi-square.

Solution We display the contingency table, but this time we include the expected frequencies just below the observed frequencies.

		Performance in training program			
		Below average	*Average*	*Above average*	*Total*
Success in job (employer's rating)	*Poor*	23 **16.80**	60 **52.64**	29 **42.56**	112
	Average	28 **25.05**	79 **78.49**	60 **63.46**	167
	Very good	9 **18.15**	49 **56.87**	63 **45.98**	121
	Total	60	188	152	400

Also, we write

$$\begin{aligned}\chi^2 &= 2.288 + 1.029 + 4.320 \\ &\quad + 0.347 + 0.003 + 0.189 \\ &\quad + 4.613 + 1.089 + 6.300 \\ &= 20.179\end{aligned}$$

From these two displays, it is clear that there is a positive dependence between performance in training and job success. For the three individual cells with the largest contributions to χ^2, the *above average-very good* cell frequency is high, whereas the *above average-poor* and *below average-very good* cell frequencies are low. ■

9.6 GOODNESS OF FIT

We speak of **goodness of fit** when we try to compare an observed frequency distribution with the corresponding values of an expected, or theoretical, distribution. To illustrate, suppose that during 400 five-minute intervals the air-traffic control of an airport received 0, 1, 2, . . . , or 13 radio messages with respective frequencies of 3, 15, 47, 76, 68, 74, 46, 39, 15, 9, 5, 2, 0, and 1. Suppose, furthermore, that we want to check whether these data substantiate the claim that the number of radio messages which they receive during a 5-minute interval may be looked upon as a random variable having the Poisson distribution with $\lambda = 4.6$. Looking up the corresponding Poisson probabilities in Table 2 and multiplying them by 400 to get the expected frequencies, we arrive at the result shown in the following table together with the original data:

Number of radio messages	*Observed frequencies*	*Poisson probabilities*	*Expected frequencies*
0	3 } 18	0.010	4.0 } 22.4
1	15 }	0.046	18.4 }
2	47	0.107	42.8
3	76	0.163	65.2
4	68	0.187	74.8
5	74	0.173	69.2
6	46	0.132	52.8
7	39	0.087	34.8
8	15	0.050	20.0
9	9	0.025	10.0
10	5 } 8	0.012	4.8 } 8.0
11	2 }	0.005	2.0 }
12	0 }	0.002	0.8 }
13	1 }	0.001	0.4 }
	400		400.0

Note that we combined some of the data so that none of the expected frequencies is less than 5.

To test whether the discrepancies between the observed and expected frequencies can be attributed to chance, we use the statistic

Statistic for test of goodness of fit

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

where the o_i and e_i are the observed and expected frequencies. The sampling distribution of this statistic is approximately the chi-square distribution with $k - m$ degrees of freedom, where k is the number of terms in the formula for χ^2 and m is the number of quantities, obtained from the observed data, that are needed to calculate the expected frequencies.

EXAMPLE With reference to the radio message data on page 303, test at the 0.01 level of significance whether the data can be looked upon as values of a random variable having the Poisson distribution with $\lambda = 4.6$.

Solution

1. *Null hypothesis*: Random variable has Poisson distribution with $\lambda = 4.6$.
   Alternative hypothesis: Random variable does not have Poisson distribution with $\lambda = 4.6$.
2. *Level of significance*: $\alpha = 0.01$

3. *Criterion*: Reject the null hypothesis if $\chi^2 > 16.919$, the value of $\chi^2_{0.05}$ for $k-m = 10-1 = 9$ degrees of freedom, where χ^2 is given by the formula above. (The number of degrees of freedom is $10 - 1 = 9$, since only one quantity, the total frequency of 400, is needed from the observed data to calculate the expected frequencies.)
4. *Calculations*: Substitution into the formula for χ^2 yields

$$\chi^2 = \frac{(18-22.4)^2}{22.4} + \frac{(47-42.8)^2}{42.8} + \cdots + \frac{(9-10.0)^2}{10.0} + \frac{(8-8.0)^2}{8.0}$$
$$= 6.749$$

5. *Decision*: Since $\chi^2 = 6.749$ does not exceed 16.919, the null hypothesis cannot be rejected; we conclude that the Poisson distribution with $\lambda = 4.6$ provides a good fit.

■

EXERCISES

9.43 The results of polls conducted two weeks and four weeks before a gubernatorial election are shown in the following table:

	Two weeks before election	*Four weeks before election*
For Republican candidate	79	91
For Democratic candidate	84	66
Undecided	37	43

Use the 0.05 level of significance to test whether there has been a change in opinion during the 2 weeks between the two polls. ($\chi^2 = 3.457$; cannot reject H_0)

9.44 A large electronics firm that hires many handicapped workers wants to determine whether their handicaps affect such workers' performance. Use the level of significance $\alpha = 0.05$ to decide on the basis of the sample data shown in the following table whether it is reasonable to maintain that the handicaps have no effect on the workers' performance:

	Performance		
	Above average	*Average*	*Below average*
Blind	21	64	17
Deaf	16	49	14
No handicap	29	93	28

9.45 Tests of the fidelity and the selectivity of 190 radio receivers produced the results shown in the following table:

		Fidelity		
		Low	*Average*	*High*
Selectivity	*Low*	6	12	32
	Average	33	61	18
	High	13	15	0

Use the 0.01 level of significance to test whether there is a relationship (dependence) between fidelity and selectivity. ($\chi^2 = 54.328$; reject H_0)

9.46 A quality-control engineer takes daily samples of $n = 4$ tractors coming off an assembly line and on 200 consecutive working days the data summarized in the following table are obtained:

Number requiring adjustments	*Number of days*
0	101
1	79
2	19
3	1

To test the claim that 10% of all the tractors coming off this assembly line require adjustments, look up the corresponding probabilities in Table I, calculate the expected frequencies, and perform the chi-square test at the 0.01 level of significance.

9.47 With reference to Exercise 9.46, verify that the mean of the observed distribution is 0.60, corresponding to 15% of the tractors requiring adjustments. Then look up the probabilities for $n = 4$ and $p = 0.15$ in Table I, calculate the expected frequencies, and test at the 0.01 level of significance whether the binomial distribution with $n = 4$ and $p = 0.15$ provides a suitable model for this situation. ($\chi^2 = 0.657$; cannot reject H_0)

9.48 Suppose that in the example on page 303 we had shown first that the mean of the distribution, rounded to one decimal, is 4.5, and then tested whether the Poisson distribution with $\lambda = 4.5$ provides a good fit. What would have been the number of degrees of freedom for the appropriate chi-square criterion?

9.49 The following is the distribution of the hourly number of trucks arriving at a company's warehouse:

Trucks arriving per hour	*Frequency*
0	52
1	151
2	130
3	102
4	45
5	12
6	5
7	1
8	2

Find the mean of this distribution, and using it (rounded to one decimal place) as the parameter λ, fit a Poisson distribution. Test for goodness of fit at the 0.05 level of significance. ($\chi^2 = 9.185$; cannot reject H_0)

9.50 Using any four columns of Table 7 (that is, a total of 200 random digits), construct a table showing how many times each of the digits 0, 1, . . . , and 9 occurred. Comparing the observed frequencies with the corresponding expected frequencies (based on the assumption that the digits are randomly generated), test at the 0.05 level of significance whether the assumption of randomness is tenable.

9.51 The following is the distribution of the sulfur oxides emission data on page 8, for which we showed that $\bar{x} = 18.85$ and $s = \sqrt{30.77} = 5.55$:

Class limits (tons)	*Frequency*
5.0– 8.9	3
9.0–12.9	10
13.0–16.9	14
17.0–20.9	25
21.0–24.9	17
25.0–28.9	9
29.0–32.9	2
	80

(a) Find the probabilities that a random variable having a normal distribution with $\mu = 18.85$ and $\sigma = 5.55$ takes on a value less than 8.95, between 8.95 and 12.95, between 12.95 and 16.95, between 16.95 and 20.95, between 20.95 and 24.95, between 24.95 and 28.95, and greater than 28.95.

(b) Multiply the probabilities obtained in part (a) by the total frequency, $n = 80$, thus getting the expected normal curve frequencies corresponding to the seven classes of the given distribution (with the first one changed to "8.9 or less" and the last one changed to "29.0 or more").

Ans (a) 0.0375 , 0.1071 , 0.2223, 0.2911, 0.2163
0.1013 , 0.0344 .

(b) 3 , 8.6 , 17.8 , 22.5 , 17.3 , 8.1 , 2.8 .

(c) Use the 0.05 level of significance to test the null hypothesis that the given data may be looked upon as a random sample from a normal population. Explain why the number of degrees of freedom for this χ^2 test is $k-3$, where k is the number of terms in the χ^2 statistic. ($\chi^2 = 1.264$; good fit.).

9.52 Among 100 purification filters used in an experiment, 46 had a service life of less than 20 hours, 19 had a service life of 20 or more but less than 40 hours, 17 had a service life of 40 or more but less than 60 hours, 12 had a service life of 60 or more but less than 80 hours, and 6 had a service life of 80 hours or more. Using steps similar to those outlined in the preceding exercise, test at the 0.01 level of significance whether the lifetimes may be regarded as a sample from an exponential population with $\mu = 40$ hours.

9.53 A chi-square test is easily implemented on a computer. The *MINITAB* commands

```
READ INTO C1 C2 C3 C4
31 42 22 25
19  8 28 25
```

place the table from the example on page 294 into columns 1–4. Then,

```
CHISQUARE C1 – C4
```

produces the output

```
Expected counts are printed below observed counts
     Method 1   Method 2   Method 3   Method 4   Total
1          31         42         22         25     120
       30.00      30.00      30.00      30.00
2          19          8         28         25      80
       20.00      20.00      20.00      20.00
Total      50         50         50         50     200
ChiSq = 0.033 + 4.800 + 2.133 + 0.833 +
        0.050 + 7.200 + 3.200 + 1.250 = 19.500
        df=3
```

Repeat the analysis using only the data from the first three methods.

9.54 The procedure in Exercise 9.53 also calculates the chi-square test for independence. Do Exercise 9.44 using the computer.

9.7 REVIEW EXERCISES

9.55 In a sample of 100 ceramic pistons made for an experimental diesel engine, 18 were cracked. Construct a 95% confidence interval for the true proportion of cracked pistons, using

(a) Table 9;

(b) the large sample confidence interval formula.

9.56 With reference to Exercise 9.55, test the null hypothesis $p = 0.20$ versus the alternative hypothesis $p < 0.20$ at the 0.05 level.

9.57 In a random sample of 160 workers exposed to a certain amount of radiation, 24 experienced some ill effects. Construct a 99% confidence interval for the corresponding true percentage, using

(a) Table 9;

(b) the large sample confidence interval formula.

9.58 With reference to Exercise 9.57, test the null hypothesis $p = 0.15$ versus the alternative hypothesis $p \neq 0.15$ at the 0.01 level.

9.59 In a random sample of 100 packages shipped by air freight, 13 had some damage. Construct a 95% confidence interval for the true proportion of damaged packages, using

(a) Table 9;

(b) the large sample confidence interval formula.

9.60 With reference to Exercise 9.59, test the hypothesis $p = 0.10$ versus the alternative hypothesis $p > 0.10$ at the 0.01 level.

9.61 In 4,000 firings of a certain kind of rocket there were 10 instances in which a rocket exploded upon ignition. Construct an upper 95% confidence limit for the probability that such a rocket will explode upon ignition.

9.62 If 26 of 200 Brand A tires fail to last 20,000 miles, whereas the corresponding figures for 200 tires each of Brands B, C, and D are 23, 15, and 32, use the 0.05 level of significance to test the null hypothesis that there is no difference in the quality of the four kinds of tires with regard to their durability.

9.63 One method of seeding clouds was successful in 57 of 150 attempts while another method was successful in 33 of 100 attempts. At the 0.05 level of significance, can we conclude that the first method is better than the second?

9.64 With reference to Exercise 9.63, find a large sample 95% confidence interval for the true difference of probabilities.

9.65 Two bonding agents, A and B, are available for making a laminated beam. Of 50 beams made with Agent A, 11 failed a stress test, whereas 19 of the 50 beams made with Agent B failed. At the 0.05 level, can we conclude that Agent A is better than Agent B?

9.66 With reference to Exercise 9.65, find a large sample 95% confidence interval for the true difference of the probabilities of failure.

9.67 Cooling pipes at three nuclear power plants are investigated for deposits that would inhibit the flow of water. From 30 randomly selected spots at each plant, 13 from the first plant, 8 from the second plant, and 19 from the third were clogged.

(a) Use the 0.05 level to test the null hypothesis of equality.

(b) Plot the confidence intervals for the three probabilities of being clogged.

9.68 Suppose that in Exercise 9.62 we had been interested also in how many of the tires lasted more than 30,000 miles and obtained the results shown in the following table:

	Brand A	*Brand B*	*Brand C*	*Brand D*
Failed to last 20,000 miles	26	23	15	32
Lasted from 20,000 to 30,000	118	93	116	121
Lasted more than 30,000 miles	56	84	69	47

(a) Use the 0.01 level of significance to test the null hypothesis that there is no difference in the quality of the four kinds of tires with regard to their durability.
(b) Plot the four individual 99% confidence intervals for proportions.

9.69 The following is the distribution of the daily number of power failures reported in a western city on 300 days:

Number of power failures	*Number of days*
0	9
1	43
2	64
3	62
4	42
5	36
6	22
7	14
8	6
9	2

Test at the 0.05 level of significance whether the daily number of power failures in this city is a random variable having the Poisson distribution with $\lambda = 3.2$.

9.70 With reference to the example on page 301, repeat the analysis after combining the categories *below average* and *average* in the training program and the categories *poor* and *average* in success. Comment on the form of the dependence.

9.71 Mechanical engineers, testing a new arc welding technique, classified welds both with respect to appearance and an X-ray inspection.

		Appearance			
		Bad	*Normal*	*Good*	*Total*
X-ray	*Bad*	20	7	3	30
	Normal	13	51	16	80
	Good	7	12	21	40
	Total	40	70	40	150

Test for independence using $\alpha = 0.05$ and find the individual cell contributions to the χ^2 statistic.

7.97 (a) Randomly select 10 cars to use the modified spark plugs. The other 10 cars use the regular spark plugs; (b) select 7 specimens by random drawing, to try in the old oven.

7.99 $t = -1.99$ we would not reject $\mu_A = \mu_B$ at the level of significance $\alpha = .05$.

CHAPTER 8

8.1 (a) $s = 4.195$; (b) 4.341.

8.3 (a) 1.787; (b) 2.144.

8.5 $.082 < \sigma < .695$.

8.7 $\chi^2 = 5.832$; cannot reject H_0.

8.9 $\chi^2 = 115.886$; reject H_0.

8.11 $\chi^2 = 10.89$; cannot reject H_0.

8.13 $F = 1.496$; we cannot reject H_0.

8.15 $F = 2.42$ we cannot reject H_0.

8.17 (a) No, samples are not normal and variances are unequal; (b) base test on the logarithms of the observations.

8.19 $1.52 < \sigma_1 < 2.20$ and $1.91 < \sigma_2 < 3.32$.

8.21 $\chi^2 = 19.21$; cannot reject H_0.

8.23 $\chi^2 = 72.22$; reject H_0.

8.25 $F = 2.797$ we cannot reject H_0.

CHAPTER 9

9.1 (a) $0.35 < p < 0.49$; (b) $0.352 < p < 0.488$.

9.3 (a) $0.52 < p < 0.64$; (b) $0.514 < p < 0.642$.

9.5 (a) $0.319 < p < 0.445$; (b) $E = 0.06$.

9.7 $E = 0.0885$.

9.9 $0.617 < p < 0.743$.

9.11 $n = 201$.

9.13 $n = 1{,}300$.

9.15 $0.513 < p < 0.639$.

9.17 $p < 0.026$.

9.19 (a) 0.75 and 0.25; (b) 0.0052 and 0.9948.

9.21 0.238.

9.23 $z = 2.19$; reject H_0.

9.25 $z = -1.83$; cannot reject H_0.

9.27 $z = 1.489$; cannot reject H_0.

9.29 At most three or at least twelve heads; 0.0352.

9.31 $\chi^2 = 2.37$; cannot reject H_0.

9.33 $\chi^2 = 9.39$; reject H_0.

9.37 $-0.005 < p_1 - p_2 < 0.145$.

9.39 $0.170 < p_1 - p_2 < 0.374$.

9.43 $\chi^2 = 3.457$; cannot reject H_0.

9.45 $\chi^2 = 54.328$; reject H_0.

9.47 $\chi^2 = 0.657$; cannot reject H_0.

9.49 $\chi^2 = 9.185$; cannot reject H_0.

9.51 (a) 0.0375, 0.1071, 0.2223, 0.2811, 0.2163, 0.1013 and 0.0344; (b) 3, 8.6, 17.8, 22.5, 17.3, 8.1 and 2.8; (c) $\chi^2 = 1.264$; good fit.

9.55 (a) $0.115 < p < 0.275$; (b) $0.105 < p < 0.255$.

9.57 (a) $0.084 < p < 0.243$; (b) $0.077 < p < 0.223$.

9.59 (a) $0.075 < p < 0.217$; (b) $0.064 < p < 0.196$.

9.61 $p < 0.00424$.

9.63 $z = 0.807$; cannot reject H_0.

9.65 $z = -1.746$; reject H_0.

9.67 (a) $\chi^2 = 8.190$; reject H_0; (b) $0.256 < p_1 < 0.611$; $0.108 < p_2 < 0.425$; $0.461 < p_3 < 0.806$.

9.69 $\chi^2 = 10.481$; cannot reject H_0.

9.71 $\chi^2 = 47.862$; reject H_0.

CHAPTER 10

10.1 $P(10 \textit{ or more}) = 0.2272$; cannot reject H_0.

10.3 $P(6 \textit{ or fewer}) = 0.3036$; cannot reject H_0.

10.5 $z = -3.91$; reject H_0.

10.7 $P(3 \textit{ or fewer}) = 0.1719$; cannot reject H_0.

10.9 $z = -0.10$; cannot reject H_0.

10.11 $z = 2.92$; difference is significant.

10.13 $z = 1.62$; cannot reject H_0.

10.15 $H = 26.0$; the populations are not identical.

10.17 $z = -0.244$; cannot reject H_0.

10.21 $z = -4.00$; reject H_0.

10.23 Maximum difference is about 0.22; cannot reject H_0.

10.25 $P(6 \textit{ or more}) = 0.0625$; reject H_0.

10.27 $z = -1.814$; reject H_0.

10.29 $H = 0.904$; cannot reject H_0.

10.31 $z = -1.797$; reject H_0.

10.33 $W_1 = 25$ so $U_1 = 19$; reject H_0.

10.35 $z = -2.248$; reject H_0.

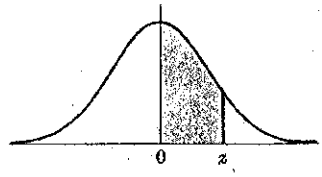
CHAPTER 11

11.1 (b) $\hat{y} = 39.05 + 0.764x$; $\hat{y} = 65.8$.

11.3 (b) $\hat{y} = 1.13 + 14.49x$; $\hat{y} = 51.8$.

Appendix II

AREAS under the STANDARD NORMAL CURVE from 0 to z



z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000

Note: 1) ถ้า $X \sim N(\mu, \sigma^2)$ แล้ว $Z = (X-\mu)/\sigma \sim N(0,1)$

2) ถ้า $Z_n = (\bar{X}-\mu)/\sigma/\sqrt{n}$ แล้ว $Z_n \longrightarrow Z$ เมื่อ $n \to \infty$ (CLT)

Table 4 Values of t_α*

ν	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$	ν
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10
11	1.363	1.796	2.201	2.718	3.106	11
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	1.341	1.753	2.131	2.602	2.947	15
16	1.337	1.746	2.120	2.583	2.921	16
17	1.333	1.740	2.110	2.567	2.898	17
18	1.330	1.734	2.101	2.552	2.878	18
19	1.328	1.729	2.093	2.539	2.861	19
20	1.325	1.725	2.086	2.528	2.845	20
21	1.323	1.721	2.080	2.518	2.831	21
22	1.321	1.717	2.074	2.508	2.819	22
23	1.319	1.714	2.069	2.500	2.807	23
24	1.318	1.711	2.064	2.492	2.797	24
25	1.316	1.708	2.060	2.485	2.787	25
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.703	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
inf.	1.282	1.645	1.960	2.326	2.576	inf.

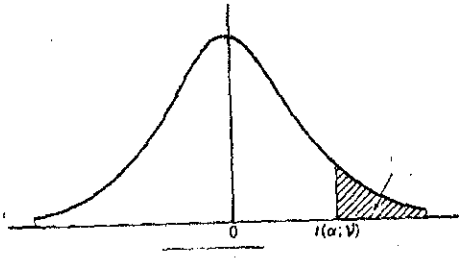
$X_i \sim N(\mu, \sigma^2)$

$(i = 1, 2, \ldots, n)$

* Abridged by permission of Macmillan Publishing Co., Inc., from *Statistical Methods for Research Workers*, 14th ed., by R. A. Fisher. Copyright © 1970 University of Adelaide.

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad \text{then}$$

$$S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 / (n-1)$$



ดังนั้น $T = \dfrac{\bar{X} - \mu}{S/\sqrt{n}} = \dfrac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{\dfrac{(n-1)S^2/(n-1)}{\sigma^2}}} = \dfrac{Z}{\sqrt{W/\nu}}$ where $\nu = n-1$

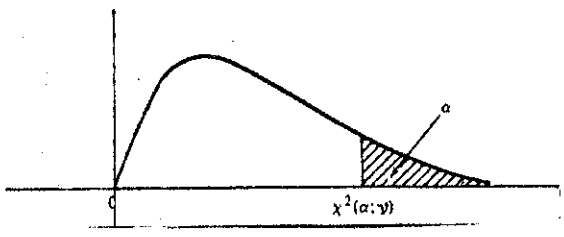
จะได้ว่า $Z \sim N(0,1)$ และ $W \sim \chi^2(\nu)$. T มี ν degree of freedom $= \nu = n-1$

ตามนิยามของ $P[T > t(\alpha; \nu)] = \alpha$

Table 5 Values of χ^2_α *

ν	$\alpha = 0.995$	$\alpha = 0.99$	$\alpha = 0.975$	$\alpha = 0.95$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$	ν
1	0.0000393	0.000157	0.000982	0.00393	3.841	5.024	6.635	7.879	1
2	0.0100	0.0201	0.0506	0.103	5.991	7.378	9.210	10.597	2
3	0.0717	0.115	0.216	0.352	7.815	9.348	11.345	12.838	3
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860	4
5	0.412	0.554	0.831	1.145	11.070	12.832	15.056	16.750	5
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548	6
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278	7
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955	8
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589	9
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188	10
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757	11
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300	12
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819	13
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319	14
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801	15
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267	16
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718	17
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156	18
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582	19
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997	20
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401	21
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796	22
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181	23
24	9.886	10.856	12.401	13.484	36.415	39.364	42.980	45.558	24
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928	25
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290	26
27	11.808	12.879	14.573	16.151	40.113	43.194	46.963	49.645	27
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993	28
29	13.121	14.256	16.047	17.708	42.557	45.772	49.588	52.336	29
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672	30
40	20.706	22.164	24.433	26.509	55.758	59.342	63.691	66.766	40
50	27.991	29.707	32.357	34.764	67.505	71.420	76.154	79.490	50
60	35.535	37.485	40.482	43.118	79.082	83.298	88.379	91.952	60
70	43.275	45.442	48.758	51.739	90.531	95.023	100.425	104.215	70
80	51.172	53.540	57.153	60.391	101.879	106.629	112.329	116.321	80
90	59.196	61.754	65.646	69.126	113.145	118.136	124.116	128.299	90
100	67.328	70.065	74.222	77.929	124.342	129.561	135.807	140.169	100

* This table is based on Table 8 of *Biometrika Tables for Statisticians*, Vol. 1, by permission of the *Biometrika* trustees.



ถ้า $X_i \sim N(\mu, \sigma^2)$
$(i = 1, 2, \ldots, n)$

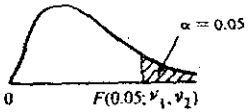
$W = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2}$; W is call $\chi^2(\nu = n-1)$

with n-1 degree of freedom.

ตามจำนวนตัวอย่าง

$P[W > \chi^2(\alpha;\nu)] = \alpha$

Table 6(a) Values of $F_{0.05}$*



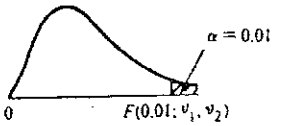
ν_2 = Degrees of freedom for denominator	ν_1 = Degrees of freedom for numerator																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.50	19.00	19.20	19.20	19.30	19.30	19.40	19.40	19.40	19.40	19.40	19.40	19.40	19.50	19.50	19.50	19.50	19.50	19.50
3	10.10	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.38	2.38	2.30	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	3.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.93
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

* This table is reproduced from M. Merrington and C. M. Thompson, "Tables of percentage points of the inverted beta (F) distribution", *Biometrika*, Vol. 33 (1943), by permission of the *Biometrika* trustees.

with ν_1 and ν_2 degrees of freedom resp. Then the ratio

$F = W/\nu_1 \,/\, Y/\nu_2$

Table 6(b) Values of $F_{0.01}$*



ν_2 = Degrees of freedom for denominator	ν_1 = Degrees of freedom for numerator																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4,052	5,000	5,403	5,625	5,764	5,859	5,928	5,982	6,023	6,056	6,106	6,157	6,209	6,235	6,261	6,287	6,313	6,339	6,366
2	98.50	99.00	99.20	99.20	99.30	99.30	99.40	99.40	99.40	99.40	99.40	99.40	99.40	99.50	99.50	99.50	99.50	99.50	99.50
3	34.10	30.80	29.50	28.70	28.20	27.90	27.70	27.50	27.30	27.20	27.10	26.90	26.70	26.60	26.50	26.40	26.30	26.20	26.10
4	21.20	18.00	16.70	16.00	15.50	15.20	15.00	14.80	14.70	14.50	14.40	14.20	14.00	13.90	13.80	13.70	13.70	13.60	13.50
5	16.30	13.30	12.10	11.40	11.00	10.70	10.50	10.30	10.20	10.10	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.70	10.90	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.20	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.30	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.83
9	10.60	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.00	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.53	2.45	2.36	2.27	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

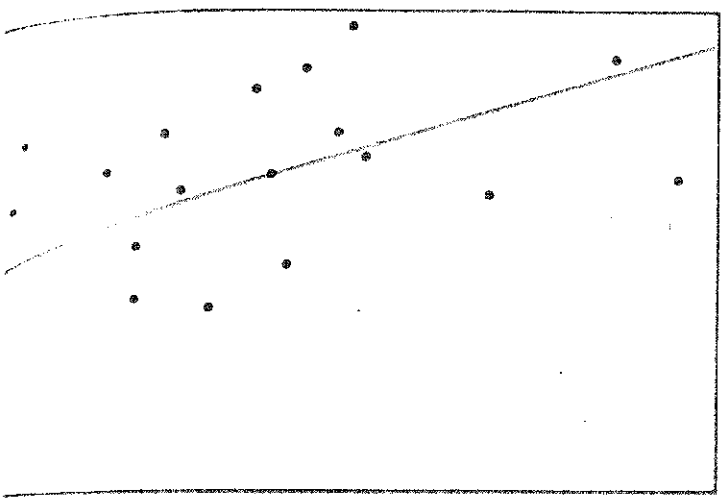
* This table is reproduced from M. Merrington and C. M. Thompson, "Tables of percentage points of the inverted beta (F) distribution," *Biometrika*, Vol. 33 (1943), by permission of the Biometrika trustees.

Appendix VII

RANDOM NUMBERS

51772	74640	42331	29044	46621	62898	93582	04186	19640	87056
24033	23491	83587	06568	21960	21387	76105	10863	97453	90581
45939	60173	52078	25424	11645	55870	56974	37428	93507	94271
30586	02133	75797	45406	31041	86707	12973	17169	88116	42187
03585	79353	81938	82322	96799	85659	36081	50884	14070	74950
64937	03355	95863	20790	65304	55189	00745	65253	11822	15804
15630	64759	51135	98527	62586	41889	25439	88036	24034	67283
09448	56301	57683	30277	94623	85418	68829	06652	41982	49159
21631	91157	77331	60710	52290	16835	48653	71590	16159	14676
91097	17480	29414	06829	87843	28195	27279	47152	35683	47280
50532	25496	95652	42457	73547	76552	50020	24819	52984	76168
07136	40876	79971	54195	25708	51817	36732	72484	94923	75936
27989	64728	10744	08396	56242	90985	28868	99431	50995	20507
85184	73949	36601	46253	00477	25234	09908	36574	72139	70185
54398	21154	97810	36764	32869	11785	55261	59009	38714	38723
65544	34371	09591	07839	58892	92843	72828	91341	84821	63886
08263	65952	85762	64236	39238	18776	84303	99247	46149	03229
39817	67906	48236	16057	81812	15815	63700	85915	19219	45943
62257	04077	79443	95203	02479	30763	92486	54083	23631	05825
53298	90276	62545	21944	16530	03878	07516	95715	02526	33537

CHAPTER 3



Looking at Data: Relationships

The focus of many statistical studies is on relationships between two or more variables. Did the Democratic party's share of presidential votes in each state change in any important way between the 1980 and 1984 elections? Does planting more corn on an acre of farm land change the yield at harvest, and if so, what is the best planting rate? Do groups of people who smoke more have a higher death rate from lung cancer? Such questions are our topic in this chapter, and these examples are among those that we will consider.

When you first examine the relationship between two variables, you should ask some preliminary questions. Most of these questions have already been raised in Chapter 1: What exactly are the variables? How were they measured? Are both variables quantitative or is at least one a categorical variable? *Quantitative variables* take numerical values for which numerical descriptions such as means and standard deviations are meaningful. Variables measured on a scale of equal units—such as length in centimeters or income in dollars—are quantitative variables. So are counts of individuals, and percents or fractions based on counts. *Categorical variables*, on the other hand, are essentially labels that tell us into which class an individual falls. The sex or occupation of a person, the make of a car, and the species of an animal are all categorical variables.

We have concentrated on quantitative variables to this point. But when data on several variables are being examined, categorical variables are usually present and are essential aids in organizing the data. Here is a small portion of a typical data set, as printed out by the statistical computing system used to analyze it.

OBS	ID	AGE	SEX	JOB	WT	SBP
1	1083	39	M	T	183	132
2	1381	27	F	E	116	117
3	1502	57	M	E	172	144
4	1481	26	M	T	139	110
5	1666	48	F	T	132	150

These data record medical measurements of the employees of a large company that offers regular free physical examinations. Each row gives data for one employee, or in statistical language, one case.

Case

A case is an individual person, animal, or thing for which values of the variables are recorded.

The employees are given an identification number (ID) so that their names will not be included on the printout. The computing system has numbered the cases consecutively under the heading OBS. Each column after the case number and ID contains the values of a specific variable. Five variables are

recorded for each case. Three of these—the employee's age, weight, and systolic blood pressure—are quantitative. The others are the employee's sex (male or female) and job category (executive or technical), which are categorical variables.

In studying relationships among variables, we must pay more attention to categorical variables than in earlier chapters. We are interested in relations between two quantitative variables (such as a person's weight and blood pressure), between a quantitative and a categorical variable (such as sex and blood pressure), and between two categorical variables (such as sex and job category). Some parts of this chapter (Sections 3.2 and 3.3) will focus on relations between quantitative variables. Categorical variables are considered in Section 3.4.

When we examine more than one variable, a new question becomes important. Is your purpose simply to explore the nature of the relationship, or do you hope to show that one of the variables can explain changes in the other? In looking at the Democrats' share of the popular vote for president in each state in 1980 and in 1984 (Table 3.1), we do not wish to explain the 1984 data by the 1980 data but rather to see a pattern that may reflect changing political conditions.[1] But in another case, the agronomists who carefully planted corn at different rates per acre and recorded the yield (Table 3.2) are indeed interested in a cause and effect relationship. They believe that the planting rate will affect the yield and their purpose is to recommend the best planting rate to farmers. In such cases, we distinguish the explanatory variable (plants per acre) from the response variable (yield of corn).

Response variable, explanatory variable

> A response variable measures an outcome of a study. An explanatory variable attempts to explain the observed outcomes.

In many studies, the goal is to show that changes in one or more explanatory variables actually cause changes in a response variable. For example, medical researchers studying a new drug to treat high blood pressure give different doses of the drug to each of several groups of patients and measure the change in blood pressure after several weeks of treatment. The explanatory variable is the dosage and the response variable is the change in blood pressure. The researchers hope to show that different doses of the drug cause changes in blood pressure. Not all explanatory–response relationships involve direct causation, however. Some of the statistical techniques in this chapter require us to distinguish explanatory from response variables; others make no use of this distinction. Explanatory variables are often called *independent variables*, and response variables are often referred to as *dependent variables*. The idea behind this language is that the response variable depends on the explanatory variable. However, since the words

"independent" and "dependent" have other meanings in statistics that are unrelated to the explanatory–response distinction, we prefer to avoid those words here.

The techniques used to study relations among variables are more complex than the methods we developed in Chapter 1 to examine the distribution of a single variable or those developed in Chapter 2 to examine the variation of a single variable over time. Fortunately, statistical analysis of several-variable data builds on the tools mastered in those chapters for examining individual variables. And the principles that guide our work remain the same:

- Combine graphical display with numerical summaries.
- Seek overall patterns and deviations from those patterns.
- Seek compact mathematical models for the data in addition to descriptive measures of specific aspects of the data.

Table 3.1 Percent of the presidential votes won by the Democratic candidate, 1980 and 1984

State	1980	1984	State	1980	1984
Ala.	48.7	38.7	Mont.	33.3	38.7
Alaska	30.1	30.9	Neb.	26.4	29.0
Ariz.	28.9	32.9	Nev.	27.9	32.7
Ark.	48.3	38.8	N.H.	28.6	31.1
Calif.	36.9	41.8	N.J.	39.2	39.5
Colo.	32.0	35.6	N.Mex.	37.5	39.7
Conn.	39.0	39.0	N.Y.	44.8	46.0
Del.	45.3	40.0	N.C.	47.5	38.0
Fla.	38.8	34.7	N.Dak.	26.7	34.3
Ga.	56.3	39.8	Ohio	35.5	40.5
Hawaii	45.6	44.3	Okla.	35.4	31.0
Idaho	25.7	26.7	Oreg.	40.1	43.9
Ill.	42.3	43.5	Pa.	43.1	46.3
Ind.	38.2	37.9	R.I.	48.1	48.2
Iowa	39.1	46.3	S.C.	48.6	35.9
Kan.	33.9	33.0	S.Dak.	32.1	36.7
Ky.	48.1	39.7	Tenn.	48.7	41.8
La.	46.4	38.6	Tex.	41.8	36.2
Maine	43.1	38.9	Utah	20.9	24.9
Md.	47.6	47.2	Vt.	39.3	41.3
Mass.	42.3	48.6	Va.	40.9	37.3
Mich.	43.1	40.4	Wash.	38.2	43.2
Minn.	47.7	50.1	W.Va.	50.1	44.7
Miss.	48.6	37.7	Wis.	44.0	45.4
Mo.	44.7	40.0	Wyo.	28.7	28.6

3.1 SCATTERPLOTS

Relationships between two quantitative variables are best displayed graphically. The most useful graph for this purpose is a *scatterplot*. Here is an example of the effective use of scatterplots.

EXAMPLE 3.1

Table 3.1 shows the percent of the popular vote that was won by the Democratic presidential candidates in the 1980 and 1984 elections. Both candidates, Jimmy Carter in 1980 and Walter Mondale in 1984, were defeated by the Republican Ronald Reagan. (In 1980 an independent candidate, John Anderson, captured 6.7% of the national vote.) We know that many states have persisting political traditions, so we expect similar behavior in two successive elections. It is possible to see this relationship in the columns of numbers in the table, but it is very difficult to assess the strength of the relationship or to see any significant changes from 1980 to 1984. A picture is needed.

Figure 3.1 is a scatterplot of the data in Table 3.1. Each point on the plot represents a single case—that is, a single state. The horizontal coordinate x_i is the percent who voted Democrat in that state's 1980 presidential vote. The vertical coordinate y_i is the percent who voted Democrat in 1984. Thus, Alabama appears as the point (48.7, 38.7), for example. Since both variables have the same units (percent), we use the same scale on both axes. The resulting plot outline is square. ■

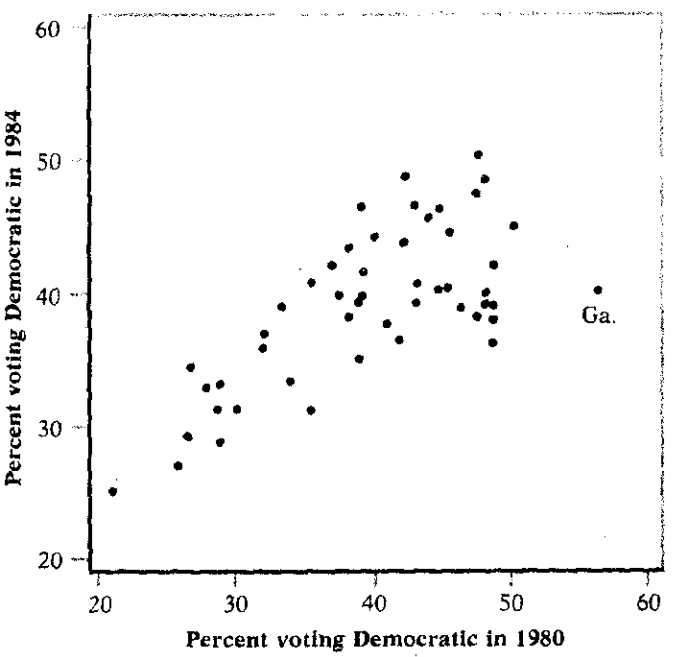


Figure 3.1 Percent of votes for Democrats in the 1980 and 1984 presidential elections, by state. See Example 3.1.

Interpreting Scatterplots

To interpret a scatterplot, look first for an overall pattern. This pattern should reveal the direction, form, and strength of the relationship between 1980 and 1984 voting in the states. The direction is clear from Figure 3.1: States with a high percent who voted Democrat in 1980 tended to also vote Democrat in 1984. That is, the percent voting Democrat in 1980 and in 1984 are positively associated.

Positive association, negative association

> Two variables are positively associated when above-average values of one tend to accompany above-average values of the other, and below-average values tend similarly to occur together. Two variables are negatively associated when above-average values of one accompany below-average values of the other, and vice versa.

In addition to the positive association, the scatterplot in Figure 3.1 shows the form of the relationship: It is roughly linear, though with much scatter about the linear pattern. The large scatter indicates that the linear relationship is not very strong.

The most important systematic deviation from the overall linear pattern in Figure 3.1 occurs at the right of the graph. A cluster of states there voted most heavily for Democrats in 1980, but were markedly less favorable to the Democrats in 1984. The single outlier on the extreme right is Georgia, President Carter's home state. Following that hint, we suspect that the south as a whole was more receptive to the southerner Carter in 1980 than to the northern liberal Mondale in 1984. To show this effect on the graph, Figure 3.2 uses a different symbol (o) to represent the 10 states south of Washington, D.C. and east of the Mississippi River; Louisiana, through which the river flows, is also included.

Figure 3.2 generally sustains our surmise about the south. In fact, we can refine our crude geographical definition of "the south" by examining the political behavior shown in Figure 3.2. The unemphasized point (•) in the middle of the southern cluster is Arkansas, which appears to be southern in voting pattern even though it lies west of the Mississippi. The two emphasized points (o) lying in the political mainstream (to the left of the cluster) are Florida and Virginia. Neither is fully southern in its behavior.

In dividing the states into "southern" and "nonsouthern," we introduced a third variable into the scatterplot. This is a categorical variable that has only two values. The two values are displayed by the two different plotting symbols. Using different symbols to plot points is a good way to incorporate a categorical variable into a scatterplot.[2]

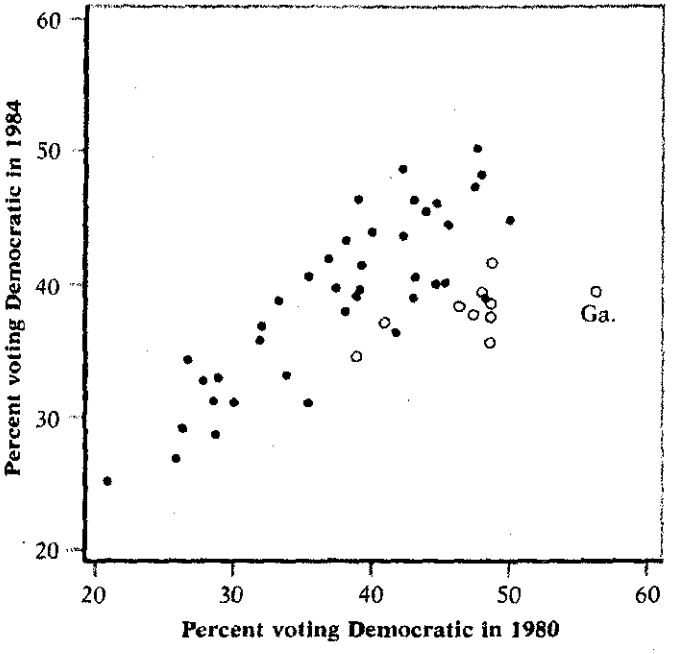


Figure 3.2 The 1980 and 1984 percent of votes for Democrats, with the south emphasized (○).

EXAMPLE 3.2

News reports sometimes rate state educational systems by comparing the mean scores of seniors in each state on college entrance examinations. This method is misleading, because the percent of high school seniors who take any particular college entrance test varies greatly from state to state. Figure 3.3 is a scatterplot of the mean score on the Scholastic Aptitude Test (SAT) mathematics examination for high school seniors in each state versus the percent of graduates in each state who took the test.

The *negative association* between these variables is evident: SAT scores tend to be lower in states where the percent of students who take the test is higher. Only 3% of the seniors in the three highest-scoring states took the SAT. The overall pattern shows two *clusters* of points. At the upper left are states where only students seeking admission to colleges that require the SAT take that test. Most students in these states take a different college entrance examination, the American College Testing (ACT) examination. The students who elect to take the SAT tend to be above average academically, so the mean SAT scores for states in this cluster are high. The other cluster, at the lower right of the scatterplot, contains states in which a high percent of college-bound seniors take the SAT. The mean scores are lower here because a less selective group of students take the test. There is little difference in mean SAT scores among these states, even though the percent of seniors who take the test varies from 30% to nearly 70%.

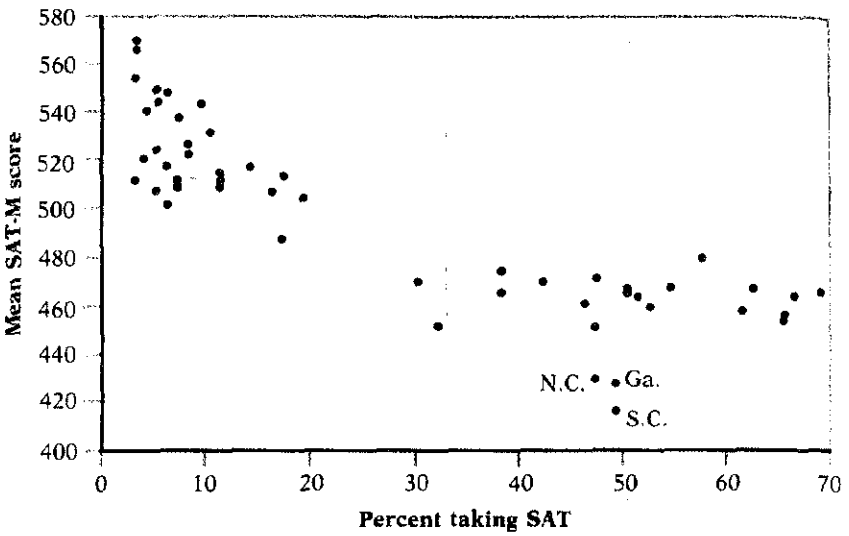


Figure 3.3 Mean SAT mathematics score and percent of high school seniors who took the test, by state. See Example 3.2.

The points that are individually labeled seem to lie a bit outside the lower cluster. The mean SAT scores in these states are lower than in other states in which a similar percent of students take the test. It is possible that the test scores point to educational deficiencies in these states. ■

Scatterplots showing relationships between other quantitative variables may appear quite different from the clouds of points in Figures 3.1 and 3.3. This is particularly true in experiments in which measurements of a response variable are taken at only a few selected levels of the explanatory variable. The following example illustrates the use of scatterplots in such a setting:

EXAMPLE 3.3

How much corn per acre should a farmer plant to obtain the highest yield? Too few plants will clearly result in a low yield. On the other hand, if there are too many plants they will compete with each other for moisture and nutrients, and yields will again fall. The amount of moisture is critical: In dry seasons the best planting rate is lower than in wet seasons. Rather than try to forecast the weather, a farmer can ensure even moisture by irrigating his fields. Table 3.2 shows the results of several years of field experiments in Nebraska.[3] Each entry is the mean yield of four small plots planted at the indicated rate per acre. All plots were irrigated, and all were fertilized and cultivated identically. The yield of each plot should therefore depend only on the planting rate—and of course on the uncontrolled aspects of each growing season such as temperature and wind. The experiment lasted several years in order to avoid misleading conclusions due to the peculiarities of a single growing season.

Table 3.2 Average irrigated corn yields (bushels per acre)

Plants per acre	1956	1958	1959	1960	Mean
12,000	150.1	113.0	118.4	142.6	131.0
16,000	166.9	120.7	135.2	149.8	143.2
20,000	165.3	130.1	139.6	149.9	146.2
24,000		134.7	138.4	156.1	143.1
28,000			119.0	150.5	134.8
Mean	160.8	124.6	130.1	149.8	

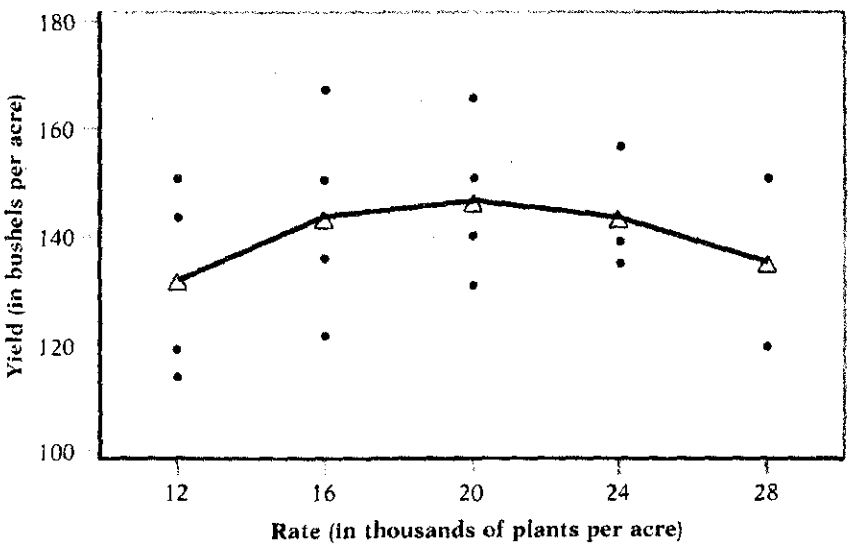


Figure 3.4 Yield of corn versus planting rate for irrigated cornfields in Nebraska. See Example 3.3.

The scatterplot in Figure 3.4 displays the results of this experiment. Since planting rate is the explanatory variable, it is plotted horizontally as the x variable. The yield is the vertical variable y. The vertical spread of points over each planting rate shows the year-to-year variation. The overall pattern is revealed by plotting the mean yield for each planting rate (averaged over all years). These means are marked by triangles and joined by line segments. As expected, the form of the relationship is not linear. Yields first increase with the planting rate, then decrease when too many plants are crowded in. Since there is no consistent direction, we cannot describe the association as either positive or negative. It appears that the best choice is about 20,000 plants per acre. ■

Plotting the means as was done in Figure 3.4 is an excellent way to summarize an experiment in which repeated observations are made at a few fixed levels of a variable such as planting rate. The scatterplot also indicates how much variability lies behind each mean. In this case, however, we must regard the means with great caution. As the gaps in Table 3.2 show, the agronomists used only the three lowest planting rates in all 4 years, adding the higher rates only as the need to study them became apparent. The means therefore cover different spans of time. In particular, 1956 was a year of very high yields, so that the means that include 1956 are biased upward relative to those that do not.

A closer look at Table 3.2 shows that 24,000 plants per acre was superior to 20,000 in 2 of the 3 years in which both rates were planted. The agronomists concluded that maximum yield is already approached at 16,000 plants per acre, but yields continue to increase somewhat up to 24,000 plants per acre. *Incomplete data*, such as those in Example 3.3, are common in practice. The agronomists' data would have been more convincing if all five planting rates had been used in all 4 years—but it was only after the first results were in that the researchers realized that higher planting rates might give better results. Incomplete data often complicate a statistical analysis, whether it is an informal analysis such as that of Example 3.3 or a formal analysis of the type that we will learn about in Chapter 11. You should therefore be alert to missing data.

Examples 3.1 and 3.3, though different in most respects, share a most important feature: *The relationship between the two variables plotted cannot be fully understood without knowledge about a third variable.* Figure 3.1 plots the percent of the presidential vote won by the Democrats in 1980 versus 1984 by state, but the relationship observed is partly explained by a political and geographic grouping of the states. Figure 3.4 plots corn yield versus planting rate, but the experiment spanned several growing seasons that differed from each other. And if the agronomists had not carefully controlled many other variables (moisture, fertilizer, etc.), these variables would have confused the situation completely. You should be cautious in drawing conclusions from a strong relationship appearing in a scatterplot until you understand what other variables may be lurking in the background.

Smoothing Scatterplots

The strength of a scatterplot is that it provides a complete picture of the relationship between two variables—at least as far as that relationship is reflected in the available data. A complete picture is often too detailed for easy interpretation, so we seek to describe the plot in terms of an overall pattern and deviations from that pattern. Though we can often do this by eye, more systematic methods of extracting the overall pattern are desirable. This is called *smoothing* a scatterplot. When we are plotting a response variable y against an explanatory variable x, Example 3.3 gives us a hint as to

how to proceed. We smoothed Figure 3.4 by averaging the y values separately for each x value. Though not all scatterplots have many y values at the same value of x, as did Figure 3.4, we can smooth a scatterplot by slicing it into vertical strips and computing the mean or median y in each strip. For initial analysis, the median is preferred to the mean because it is more resistant.

Median trace

To construct the median trace of a scatterplot, first slice the plot into vertical strips of equal width. Compute the median of the y values in each strip and plot the median vertically above the horizontal midpoint of the strip. Connect the medians by straight line segments to form the median trace.

The median trace displays the overall pattern of the dependence of y on x. The scatter of the observations above and below the median trace displays deviations from the pattern.

EXAMPLE 3.4

Prior to 1970, American young men were drafted into military service by local draft boards who followed a complex system of preferences and exemptions. Congress decided that a random selection process—a draft lottery—would be fairer. The first draft lottery was held in 1970. The 366 possible birth dates were placed into identical plastic capsules, poured into a rotating drum, and picked out one by one. The first birth date drawn won draft number 1, the next 2, and so on. Eligible men were then drafted in order of their draft numbers, those with the lowest numbers first.

The outcome of the 1970 draft lottery appears in the scatterplot in Figure 3.5. Birth dates, numbered 1 to 366 beginning with January 1, are plotted horizontally. The draft number assigned each date by the lottery is plotted on the vertical scale. A properly conducted lottery should produce *no* systematic relationship between these variables. Figure 3.5 certainly shows no clear overall pattern. Yet it was charged that the lottery was biased against men born late in the year, that these men received systematically lower draft numbers than men born earlier. An investigation showed that birth dates had indeed been inserted into capsules and poured into a box 1 month at a time before being placed into the drum.

With this hint that a month effect might be present, we will smooth Figure 3.5 by a median trace. Imagine that the scatterplot is sliced vertically like a loaf of bread. Each slice contains 1 month's birth dates. We calculate the median draft number for each month, and plot it vertically above the horizontal midpoint of each slice. Figure 3.6 shows the median trace superimposed on the scatterplot. The downward trend late in the year is now apparent. In 1971, the Department of Defense reassigned the officers who had conducted the 1970 lottery and asked statisticians from the National Bureau of Standards to design a truly random selection procedure.[4] ■

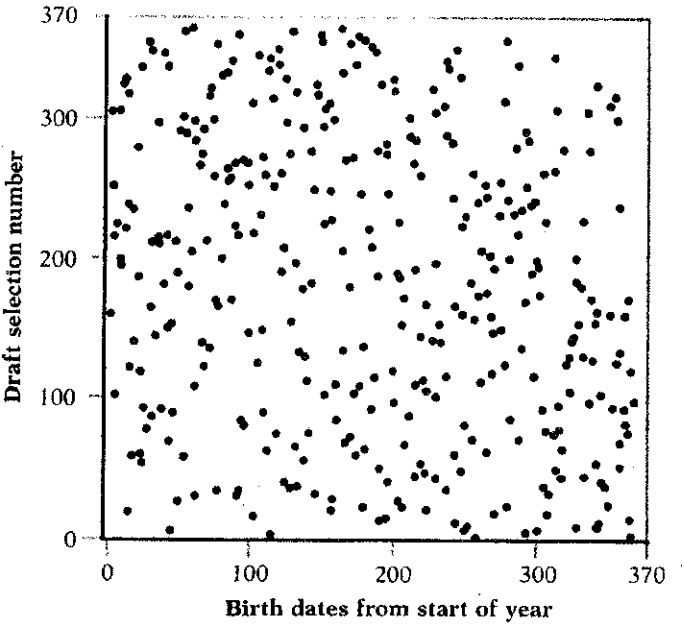


Figure 3.5 The 1970 draft lottery: draft selection numbers versus birth dates. See Example 3.4.

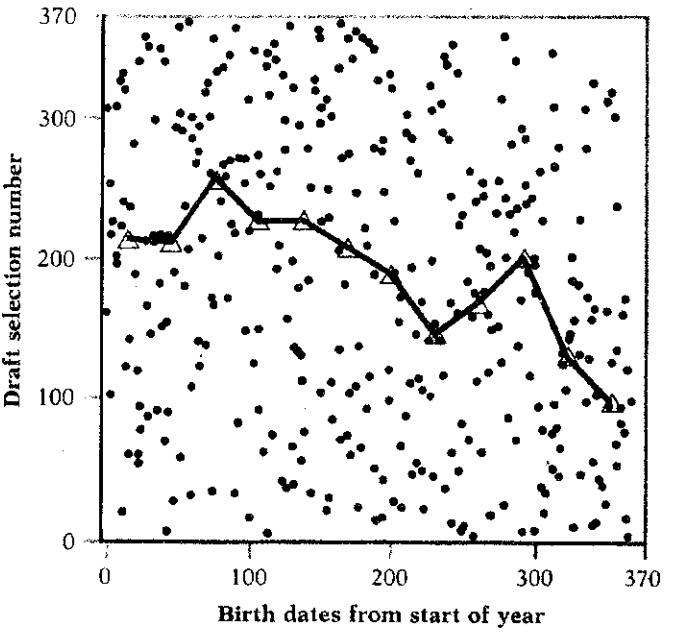


Figure 3.6 A median trace by month of birth for the 1970 draft lottery.

Example 3.4 demonstrates that smoothing can reveal relationships that are not obvious from a scatterplot alone. In this case there is a *negative association* between birth date and draft number: Later birth dates tend to have lower draft numbers. Once the median trace shows us what to look for, we can see that the point cloud in Figure 3.5 is a bit thin in the upper right region, indicating that men born late in the year won few high draft numbers. The combination of graphing and calculating once again proves its effectiveness.

To draw a median trace, first slice the scatterplot. The number of slices chosen determines the degree of smoothing provided by the median trace. Fewer slices smooth the data more, while more slices allow the median trace to follow the ups and downs of y more closely. In general, the number of slices should increase with the number of cases; beyond this, choosing the slices is a matter of judgment. It is often best to take advantage of naturally occurring slices such as the months in Example 3.4.

Slicing a scatterplot into vertical strips is the basis for other graphical displays as well. Figure 3.7 shows *boxplots* for each of the monthly slices of the draft lottery data. The sequence of medians shows the smoothed pattern; the quartiles and extremes show the scatter about the pattern portrayed by the individual points in Figure 3.6. Smoothing a scatterplot either by the median trace or by side-by-side boxplots thus displays both an overall pattern and deviations from it.

The 1970 draft lottery raises one final question. How can we be confident that the lower draft numbers assigned to men born later in the year were not merely the play of chance? After all, repeated random drawings of birth

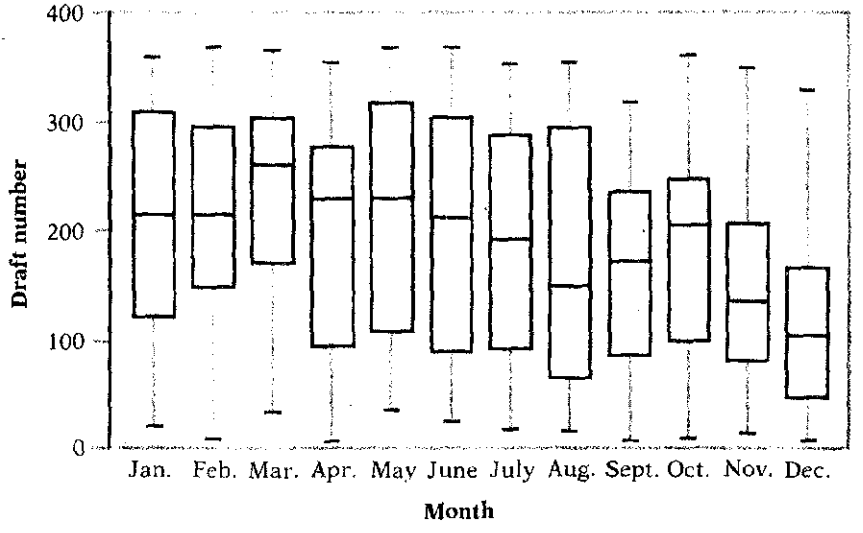


Figure 3.7 Boxplots by month for selection numbers in the 1970 draft lottery.

dates would give a different order each time. Some would appear—after the fact—to favor January, and others to favor December. So the pattern that Figure 3.5 seemed to show may be an accident. Quite true. We must judge whether this pattern is stronger than could reasonably arise from the play of chance in a truly random drawing. This judgment requires a calculation of probabilities. Such a calculation shows that an association between birth date and draft number as strong as the one observed in 1970 would occur less than once in 1000 random drawings. There is in fact good evidence that the 1970 lottery was unfair.

Categorical Explanatory Variables

Variations on scatterplots are also the preferred means for showing relations between a categorical explanatory variable and a quantitative response. These displays are very similar to those already discussed. Suppose that the agronomists of Example 3.3 had compared the yields of five varieties of corn rather than five planting rates. The plot in Figure 3.4 remains helpful if the varieties A, B, C, D, and E are marked at equal intervals on the horizontal axis in place of the planting rate. In particular, a graph of the mean (or median) responses for each category will show the overall nature of the relationship. If there are too many observations in each category to plot individually, as in Figure 3.4, side-by-side boxplots or stemplots can replace the scatterplot of response values above each category label in the graph. Figure 1.9 is such a graph. There the categorical explanatory variable is hot dog type (beef, meat, or poultry), and the response is the number of calories in each hot dog.

Many categorical variables, like corn variety or hot dog type, have no natural order from smallest to largest. In this case we cannot speak of a positive or negative association with the response variable. If the mean responses increase as we go from right to left in the plot, we could make them decrease by writing the categories in the opposite order. In such cases the plot simply presents a side-by-side comparison of several distributions. The categorical variable labels the distributions. Some categorical variables do have a least-to-most order. We can then speak of the direction of the association between the categorical explanatory variable and the quantitative response. Here is an example.

EXAMPLE 3.5

What is the relationship between family income and the educational level of the householder? (In government records, the householder is the adult who owns or rents a dwelling.) Educational level is the explanatory variable. It appears in government data as a categorical variable with these values:

A Less than 8 years of elementary school

B 8 years of elementary school

C 1 to 3 years of high school

D 4 years of high school

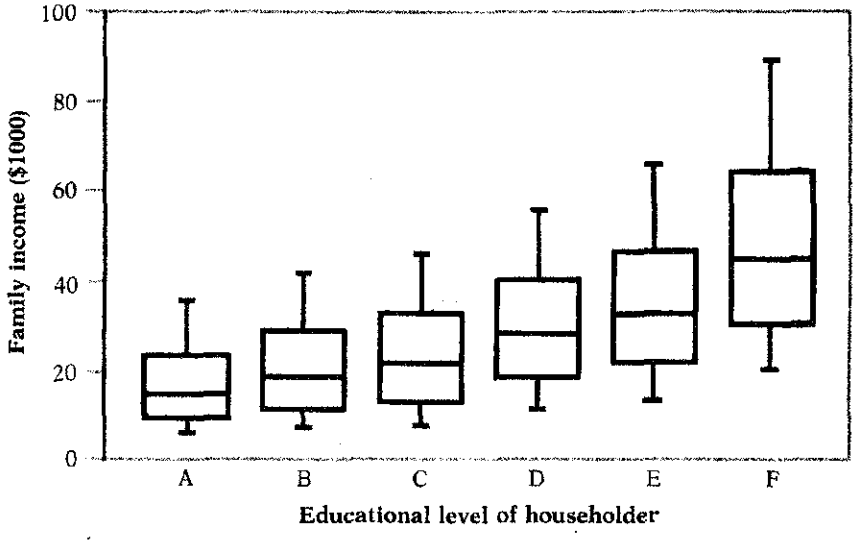


Figure 3.8 Boxplots by educational level for family incomes in 1985. See Example 3.5.

E 1 to 3 years of college

F 4 or more years of college

The categories A to F are ordered from least education to most education. The side-by-side boxplots of Figure 3.8 show the distributions of family income in 1985 for white families in these educational categories. The plot shows a positive association between educational level and income. That is, family income tends to increase as the educational level of the householder increases. What is more, the variability of income also increases with education. Some college-educated households have quite low incomes. Additional education offers the opportunity to earn more but does not guarantee a high income. ■

SUMMARY

A **categorical variable** records into which of two or more groups an observation falls, while a **quantitative variable** takes numerical values for which arithmetic operations make sense.

When changes in a variable x are thought to explain or even cause changes in a second variable y, x is called an **explanatory variable** and y is called a **response variable**.

A **scatterplot** is a plot of observations x_i and y_i as points in the plane, where x_i and y_i are the values of quantitative variables x and y for the same **case**, that is, the same person, animal, or object.

The explanatory variable, if any, is always plotted on the horizontal scale of a scatterplot. Plotting points with different symbols allows us to see the effect of a categorical variable in a scatterplot.

In examining a scatterplot, look for an overall pattern showing the form, direction, and strength of the relationship, and then for outliers or other deviations from that pattern. **Linear relationships** are an important form.

If the relationship has a clear direction, we speak of either **positive association** or **negative association**.

Smoothing a scatterplot by a **median trace** or other method helps reveal the nature of the dependence of y on x.

SECTION 3.1 EXERCISES

3.1 In each of the following cases, tell whether the variable is quantitative or categorical:

(a) The name of the manufacturer of a TV set
(b) The number of insects on a corn plant
(c) The score on a test of math anxiety for a student taking a statistics course
(d) The major area of study for the student in (c)
(e) The number of pages in a book
(f) Your height in inches

3.2 In each of the following cases, tell whether you would be interested simply in exploring the relationship between the two variables or whether you would want to view one of the variables as an explanatory variable and the other as a response variable. In the latter case, state which is the explanatory variable and which is the response variable.

(a) The amount of time spent studying for a statistics exam and the grade on the exam
(b) The height and weight of a person
(c) The amount of yearly rainfall and the yield of a crop
(d) A student's scores on the SAT math exam and on the SAT verbal exam
(e) The occupational class of a father and of his son

3.3 Vehicle manufacturers are required to test their vehicles for the amount of each of several pollutants in the exhaust. Even among identical vehicles the amount of a pollutant varies, so several vehicles must be tested. Figure 3.9 plots the amounts of two pollutants, carbon monoxide and nitrogen oxides, for 46 identical vehicles. Both variables are measured in grams of the pollutant per

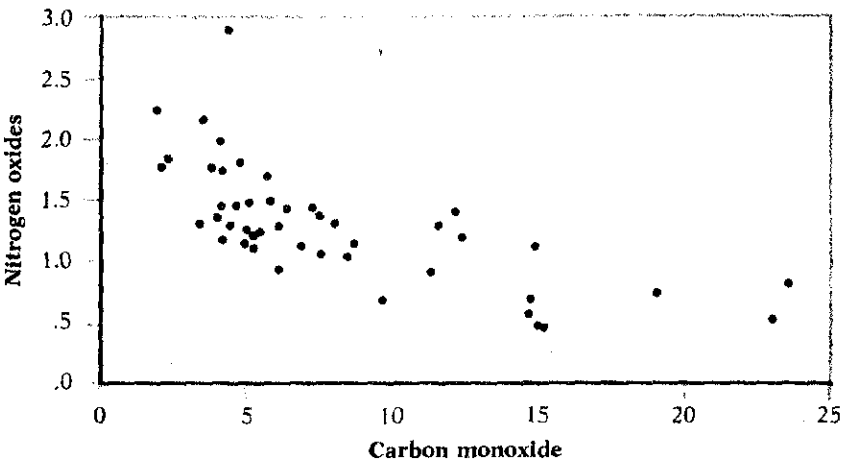


Figure 3.9 Nitrogen oxides versus carbon monoxide in the exhaust of 46 vehicles (Exercise 3.3).

mile driven. (Data from Thomas J. Lorenzen, "Determining statistical characteristics of a vehicle emissions audit procedure," *Technometrics*, 22 (1980), pp. 483–493.)

(a) Describe the nature of the relationship. Is the association positive or negative? Is it nearly linear or clearly curved? Are there any outliers?

(b) A writer on automobiles says, "When an engine is properly built and properly tuned, it emits few pollutants. If the engine is out of tune, it emits more of all the important pollutants. You can find out how badly a vehicle is polluting the air by measuring any one pollutant. If that value is acceptable, the other emissions will also be OK." Do the data in Figure 3.9 support this claim?

3.4 The following are the golf scores of 12 members of a women's golf team in two rounds of tournament play. (A golf score is the number of strokes required to complete the course, so low scores are better.)

Player	1	2	3	4	5	6	7	8	9	10	11	12
Round 1	89	90	87	95	86	81	102	105	83	88	91	79
Round 2	94	85	89	89	81	76	107	89	87	91	88	80

(a) Plot the scores from round 2 versus the scores from round 1.

(b) Is there an association between the two scores? If so, is it positive or negative? Explain why you would expect scores in

two rounds of a tournament to have an association like the one you observed.

(c) A good golfer can have an unusually bad round and a weaker golfer can have an unusually good round. Either of these situations could produce an outlier. Circle the outlier in your scatterplot. Can you tell from the data given whether the unusual value is produced by a good player or a poor player? What other data would you need to distinguish between the two possibilities?

3.5 When water flows across farm land, some of the soil is washed away, resulting in erosion. An experiment was conducted to investigate the effect of the rate of water flow on the amount of soil washed away. Flow is measured in liters per second and the eroded soil is measured in kilograms. The data are given in the following table. (From G. R. Foster, W. R. Ostercamp, and L. J. Lane, *Effect of Discharge Rate on Rill Erosion*, presented at the 1982 Winter Meeting of the American Society of Agricultural Engineers.)

Flow rate	.31	.85	1.26	2.47	3.75
Eroded soil	.82	1.95	2.18	3.01	6.07

(a) Plot the data.

(b) Describe the pattern that you see. Would it be reasonable to describe the overall pattern by a straight line? Is the association positive or negative?

3.6 McDonald's "Big Mac" hamburger is sold in many countries around the world. By comparing the cost of a Big Mac in the local currency to its cost in the United States, we can find the exchange rate between the American dollar and that currency that would make the cost of a Big Mac the same in both countries. *The Economist* did this, and compared the result with the actual exchange rates between the dollar and foreign currencies. The following table gives some of the data. The entries are the value of 1 dollar in foreign currency; for example, at the official exchange rate 1 U.S. dollar was worth 1.64 Australian dollars. (From *The Economist*, Sept. 6, 1986, p. 77.)

(a) Make a scatterplot of the official exchange rate y versus the Big Mac exchange rate x.

(b) Describe the overall pattern of your scatterplot. Is the association positive or negative? Is there a clearly linear pattern? Are there any distinct outliers? How well does comparing the prices of a Big Mac predict the official value of the dollar in foreign currencies?

Section 3.1 Exercises

Country	Big Mac dollar value	Official dollar value
Australia	1.09	1.64
Brazil	7.80	13.80
Britain	.69	.67
Canada	1.18	1.39
France	10.30	6.65
Hong Kong	4.75	7.80
Ireland	.74	.74
Holland	2.72	2.28
Singapore	1.75	2.15
Sweden	10.30	6.87
W. Germany	2.66	2.02

3.7 Do heavier cars cost more than lighter cars? Figure 3.10 is a plot of the base price in dollars and the weight in pounds for all 1986 model four-door sedans listed in an auto guide. Cars made by American manufacturers are plotted with a "•" and cars of foreign make are plotted with an "o."

(a) Describe the overall relationship between the weight of a car and its price. Is the association strong (weight and price closely connected) or weak? Is it generally positive or negative?

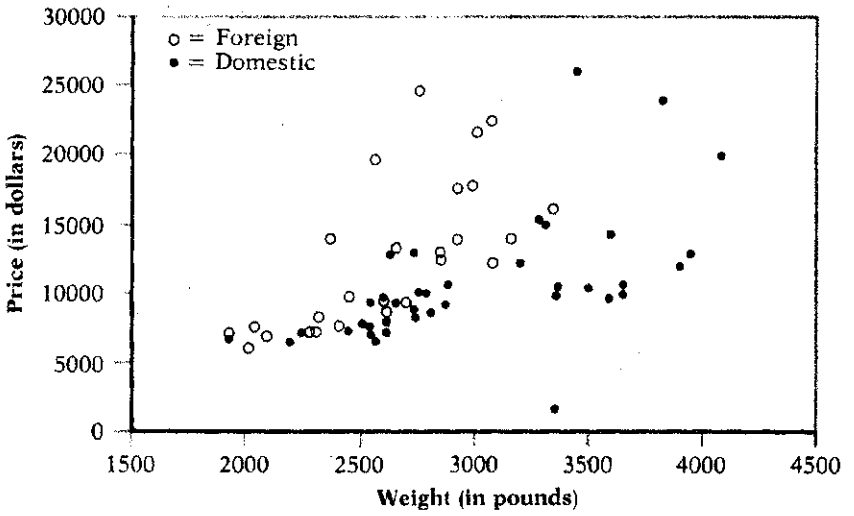


Figure 3.10 Base price in dollars versus weight in pounds for 1986 four-door sedans (Exercise 3.7).

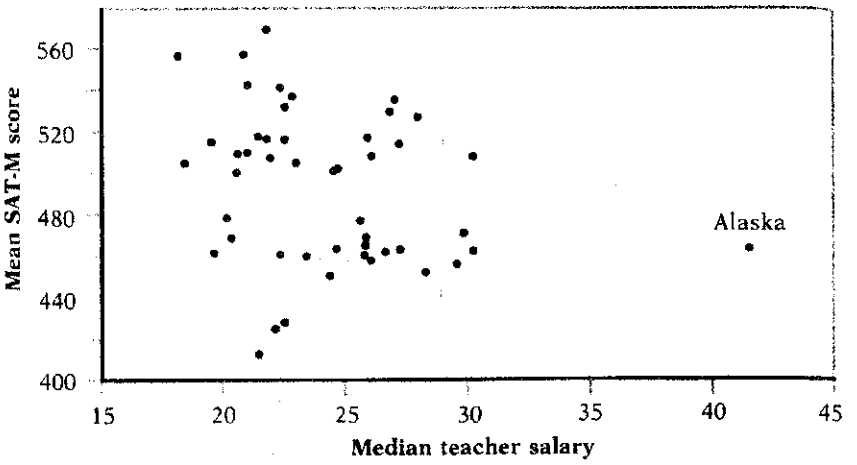


Figure 3.11 Mean SAT mathematics score versus median teacher salary by state (Exercise 3.8).

(b) Describe the major differences between domestic and foreign cars as they appear in the plot.

3.8 Does increasing public spending on education improve student performance? Figure 3.11 plots one measure of academic performance, the mean SAT mathematics test score in each state, against one measure of spending, the median salary paid to teachers in the state. The outlier is identified as Alaska.

(a) Is Alaska an outlier in both variables or only in one? Why do you think Alaska is an outlier?

(b) Describe the overall pattern of the relationship between teachers' salaries and students' SAT mathematics scores.

(Experts in education say there is not a clear association between spending on public education and SAT scores. As Example 3.2 shows, the percent of high school students who take the SAT varies greatly from state to state, so the mean SAT score is based on different kinds of students in different states.)

3.9 The following data refer to an outbreak of botulism, a form of food poisoning that may be fatal. Each case is a person who contracted botulism in the outbreak. The variables recorded are the subject's age in years, the incubation period (the time in hours between eating the infected food and the first signs of illness), and whether the subject survived (S) or died (D). (Modified from data provided by Professor Dana Quade, University of North Carolina, Chapel Hill, N.C.)

Section 3.1 Exercises

Case	1	2	3	4	5	6	7	8	9
Age	29	39	44	37	42	17	38	43	51
Incubation	13	46	43	34	20	20	18	72	19
Outcome	D	S	S	D	D	S	D	S	D
Case	10	11	12	13	14	15	16	17	18
Age	30	32	59	33	31	32	32	36	50
Incubation	36	48	44	21	32	86	48	28	16
Outcome	D	D	S	D	D	S	D	S	D

(a) Make a scatterplot of incubation period against age, using different symbols for cases that were fatal and cases where the victim survived.

(b) Is there an overall relationship between age and incubation period? If so, describe it.

(c) More important, is there a relationship between either age or incubation period and whether the victim survived? Describe any relations that seem important here.

(d) Are there any unusual cases that may require individual investigation?

3.10 We wish to predict the level of nitrogen oxide (NOX) emissions from the level of carbon monoxide (CO) emissions for a vehicle of the type considered in Exercise 3.3. Figure 3.9 displays a moderately strong relationship between the two variables. To construct a median trace, we compute the median NOX level for the vehicles in each 5-g/mile-wide slice of CO levels. Here is the result of these calculations:

CO level	Count	Median NOX
$0 \le CO < 5$	16	1.785
$5 \le CO < 10$	18	1.270
$10 \le CO < 15$	8	1.055
$15 \le CO < 20$	2	.635
$20 \le CO < 25$	2	.715

Draw the median trace on a graph with the same scales as those used in Figure 3.9. Describe the overall relation displayed by the median trace.

3.11 Refer back to the flea data given in Exercise 1.19 for Thomas the cat.

(a) Divide the data into consecutive 5-day periods and find the median for each period (discard the data for the final two days).

Plot all of the data and superimpose the median trace. Describe the pattern displayed by the trace.

(b) Repeat (a) using 3-day periods. For this problem do you prefer 5-day periods or 3-day periods? Why?

3.12 Compute the median trace for 5-day periods from the flea data given in Exercise 2.57 for Creole the cat. Compare the overall patterns of flea reproduction for Creole and Thomas as displayed by the median traces.

3.13 To demonstrate the effect of nematodes (microscopic worms) on plant growth, a botanist prepares 16 identical planting pots and then introduces different numbers of nematodes into the pots. A tomato seedling is transplanted into each plot. Here are data on the increase in height of the seedlings (in centimeters) 16 days after planting. (Data provided by Matthew Moore.)

Nematodes	Seedling growth			
0	10.8	9.1	13.5	9.2
1000	11.1	11.1	8.2	11.3
5000	5.4	4.6	7.4	5.0
10,000	5.8	5.3	3.2	7.5

(a) Make a scatterplot of the response variable (growth) against the explanatory variable (nematode count). Then compute the mean growth for each group of seedlings, plot the means against the nematode counts, and connect these four points with line segments.

(b) Briefly describe the conclusions about the effects of nematodes on plant growth that these data suggest.

3.14 The presence of harmful insects in farm fields is detected by erecting boards covered with a sticky material and then examining the insects trapped on the board. Some colors are more attractive to insects than others. In an experiment aimed at determining the best color for attracting cereal leaf beetles, six boards in each of four colors were placed in a field of oats in July. The following table gives data on the number of cereal leaf beetles trapped. (Modified from M. C. Wilson and R. E. Shade, "Relative attractiveness of various luminescent colors to the cereal leaf beetle and the meadow

Board color	Insects trapped					
Lemon yellow	45	59	48	46	38	47
White	21	12	14	17	13	17
Green	37	32	15	25	39	41
Blue	16	11	20	21	14	7

spittlebug," *Journal of Economic Entomology*, 60 (1967), pp. 578–580.)

(a) Make a plot of the counts of insects trapped against board color (space the four colors equally on the horizontal axis). Compute the mean count for each color, add the means to your plot, and connect the means with line segments.

(b) Based on the data, state your conclusions about the attractiveness of these colors to the beetles.

(c) Does it make sense to speak of a positive or negative association between board color and insect count?

3.15 When animals of the same species live together, they often establish a clear "pecking order." Lower ranking individuals defer to higher ranking animals in many ways, usually avoiding open conflict. A researcher on animal behavior wants to study the relationship between pecking order and physical characteristics such as weight. He confines four chickens in each of seven pens and observes the pecking order that emerges in each pen. Here is a table of the weights (in grams) of the chickens, arranged by pecking order. That is, the first row gives the weights of the dominant chicken in the seven pens, the second row the weights of the number 2 chicken in each pen, and so on. (Data collected by D. L. Cunningham, Cornell University, Ithaca, N.Y.)

	Weight (g)						
Pecking order	Pen 1	Pen 2	Pen 3	Pen 4	Pen 5	Pen 6	Pen 7
1	1880	1300	1600	1380	1800	1000	1680
2	1920	1700	1830	1520	1780	1740	1460
3	1600	1500	1520	1520	1360	1520	1760
4	1830	1880	1820	1380	2000	2000	1800

(a) Make a plot of these data that is appropriate to study the effect of weight on pecking order. Include in your plot any means that might be helpful.

(b) We might expect that heavier chickens would tend to stand higher in the pecking order. Do the data give clear evidence for or against this expectation?

3.16 The tensile modulus of elasticity is a measure of strength for wood products. In an experiment 50 strips of yellow poplar wood were each measured twice. The units of measurement are millions of pounds per square inch, the two measures are denoted by T1 and T2, and the number of the strip is denoted by S. The data are given in the following table. Analysis of this larger data set is best done using statistical software. (Data provided by Michael Triche and

Professor Michael Hunt, Forestry Department, Purdue University, West Lafayette, Ind.)

S	T1	T2	S	T1	T2	S	T1	T2	S	T1	T2
1	1.58	1.55	2	1.53	1.54	3	1.38	1.37	4	1.24	1.25
5	1.59	1.57	6	1.52	1.51	7	1.47	1.45	8	1.45	1.44
9	1.54	1.53	10	1.49	1.49	11	1.51	1.51	12	1.60	1.58
13	1.34	1.32	14	1.33	1.33	15	1.72	1.70	16	1.63	1.62
17	1.51	1.50	18	1.42	1.42	19	1.90	1.87	20	1.81	1.79
21	1.88	1.86	22	1.68	1.65	23	1.56	1.54	24	1.68	1.67
25	1.75	1.74	26	1.63	1.62	27	1.65	1.65	28	1.76	1.75
29	1.58	1.56	30	1.69	1.67	31	1.60	1.58	32	1.57	1.56
33	1.70	1.69	34	1.41	1.40	35	1.33	1.33	36	1.59	1.59
37	1.64	1.66	38	1.74	1.74	39	1.89	1.91	40	1.73	1.75
41	1.80	1.81	42	1.67	1.68	43	1.63	1.64	44	1.87	1.88
45	2.02	2.04	46	1.79	1.81	47	1.86	1.88	48	1.72	1.73
49	1.83	1.83	50	1.68	1.70						

(a) Plot T2 versus T1.
(b) Is the association between T1 and T2 positive or negative? Is this the direction of association that you would expect before looking at your plot? Why?
(c) Plot T1 versus S and T2 versus S. Do you see any pattern? If so, describe it. What questions would you ask the experimenters in seeking an explanation for the pattern observed?

3.2 LEAST SQUARES REGRESSION

When we smooth a scatterplot by using a median trace or some other tool, we are attempting to summarize the dependence of the response y on the explanatory variable x without specifying what form that dependence should take. A single equation that describes the dependence of y on x provides a more compact summary. The simplest such equation is a linear equation of the form

$$y = a + bx$$

whose graph is a straight line. A strong linear pattern in a scatterplot can be summarized by a linear equation.

EXAMPLE 3.6 Warren heats his home with natural gas. The amount of gas required to heat the home depends on the outdoor temperature—the colder the weather, the more gas will be consumed. As long as the family's habits, the insulation of the house,

and other such factors do not change, Warren should be able to predict gas consumption from the outdoor temperature. He therefore measures his household's natural gas consumption each month during one heating season, from October to the following June. Because months differ in length, he divides gas consumption by the number of days in the month to obtain the daily gas consumption.

Outdoor temperature influences gas consumption only when it is cold enough to require heating. The usual measure of the need for heating is *heating degree days*. To find the number of heating degree days for a given day, first record the high and low temperature for the day. The average temperature is taken to be the mean of the high and low temperatures. One heating degree day is accumulated for each degree this average falls below 65°F. An average temperature of 20°F, for example, corresponds to 45 degree days. Warren obtains the average number of degree days per day for each month from weather records for his community.

Here are the data.[5] The explanatory variable x is heating degree days per day for the month, and the response variable y is gas consumption per day in hundreds of cubic feet.

Month	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June
x	15.6	26.8	37.8	36.4	35.5	18.6	15.3	7.9	.0
y	5.2	6.1	8.7	8.5	8.8	4.9	4.5	2.5	1.1

A scatterplot (Figure 3.12) shows that the relationship is strongly linear. The deviations from the straight-line pattern reflect the use of gas for cooking, windows left open, and other influences on gas consumption. ■

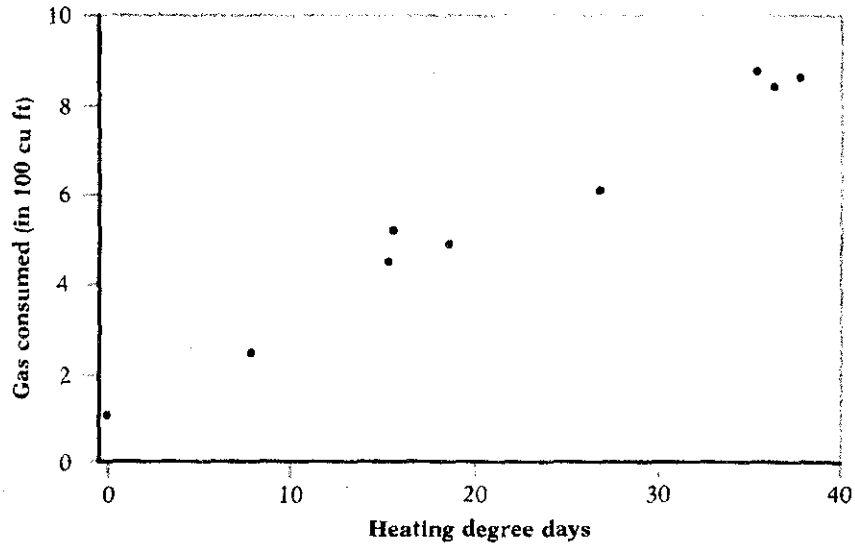


Figure 3.12 Residential natural gas consumption versus heating degree days. See Example 3.6.

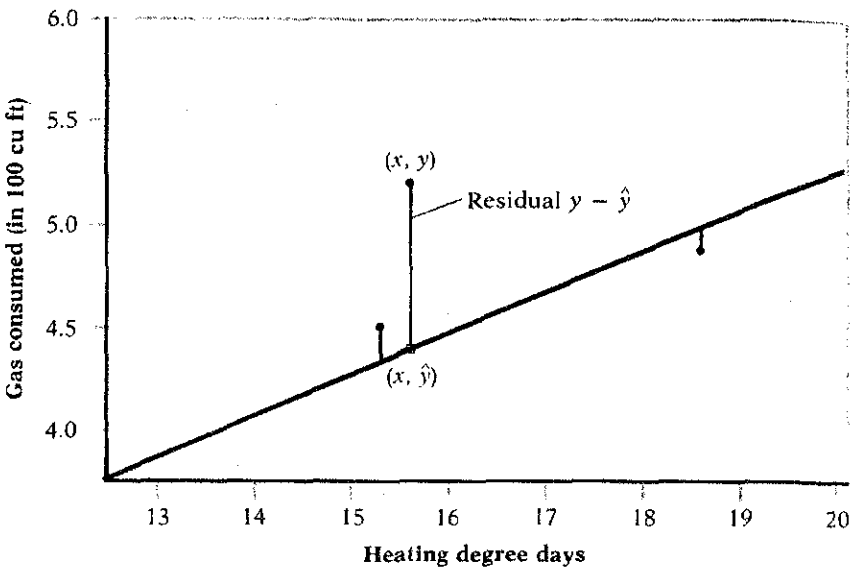


Figure 3.13 The least squares method: fitted line and residuals.

Computing the Regression Line

regression line

Warren's goal is to use his data to predict gas consumption at different temperatures. To do this, he will draw a straight line through the scatterplot in Figure 3.12. A line that shows the dependence of gas consumption on degree days is called a *regression line* of the response variable (consumption) on the explanatory variable (degree days).* Warren could find a satisfactory regression line by moving a transparent straightedge through the points until the fit seems good. We prefer a more objective method. When a scatterplot shows a less strong linear pattern than the one in Figure 3.12, it is difficult to fit by eye a line that is appropriate for predicting y from x. Chapter 2 introduced the most common technique for fitting a line to data, the *method of least squares*. That chapter dealt only with linear growth over time; now we will use least squares regression lines to describe the overall pattern of any linear relationship.

To develop the least squares idea in detail, we magnify the center portion of Figure 3.12 and draw a line through the points. See Figure 3.13. How well does this line fit? Since our goal is to predict y when we are given a

* The term "regression" and the general methods for studying relationships now included under this term were introduced by the English gentleman scientist Francis Galton (1822–1911). Galton was engaged in the study of heredity. One of his observations was that children of tall parents tended to be taller than average but not as tall as their parents. This "regression toward mediocrity" gave these statistical methods their name.

value of x, the error in our prediction is measured in the y, or vertical direction. We therefore concentrate on the deviations of the data points from the line in the vertical direction. These deviations are drawn in Figure 3.13. They are the residuals when this particular line is used to predict y from x.

We represent n observations on two variables x and y as

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

If a line $y = a + bx$ is drawn through the scatterplot of these observations, the line gives the value of y corresponding to x_i as $\hat{y}_i = a + bx_i$. The notation $\hat{y}$ (read "y hat") will be used to distinguish the y-value predicted by the line from the actually observed y. The ith residual is the vertical deviation of the ith data point from the line, which is

$$\begin{aligned} \text{residual} &= \text{observed } y - \text{predicted } y \\ &= y_i - \hat{y}_i \\ &= y_i - a - bx_i \end{aligned}$$

Some of these residuals are positive and some are negative. *The method of least squares chooses the line that makes the sum of the squares of the residuals as small as possible.* Finding this line amounts to finding the values of the intercept a and the slope b that minimize the quantity

$$\sum(y_i - a - bx_i)^2$$

for the given observations x_i and y_i. In the case of the heating data given in Example 3.6, we must choose the a and b that minimize

$$(5.2 - a - 15.6b)^2 + (6.1 - a - 26.8b)^2 + \cdots + (1.1 - a - .0b)^2$$

Here is the solution to this mathematical problem.

Least squares regression line

The least squares regression line of y on x calculated from n observations on these two variables is given by $\hat{y} = a + bx$, where

$$b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} \qquad (3.1)$$

$$a = \bar{y} - b\bar{x}$$

The sums in Equation 3.1 run over all of the observed values x_i or y_i; the subscripts were omitted for easier reading. There are many alternative

expressions for a and b. Many calculators and all statistical software systems will compute a and b for least squares regression lines. Use these resources if they are available to you rather than calculate by hand.

If you do statistical calculations with a basic calculator, you will want to use an alternative *computing formula* for the slope b of the least squares line. Like the similar Equation 1.3 for the variance, this formula avoids subtracting the mean from each observation and makes efficient use of basic sums and sums of squares. The computing formula is

computing formula for *b*

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} \tag{3.2}$$

EXAMPLE 3.7

To calculate the least squares regression line of gas consumption on degree days from the data in Example 3.6, proceed as follows. First compute the building block sums

$$\sum x = 15.6 + \cdots + .0 = 193.9$$
$$\sum x^2 = 15.6^2 + \cdots + .0^2 = 5618.11$$
$$\sum y = 5.2 + \cdots + 1.1 = 50.3$$
$$\sum xy = (15.6)(5.2) + \cdots + (.0)(1.1) = 1375.0$$

Then from Equation 3.2

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$
$$= \frac{(9)(1375.0) - (193.9)(50.3)}{(9)(5618.11) - (193.9)^2}$$
$$= \frac{2621.83}{12{,}965.78} = .202$$
$$a = \bar{y} - b\bar{x}$$
$$= \frac{50.3}{9} - .202\frac{193.9}{9} = 1.23$$

The resulting line is

$$\hat{y} = 1.23 + .202x$$

It is shown in Figure 3.14 superimposed on the original scatterplot. ■

slope

In the regression line $\hat{y} = 1.23 + 0.202x$, the *slope* of the line is $b = 0.202$. The slope is the increase in y corresponding to an increase of one unit in x. Warren estimates that each additional degree day increases his gas consumption by 0.202 hundred cubic feet (or 20.2 cubic feet) per day.

intercept

The number $a = 1.23$ is the *intercept*, the value of $\hat{y}$ when $x = 0$. In this case, $x = 0$ indicates an average temperature of 65° or higher, so there is no demand for heating. Warren expects to use 123 cubic feet of gas per day when $x = 0$. This represents gas used for heating water and for cooking.

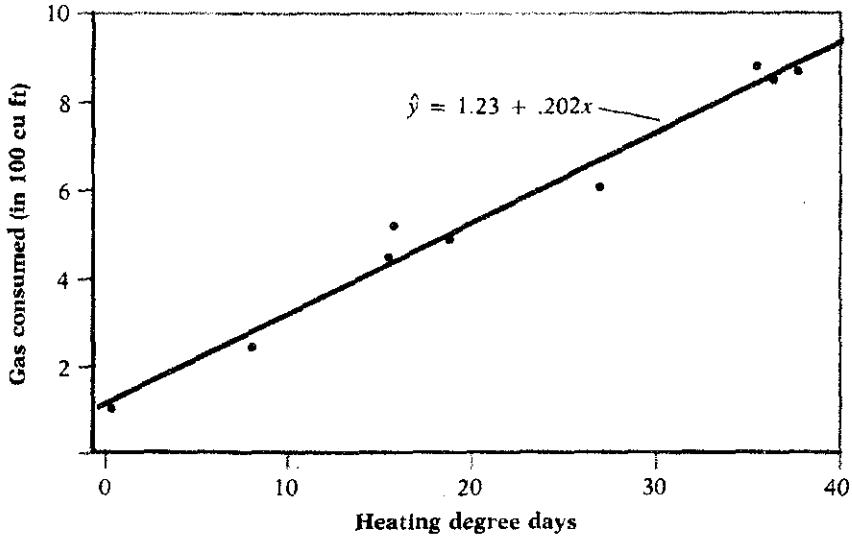


Figure 3.14 The least squares regression line of natural gas consumption on heating degree days.

Warren adds insulation in his attic during the summer, hoping to reduce his gas consumption for heating his home. The next February, gas consumption is 870 cubic feet per day. Was the added insulation effective? We cannot simply compare this year's gas consumption with last February's rate (880 cubic feet per day) because the average temperatures for the two months were not the same. In fact, this February had an average of 40 degree days per day. We therefore *predict* from the regression equation how much gas the house would have used at 40 degree days per day last winter. Our prediction is

$$\hat{y} = 1.23 + (.202)(40) = 9.31$$

or 931 cubic feet. Warren estimates that he saved about 61 cubic feet per day by adding insulation.

The accuracy of predictions from a regression line depends on how much scatter about the line the data show. When, as in this example, the data points are all very close to the line, we are confident that our prediction is reliable. If, however, the data show a linear pattern with considerable spread (as in Figure 3.1), we may agree that a regression line summarizes the pattern but we will also put less confidence in a prediction based on the line. We will learn in Chapter 10 how to give a numerical statement of our level of confidence in predictions. Of course, the usefulness of any

regression line is limited by the data from which it was computed. Warren must collect new data and compute a new regression line if he wishes to predict his gas consumption with the new insulation in place.

Fit and Residuals

The dependence of gas consumption on degree days (Examples 3.6 and 3.7) was so strong that we can be confident that the least squares regression line tells almost the whole story about these data. This happy circumstance is not always the case. The deviations from the fitted regression line often give important additional information. These deviations are best studied by calculating and plotting the n residuals $y_i - \hat{y}_i$. As we saw in Chapter 2, the residuals from a least squares regression line have the property that their sum (and therefore their mean) is always 0.

EXAMPLE 3.8

In Example 3.6, the observed consumption of gas for October, when $x = 15.6$ degree days, was $y = 5.2$. The predicted consumption for this x from the least squares regression line of Example 3.7 is

$$\hat{y} = 1.23 + (.202)(15.6) = 4.38$$

The October residual is therefore

$$e_t = 5.2 - 4.38 = .82$$

This residual appears in Figure 3.13 as the length of the vertical line joining y and $\hat{y}$. ■

The distribution of the residuals can be examined using the tools presented in Chapter 1. A stemplot helps to check the symmetry of the distribution and to spot outliers. A normal probability plot can tell whether the residuals are approximately normally distributed. Since formal statistical methods for predicting y from x using a regression line depend on the shape of the distribution of residuals, questions of symmetry and normality will eventually interest us. Our present concerns, however, are descriptive. Is the relationship between y and x close to linear? Are there outlying observations that require special attention? Is there evidence that y is influenced by variables other than x? Scatterplots of the residuals against other variables are used to investigate questions like these.

You should always plot the residuals against the corresponding values of the explanatory variable x. If all is well, the pattern of this plot will be an unstructured horizontal band centered at 0 (the mean of the residuals) and symmetric about 0. Figure 3.15(a) displays an idealized version of this pattern. A curved pattern like the one in Figure 3.15(b) shows a nonlinear dependence of the response y on x. A fan-shaped pattern like the one in Figure 3.15(c) shows that the variation of y about the line increases as x increases; predictions of y will therefore be more precise for smaller values of x, where y shows less spread about the line.

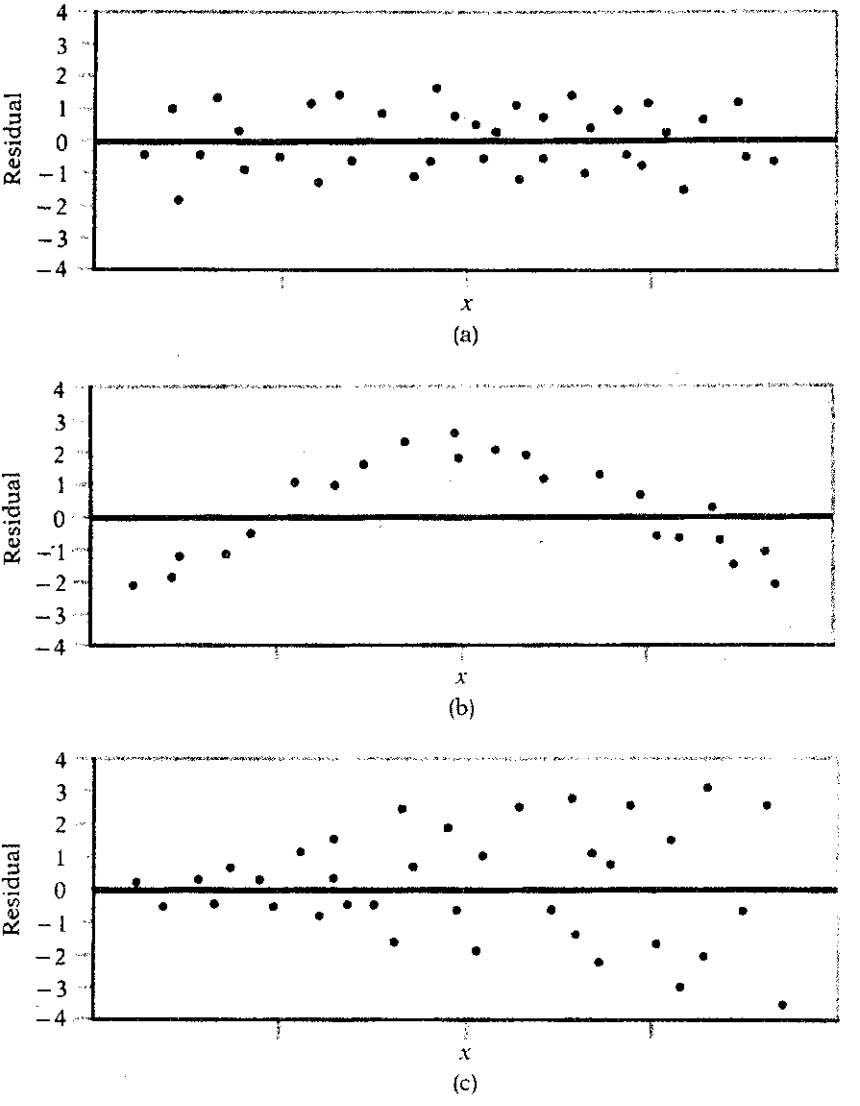


Figure 3.15 Idealized patterns in plots of least squares residuals.

Other residual plots can give other kinds of information. A plot of the residuals against the time order in which the data were collected may reveal a dependence of y on some variable that changes over time. Since we have seen that background variables can strongly affect the appearance of a scatterplot, residuals should be plotted against time and against any other variables. In all such plots, the residuals are plotted vertically and time or

any other potential explanatory variable horizontally. A sloping linear band or other strong deviation from the pattern shown in Figure 3.15(a) indicates a dependence of y on the new variable. A sloping linear band cannot appear when the residuals are plotted against x because the residuals have no remaining linear dependence on x. Of course, in practice, patterns will not be as neat as the schematic drawings in Figure 3.15. Any set of data will show some irregularity; don't overreact to vague patterns in a residual plot. Here are two examples of plotting and interpreting residuals.

EXAMPLE 3.9

The mathematics department of a large state university must plan in advance the number of sections and instructors required for elementary courses. The department hopes that the number of students in these courses can be predicted from the number of entering freshmen, which is known before the new students actually choose courses. The table below contains the data for recent years.[6] The explanatory variable x is the number of students in the freshman class, and the response variable y is the number of students who enroll in mathematics courses at the 100 level.

Year	1980	1981	1982	1983	1984	1985	1986	1987
x	4595	4827	4427	4258	3995	4330	4265	4351
y	7364	7547	7099	6894	6572	7156	7232	7450

Equation 3.2 could be used to compute the regression line from these data. Instead, we enter the data into the MINITAB statistical computing system, calling x FRESH and y MATH for easy recall. The regression command and the first part of the printout appear as follows:

```
MTB > REGRESS 'MATH' ON 1, 'FRESH'

The regression equation is

MATH = 2492.69 + 1.0663 FRESH
```

Other information, including a table of the residuals, follows on the printout. We see from the output that the least squares regression line is

$$\hat{y} = 2492.69 + 1.0663x$$

■

A scatterplot (Figure 3.16) of the data from Example 3.9 with the regression line shows a reasonably linear fit. There is a cluster of points with similar values near the center. A plot of the residuals against x (Figure 3.17) magnifies the vertical deviations of the points from the line. It is apparent from either graph that a slightly different line would fit the five lower points very well, so that the three points above the line represent a somewhat different relation between the number of freshmen x and mathematics enrollments y.

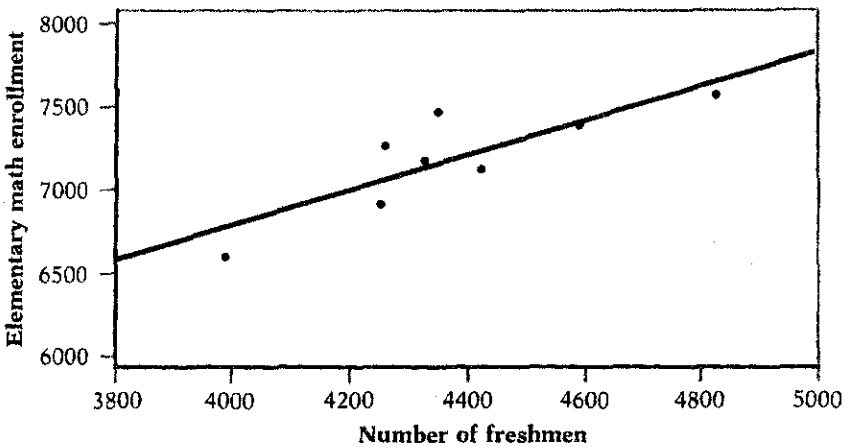


Figure 3.16 The regression of elementary mathematics enrollment on number of freshmen at a large university. See Example 3.9.

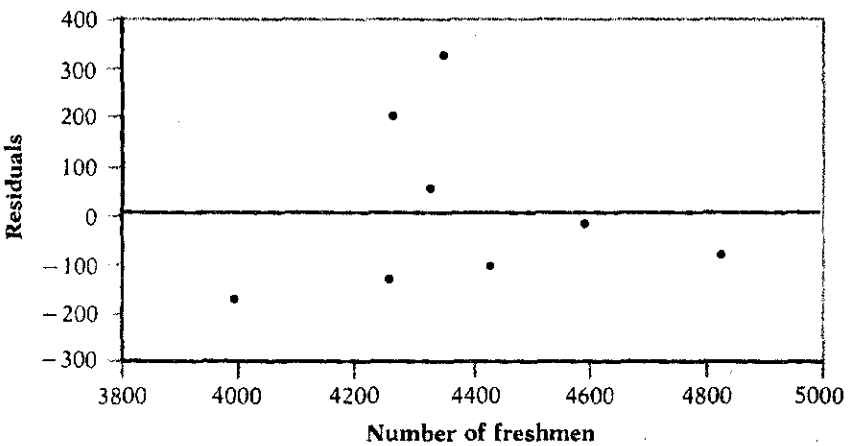


Figure 3.17 Residual plot for the regression of mathematics enrollment on number of freshmen.

A second plot of the residuals clarifies the situation. Figure 3.18 is a plot of the residuals against time. We now see that the five negative residuals are for the years 1980 to 1984, and the three positive residuals represent the years 1985 to 1987. This pattern suggests that a change took place between 1984 and 1985 causing a higher proportion of students to take mathematics courses beginning in 1985. In fact, one of the schools in the university

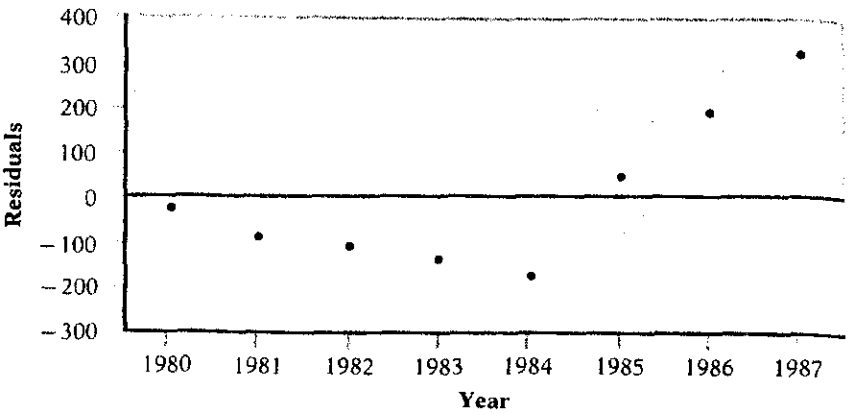


Figure 3.18 Plot of the residuals versus time for the regression of mathematics enrollment on number of freshmen.

changed its program to require that entering students take another mathematics course. Notice that Figure 3.18 does *not* say that mathematics enrollments changed linearly over time up to 1984—a glance at the table in Example 3.9 shows that they did not. Figure 3.18 is a plot of the residuals from the regression of the enrollments on the count of freshmen, not a plot of the enrollments themselves.

The contrast in Figure 3.18 between the years up to 1984 and the following years is another reminder that an observed relationship may not be trustworthy when changes occur in the underlying situation. Because the least squares regression line from Example 3.9 makes use of data from 1984 and earlier, the mathematics department should not use it in future years to predict elementary mathematics enrollments from the count of freshmen.

EXAMPLE 3.10

A study of cognitive development in young children recorded the age (in months) at which each of 21 children spoke their first word and their Gesell score, the result of an aptitude test taken much later. The data appear in Table 3.3 and in the scatterplot of Figure 3.19.[7] Notice that cases 3 and 13, and cases 16 and 21, have identical values for both variables. When drawing the scatterplot, we used a different symbol to show that one point stands for two cases.

The purpose of the study was to ascertain if age at first word (x) could predict the later test score y. It is also plausible, however, to try to guess the age when a child first spoke from the test score. Both the regression line of y on x and the regression line of x on y are therefore displayed in Figure 3.19. *These regression lines are very different.* One minimizes vertical deviations and the other minimizes horizontal deviations. The lines intersect at the point $(\bar{x}, \bar{y})$, through which any least squares regression line always passes. It is essential in a regression setting to clearly distinguish the explanatory from the response variable. This example also indicates why it is hard to fit an appropriate line by eye—we cannot easily concentrate on lack of fit in one direction only. ■

Table 3.3 Age (in months) at first word and Gesell adaptive score

Case	Age	Score	Case	Age	Score
1	15	95	12	9	96
2	26	71	13	10	83
3	10	83	14	11	84
4	9	91	15	11	102
5	15	102	16	10	100
6	20	87	17	12	105
7	18	93	18	42	57
8	11	100	19	17	121
9	8	104	20	11	86
10	20	94	21	10	100
11	7	113			

Figure 3.20 shows the regression of Gesell score on age at first word (the colored line) and highlights two cases, the children numbered 18 and 19 in Table 3.3. Case 19 is an outlier that lies well away from the general straight-line pattern of the other points. Case 18 is not an outlier if we take the overall pattern to be roughly linear, for this point falls in that pattern. The residual plot (Figure 3.21) reinforces these conclusions. Case 19 produces

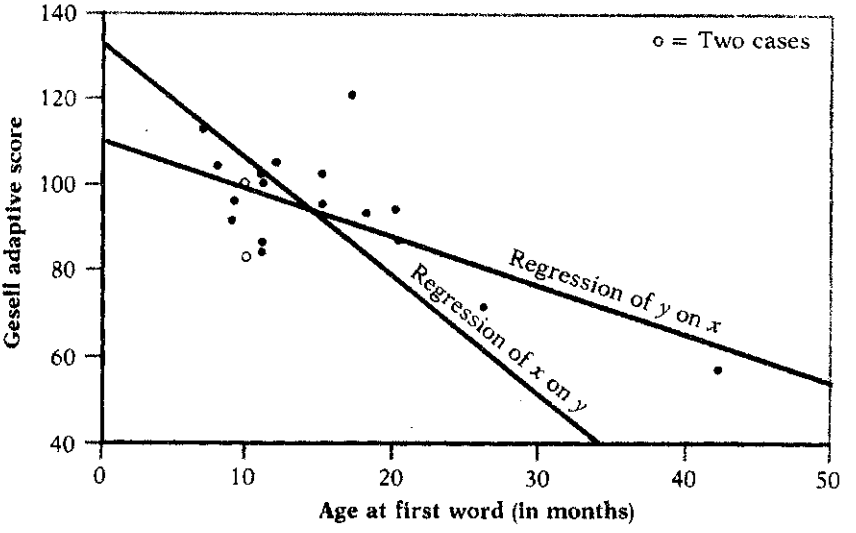


Figure 3.19 The two least squares regression lines of Gesell score on age at first word and of age at first word on Gesell score. See Example 3.10.

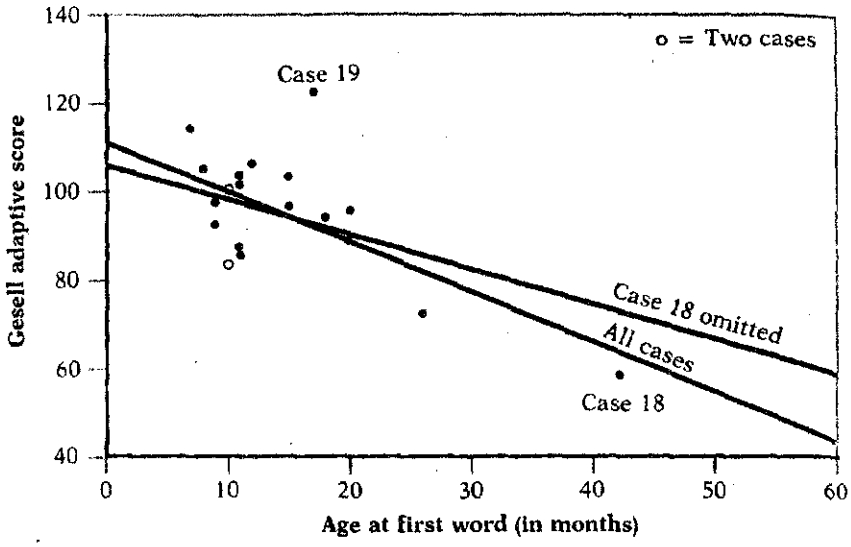


Figure 3.20 The regression of Gesell score on age at first word, with and without an influential case.

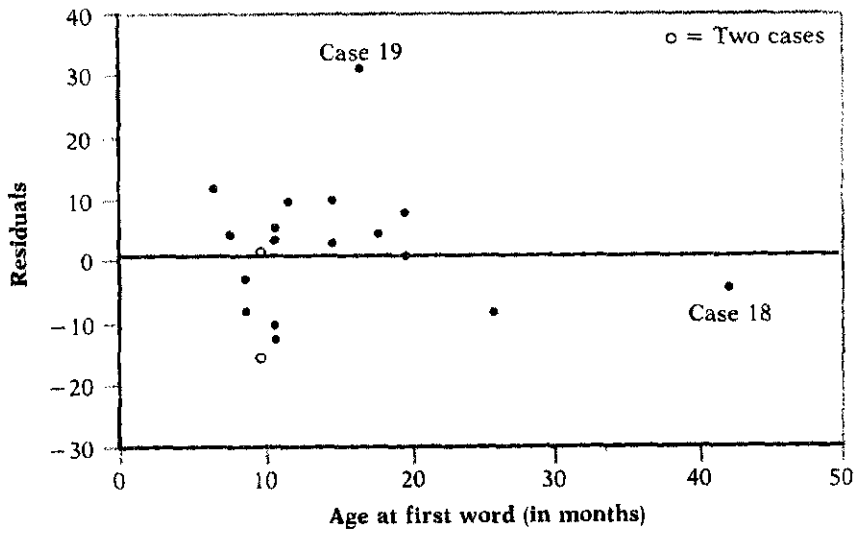


Figure 3.21 Residual plot for the regression of Gesell score on age at first word: case 19 is an outlier and case 18 is an influential case that does not have a large residual.

the largest residual and case 18 a rather small residual. Yet case 18 is very important. Because of its extreme position on the age scale, it has a strong influence on the position of the regression line. The black line in Figure 3.20 is the least squares regression line computed from all 20 cases except case 18. Notice how this single case pulls the colored regression line toward itself. Case 19, on the other hand, has less influence on the position of the regression line because this child's age at first word (17 months) lies close to the mean age $\bar{x} = 14.4$ months. The regression line must pass through the point $(\bar{x}, \bar{y})$, so it is difficult for an observation at an age close to $\bar{x}$ to move the line to any great degree.

Outliers and Influential Observations

Example 3.9 illustrates the power of residuals to detect deviations from linear dependence of y on x and even to suggest explanations for departures from linearity. Example 3.10, however, points out a weakness of regression residuals. Residuals call attention to outliers but not to other influential cases.

Outlier, influential observation

> An outlier in a regression is a data point that produces a large residual. An observation is influential if removing it would markedly change the position of the regression line.

In other words, an outlier lies outside the linear pattern, and so has a large residual. But you must also be on the alert for influential observations that are not outliers. Least squares regression is not resistant. The position of the least squares regression line is heavily influenced by observations that are extreme in x. The influence of these points often guarantees that they are not outliers, however, because they draw the regression line toward themselves. Influential observations are easy to detect (as long as there is only a single explanatory variable), but merely looking for large residuals doesn't do the job. Many statistical computing systems produce tables and plots of regression residuals, but if you rely on these alone, you may miss the most influential cases. The basic scatterplot of y versus x will alert you to observations that are extreme in x and may therefore be influential.

An influential observation should be investigated to ensure that it is correct. Even if no error is found, you should ask whether this observation belongs to the population you are studying. A statistician would ask the child development researcher in Example 3.10 whether the child of case 18 is so slow to speak that this case should not be allowed to influence the analysis. If this case is excluded, much of the evidence for a connection between the age at which a child begins to talk and later aptitude scores vanishes. If the case is retained, we need data on other children who were slow to begin talking, so that the analysis is no longer so heavily dependent on a single child.

In addition to showing outliers and influential observations, examination of residuals may also disclose the influence of lurking variables.

Lurking variable

> A lurking variable is a variable that has an important effect on the response but that is not included among the explanatory variables studied.

Since lurking variables are often unrecognized and unmeasured, detecting their effect is a challenge. The most effective tool is an understanding of the background of the data that allows you to guess what lurking variables might be present. One useful method for detecting lurking variables is to plot both the response variable and the regression residuals against the time order of the observations, since many variables change with time. This was the case in Figure 3.18, where the change in course requirements over time was a lurking variable. Here is another example.

EXAMPLE 3.11

A study of the manufacture of molded plastic parts examined the effect of time spent in the mold (x) on the strength (y) of the part. Several batches of hot plastic were pressed for 10 seconds, then several more batches for 20 seconds, and so on. Regression showed a strong dependence of strength on molding time. A plot of strength against the order in which the batches were molded, however, showed an even stronger dependence. This led the engineers to realize that the mold grew constantly warmer as more batches were processed, and that this lurking variable explained the changing strength.[8]

This experiment was poorly designed; the effect of time in the mold cannot be separated from the effect of any lurking variable that was changing over time because all 10-second batches were molded first, then all 20-second batches, and so on. As we shall see in Chapter 4, properly designed experiments can prevent the effects of lurking variables from being confused with the effects of the explanatory variable or variables. ■

A regression line is a compact description of the overall pattern of dependence of y on x. Since the straight line is fitted to the data, it can be described briefly as a FIT. The definition of residuals then says that

$$\text{DATA} = \text{FIT} + \text{RESIDUAL}$$

There are many ways of obtaining an overall FIT for two-variable data. The median trace produces a FIT that is not forced to be a straight line. Some statistical software systems provide alternative methods of fitting a straight line to data that are more resistant than least squares regression. Examining residuals, especially through plots, can help assess the success of a FIT produced by any method. Residuals from least squares regression have several special properties (such as mean 0) that make them simpler to work with.

Section 3.2 Exercises

If your software allows resistant fitting of a line to data, you can check the results of the least squares fit by also doing a resistant fit. If the two lines do not agree closely, search for influential observations or other explanations for the lack of agreement. Despite its lack of resistance, the least squares method continues to dominate statistical practice. To use this—and other—statistical methods wisely, you must never neglect the detailed examination of your data.

SUMMARY

When a scatterplot suggests that the dependence of y on x can be summarized by a straight line, the **least squares regression line** of y on x can be calculated. This fitted line can be used to predict y for a given value of x.

The fit of a regression line is examined by plotting the **residuals**, or differences between the observed and predicted values of y. Be on the lookout for **outliers**, which are points with unusually large residuals, and also for nonlinear patterns and uneven variation about the line.

Influential observations, individual points that substantially change the regression line, must also be spotted and examined. Influential observations are often outliers in the x variable, but they may not have large residuals.

Evidence of the effects of **lurking variables** on y may be provided by plots of y and the residuals against the time order of the observations.

SECTION 3.2 EXERCISES

3.17 Exercise 2.18 presented the following data on Sarah's growth between the ages of 36 months and 60 months:

Age (months)	36	48	51	54	57	60
Height (cm)	86	90	91	93	94	95

(a) Use Equation 3.2 to compute the slope b of the least squares regression line of height on age; then find the intercept a from $a = \bar{y} - b\bar{x}$. According to the regression line, how much does Sarah grow each month?

(b) Make a scatterplot of the data and draw the least squares line on the plot.

(c) Use your regression line to estimate Sarah's height at 42 months of age.

3.18 A student who waits on tables at a Chinese restaurant in a college neighborhood records the cost of meals and the tip left by single diners. Here are some of the data.

Meal	\$3.50	\$4.79	\$5.24	\$3.62	\$5.35
Tip	\$ .25	\$ .50	\$ .60	\$ .40	\$ .75

(a) Compute the least squares regression for these data starting with Equation 3.2.
(b) Make a scatterplot of the data and draw the regression line on your plot.
(c) The next diner orders a meal costing \$3.89. Use your regression line to predict the tip.

3.19 Use your regression line from Exercise 3.17 to predict Sarah's height at each of 36, 48, 51, 54, 57, and 60 months. Then compute the residuals for the regression line by subtracting these predicted heights from the actually observed heights at these ages. Verify that the residuals have sum 0 (up to roundoff error). Make a plot of the residuals against age and briefly describe the pattern of residuals.

3.20 Compute the five residuals for the data in Exercise 3.18, using the regression line that you have already computed. Verify that the residuals have sum 0 (up to roundoff error). Plot the residuals against the cost of the meals and comment on the pattern of the residuals.

3.21 Runners are concerned about their form when racing. One measure of form is the stride rate, defined as the number of steps taken per second. A runner is inefficient when the rate is either too high or too low. Of course, as the speed increases, the stride rate should also increase. In a study of 21 of the best American female runners, the stride rate was measured for different speeds. The following table gives the speeds (in feet per second) and the average stride rates for these women. (Data from R. C. Nelson, C. M. Brooks, and N. L. Pike, "Biomechanical comparison of male and female distance runners," in P. Milvy (ed.), *The Marathon: Physiological, Medical, Epidemiological, and Psychological Studies*, New York Academy of Sciences, 1977, pp. 793–807.)

Speed	15.86	16.88	17.50	18.62	19.97	21.06	22.11
Stride rate	3.05	3.12	3.17	3.25	3.36	3.46	3.55

(a) Plot the data with speed on the x axis and stride rate on the y axis. Does a straight line adequately fit these data?
(b) Compute the slope and intercept for the least squares line, using

the computing formula 3.2 or a statistical calculator or software. Graph the least squares line on your plot from (a).

(c) For each of the speeds given, compute the predicted value using the least squares line.

(d) Using the results of (c), compute the residuals. Verify that the residuals add to zero.

(e) Plot the residuals versus speed. Describe the pattern. What does the plot indicate about the adequacy of the linear fit?

3.22 Refer to the previous exercise. The corresponding data for a group of 24 elite American male runners are given in the following table:

Speed	15.86	16.88	17.50	18.62	19.97	21.06	22.11
Stride rate	2.92	2.98	3.03	3.11	3.22	3.31	3.41

Explain why the stride rate at each speed is lower for men than for women. Then answer the questions given in the previous exercise using the data for males.

3.23 Research on digestion requires accurate measurements of blood flow through the lining of the stomach. A promising way to make such measurements easily is to inject mildly radioactive microscopic spheres into the bloodstream. The spheres lodge in tiny blood vessels at a rate that is in proportion to blood flow; their radioactivity allows blood flow to be measured from outside the body. Medical researchers compared blood flow in the stomachs of dogs, measured by use of microspheres, with simultaneous measurements taken using a catheter inserted into a vein. The data, in milliliters of blood per minute (ml/minute), appear below. (Based on L. H. Archibald, F. G. Moody, and M. Simons, "Measurement of gastric blood flow with radioactive microspheres," *Journal of Applied Physiology*, 38 (1975), pp. 1051–1056.)

Spheres	4.0	4.7	6.3	8.2	12.0	15.9	17.4	18.1	20.2	23.9
Vein	3.3	8.3	4.5	9.3	10.7	16.4	15.4	17.6	21.0	21.7

(a) Make a scatterplot of these data, with the microsphere measurement as the explanatory variable. There is a strongly linear pattern.

(b) Calculate the least squares regression line of venous flow on microsphere flow. Draw your regression line on the scatterplot.

(c) Predict the venous measurement for microsphere measurements of 5, 10, 15, and 20 ml/minute. If the microsphere measurements are within about 10% to 15% of the predicted venous measurements, the researchers will simply use the microsphere

measurements in future studies. Is this condition satisfied over this range of blood flow?

3.24 Suppose that the last customer in Exercise 3.18 had ordered a large meal and left no tip. The data are now as follows:

Meal	\$3.50	\$4.70	\$5.24	\$3.62	\$15.69
Tip	\$ .25	\$ .50	\$ .60	\$ .40	0

(a) Make a scatterplot of these data. Which observation will be most influential? Why?

(b) Fit a line by eye to the first four observations and draw this line on your plot. The least squares regression line fitted to all five observations is $\hat{y} = 0.57 - 0.034x$. Draw this line on your graph as well. A comparison of the two lines shows the influence of the final observation.

(c) Does it appear from the graph that the influential observation has a larger residual from the least squares line than the other observations? (You need not actually compute the residuals.)

3.25 One component of air pollution is airborne particulate matter such as dust and smoke. Particulate pollution is measured by using a vacuum motor to draw air through a filter for 24 hours. The filter is weighed at the beginning and end of the period, and the weight gained is a measure of the concentration of particles in the air. In a study of pollution, measurements were taken every 6 days with identical instruments in the center of a small city and at a rural location 10 miles southwest of the city. Because the prevailing winds blow from the west, it was suspected that the rural readings would be generally lower than the city readings, but that the city readings could be predicted from the rural readings. The following table gives readings taken every 6 days between May 2 and November 26, 1986. The entry NA means that the reading for that date is not available, usually because of equipment failure. (Data provided by Matthew Moore.)

Rural	NA	67	42	33	46	NA	43	54	NA	NA	NA	NA
City	39	68	42	34	48	82	45	NA	NA	60	57	NA
Rural	38	88	108	57	70	42	43	39	NA	52	48	56
City	39	NA	123	59	71	41	42	38	NA	57	50	58
Rural	44	51	21	74	48	84	51	43	45	41	47	35
City	45	69	23	72	49	86	51	42	46	NA	44	42

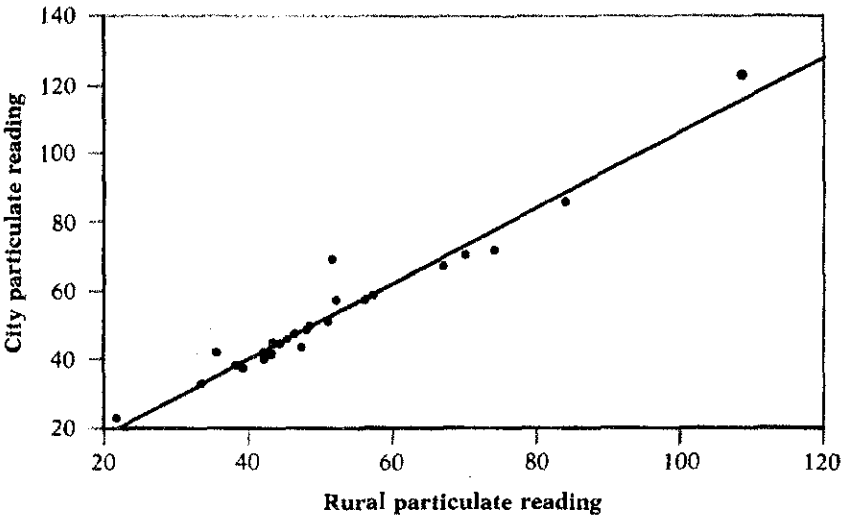


Figure 3.22 The regression of city particulate concentration on rural particulate concentration for the same days (Exercise 3.25).

To assess the success of predicting the city particulate reading from the rural reading, the 26 complete cases (with both readings present) are analyzed. Computer work finds the least squares regression line of the city reading y on the rural reading x to be

$$\hat{y} = -2.580 + 1.0935x$$

Figure 3.22 is a scatterplot for the 26 complete cases with the regression line drawn.

(a) Which observation in Figure 3.22 appears to be the most influential? Is this the observation with the largest residual (vertical distance from the line)?

(b) Locate in the table the observation you chose from the graph in (a) and compute its residual.

(c) Do the data suggest that using the least squares line for prediction will give approximately correct results over the range of values appearing in the data? (The incompleteness of the data does not seriously weaken this conclusion if equipment failures are independent of the variables being studied.)

(d) On the fourteenth date in the series, the rural reading was 88 and the city reading was not available. What do you estimate the city reading to be for that date?

3.26 To study the energy savings resulting from adding solar heating panels to a house, researchers measured the natural gas consumption of the house for more than a year, then installed solar panels and observed

the natural gas consumption for almost 2 years. The variables are as in Example 3.6: The explanatory variable x is degree days per day during the several weeks covered by each observation, and the response variable y is gas consumption (in hundreds of cubic feet) per day during the same period. Figure 3.23 plots y against x, with separate symbols for observations taken before and after the installation of the solar panels. The least squares regression lines were computed separately for the before and after data, and are drawn on the plot. The regression lines are

$$\text{Before:} \quad \hat{y} = 1.089 + .189x$$
$$\text{After:} \quad \hat{y} = 0.853 + .157x$$

(Data provided by Professor Robert Dale, Purdue University, West Lafayette, Ind.)

(a) Does the scatterplot suggest that a linear model is appropriate for the relationship between degree days and natural gas consumption? Do any individual observations appear to be outliers (large residuals) or highly influential?

(b) About how much additional natural gas was consumed per day for each additional degree day before the panels were added? After the panels were added?

(c) The average daily temperature during January in this location is about 30°, which corresponds to 35 degree days per day. Use the

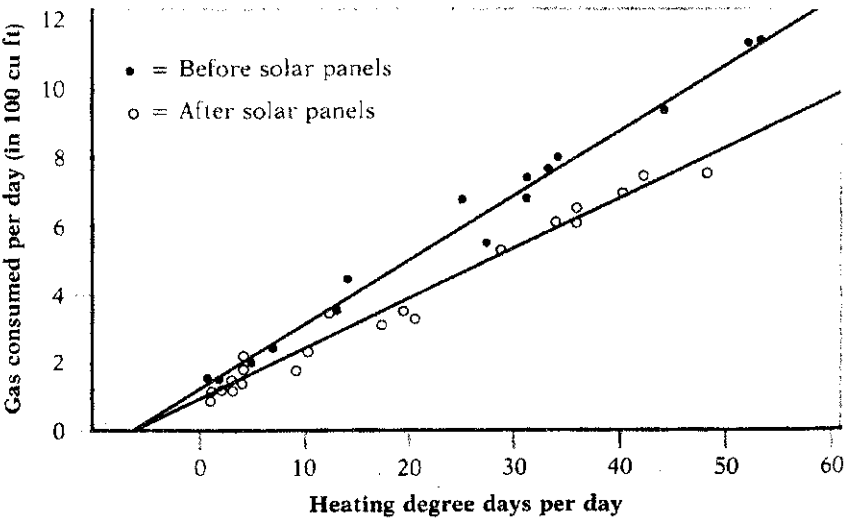


Figure 3.23 The regression of residential natural gas consumption on heating degree days before and after installation of solar heating panels (Exercise 3.26).

regression lines to predict daily gas usage for a day with 35 degree days before and after installation of the panels. If the price of natural gas is \$0.60 per hundred cubic feet, how much money do the solar panels save in the 31 days of a typical January?

3.27 Table 1.4 gives the calories and sodium content for each of 17 brands of meat hot dogs. The distribution of calories was examined in the discussion following Example 1.11, and the distribution of sodium in Exercise 1.58. Now we examine the relation between calories and sodium. Figure 3.24 is a scatterplot of the data from Table 1.4.

(a) Describe the main features of the relationship. (The discussion following Example 1.11 may help.)

(b) The plot shows two least squares regression lines. One was calculated using all of the observations, while the other omitted the brand of veal hot dogs that is an outlier in both variables measured. Which line (colored or black) was calculated from all of the data? Explain your answer.

(c) The regression line that ignores the outlier is

$$\hat{y} = 46.90 + 2.401x$$

A new brand of meat hot dog (not made with veal) has 150 calories per frank. How many milligrams of sodium do you estimate that one of these hot dogs contains?

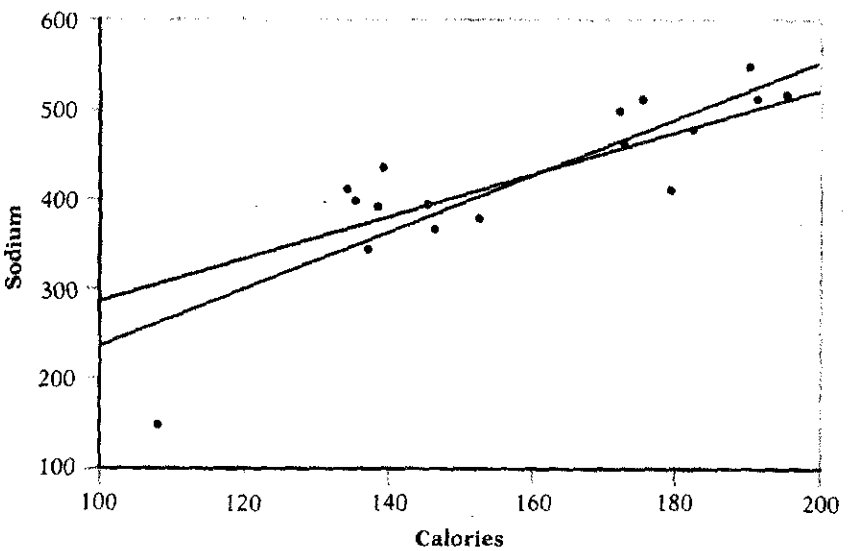


Figure 3.24 The regression of sodium content on calories for brands of hot dogs, calculated with and without an outlier (Exercise 3.27).

3.28 Are baseball players paid according to their performance? To study this question, a statistician analyzed the salaries of over 260 major league hitters along with such explanatory variables as career batting average, career home runs per time at bat, and years in the major leagues. This is a *multiple regression* with several explanatory variables. More detail on multiple regression appears in Chapter 10, but the fit of the model is assessed just as we have done in this chapter, by calculating and plotting the residuals

$$\text{residual} = \text{observed } y - \text{predicted } y$$

(This analysis was done by Crystal Richard.)

(a) Figure 3.25(a) is a plot of the residuals versus the predicted salary. This plot was produced by the SAS statistical software system that was used to analyze the data. Notice that when points are too close together to plot separately, SAS uses letters of the alphabet to show how many points there are at each position. Describe the pattern that appears on this residual plot. Will the regression model predict high or low salaries more precisely?

(b) After studying the residuals in more detail, the statistician decided to predict the logarithm of salary rather than the salary itself. One reason was that while salaries are not normally distributed (the distribution is skewed to the right), their logarithms are nearly normal. When the response variable is the logarithm of salary, a plot of the residuals against the predicted value is satisfactory—it looks like Figure 3.15(a). Figure 3.25(b) is a plot of the residuals against the number of years the player has been in the major leagues. Describe the pattern that you see. Will the model overestimate or underestimate the salaries of players who are new to the majors? Of players who have been in the major leagues about 8 years? Of players with more than 15 years in the majors?

3.29 Refer to the golf data given in Exercise 3.4.

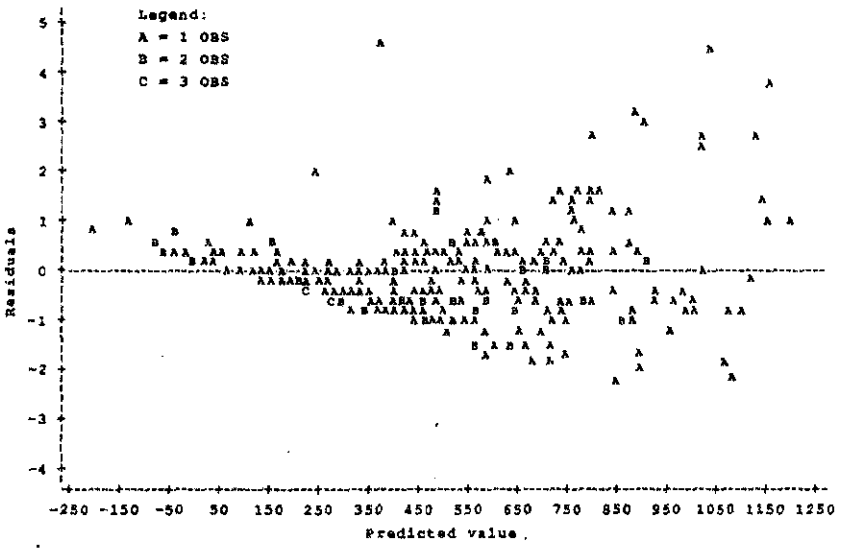
(a) Plot the data with the round 1 scores on the x axis and the round 2 scores on the y axis. Does it appear reasonable to fit a straight line to these data? Are there any apparent outliers or influential observations?

(b) Compute the slope and intercept for the least squares line. Graph the least squares line on your plot from (a).

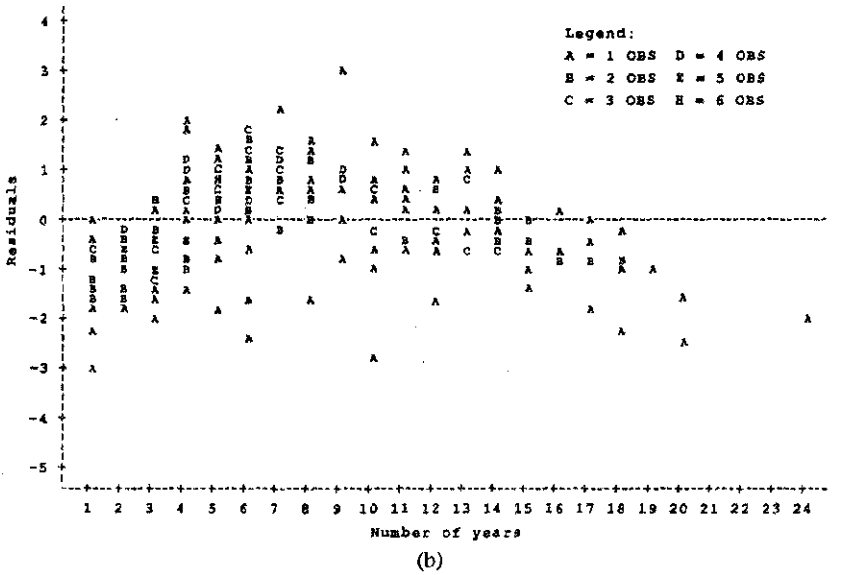
(c) For each of the round 1 scores, compute the predicted value for the round 2 score using the least squares line.

(d) Using the result of (c), compute the residuals. Verify that the residuals add to zero.

(e) Plot the residuals versus the round 1 scores. Describe the pattern. What does the plot indicate about the adequacy of the linear fit?



(a)



(b)

Figure 3.25 Two residual plots for the regression of baseball players' salaries on their performance (Exercise 3.28).

3.30 Refer to the erosion data given in Exercise 3.5.

(a) Compute the slope and intercept for the least squares line. Graph the least squares line on your plot from (a) of that exercise.

(b) For each of the flow rates, compute the predicted value using the least squares line.

(c) Using the results of (b), compute the residuals. Verify that the residuals add to zero.

(d) Plot the residuals versus flow rate. Describe the pattern. What does the plot indicate about the adequacy of the linear fit?

3.31 Refer to the data in Exercise 3.16 giving two measurements of the strength of wood strips. As before, you should use a statistical computing system to analyze this larger data set.

(a) Plot the data with T1 on the x axis and T2 on the y axis. Does it appear reasonable to fit a straight line to these data?

(b) Compute the least squares regression line. Graph the least squares line on your plot from (a).

(c) For each case compute the predicted value using the least squares line and find the residuals.

(d) Plot the residuals versus T1. Describe the pattern. What does the plot indicate about the adequacy of the linear fit?

(e) Plot the residuals versus S, which gives the time order of the measurements taken. Describe the plot.

(f) Make a stemplot (or, if your system permits, a normal quantile plot) of the residuals. Describe the distribution of the residuals. Is it nearly symmetric? Approximately normal?

3.32 Refer to the previous exercise.

(a) Plot the data with T2 on the x axis and T1 on the y axis.

(b) Compute the slope and intercept for the least squares line which views T1 as the response variable and T2 as the explanatory variable. Graph the least squares line on your plot from (a).

(c) On a plot of the data with T1 on the x axis and T2 on the y axis, graph the least squares line from the previous exercise and also the one that you just found in part (b) of this exercise. (Hint: For each line simply find two points on it, plot the points and connect them with a line.)

(d) Find the mean of each variable and verify the fact that the two lines intersect at the point corresponding to these means.

3.33 If a statistical computing system is available, a more thorough analysis of the particulate pollution data in Exercise 3.25 is possible. Enter the data into the computer.

(a) Do a regression of the city readings y on the rural readings x; verify the regression equation given in Exercise 3.25 and obtain the residuals.

(b) Plot the residuals against both x and the time order of the observations, and comment on the results.

(c) Make a stemplot (or, if your system allows, a normal quantile plot) of the residuals. Is the distribution of the residuals nearly symmetric? Does it appear to be approximately normal?

3.34 The following table gives the results of a study of a sensitive chemical technique called gas chromatography which is used to detect very small amounts of a substance. Five measurements were taken for each of four amounts of the substance being investigated. The explanatory variable x is the amount of substance in the specimen, measured in nanograms (ng), or units of 10^{-9} gram. The response variable y is the output reading from the gas chromatograph. The purpose of the study is to calibrate the apparatus by relating y to x. (Data from D. A. Kurtz (ed.), *Trace Residue Analysis*, American Chemical Society Symposium Series No. 284, 1985, Appendix.)

Amount (ng)	Response				
.25	6.55	7.98	6.54	6.37	7.96
1.00	29.7	30.0	30.1	29.5	29.1
5.00	211	204	212	213	205
20.00	929	905	922	928	919

(a) Make a scatterplot of these data. The relationship appears to be approximately linear, but the wide variation in the response values makes it hard to see detail in this graph.

(b) Compute the least squares regression line of y on x, and plot this line on your graph.

(c) Now compute the residuals and make a plot of the residuals against x. It is much easier to see deviations from linearity in the residual plot. Describe carefully the pattern displayed by the residuals.

3.3 CORRELATION

We have to this point concentrated on analyzing data having a clear explanatory–response structure. In order to fit a regression line, we must know which is the explanatory variable and which is the response. What tools are available when we are interested in the relation or association between two variables but do not wish to claim that one explains the other?

The basic scatterplot, of course, continues to portray the direction, form, and strength of any relationship between two quantitative variables. But the

interpretation of a scatterplot by eye is surprisingly subjective. Changing the horizontal or vertical scale, for example, greatly affects our perception of the strength of a linear or other pattern. Even the amount of white space around the point cloud in a scatterplot can fool us. As Figure 3.26 illustrates, a scatterplot appears to show a stronger relationship when the point cloud is reduced in size relative to its surroundings. The two scatterplots in this figure

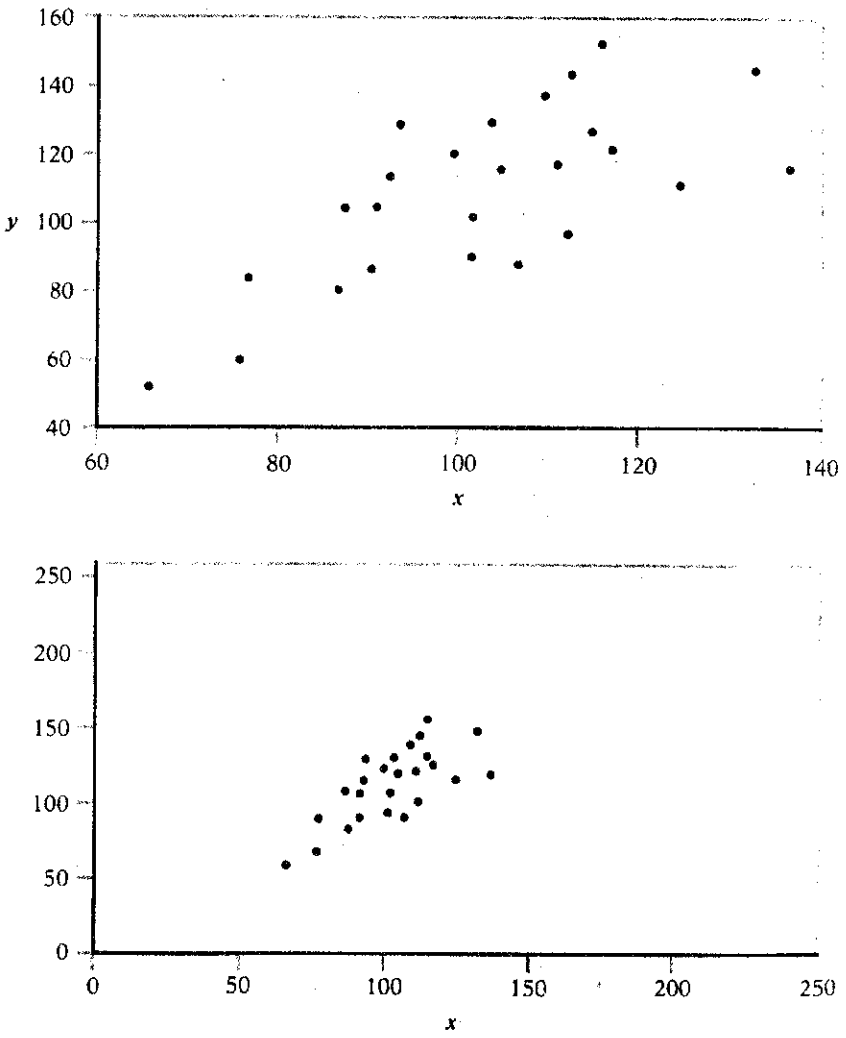


Figure 3.26 Two identical scatterplots; the linear pattern in the lower plot appears stronger because of the surrounding white space.

are identical in every respect except that the lower plot is drawn smaller in a large field and therefore appears to show a stronger linear pattern.[9]

When we are concerned with linear patterns in a scatterplot, we can use an important numerical measure to aid our visual perception. The *correlation coefficient* measures the strength of the linear association between two quantitative variables. It does not distinguish explanatory from response variables, and it is not affected by changes in the unit of measurement of either or both variables.* The word "correlation" is often used as a vague synonym for "association." Because correlation is a specific numerical measure that applies only to linear association and only to quantitative variables, we will use the word only in this sense. There is a positive association between educational level and income in Example 3.5, for example, but correlation is not meaningful because educational level is a categorical variable.

Computing the Correlation

We again have n observations on two variables x and y, denoted by

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

Unlike the regression setting, x and y are not necessarily explanatory and response variables, although they may be. Here is the definition of the correlation coefficient.

Correlation coefficient

The correlation coefficient for variables x and y computed from n cases is

$$r = \frac{1}{n-1} \sum \left(\frac{x - \bar{x}}{s_x} \right) \left(\frac{y - \bar{y}}{s_y} \right) \qquad (3.3)$$

Here $\bar{x}$ and s_x are the mean and standard deviation of the x observations alone, and similarly $\bar{y}$ and s_y refer to the y observations. As usual, the sum runs over all of the cases for which the variables x and y have been measured.

We can also give a *computing formula* for the correlation coefficient r that eliminates the need to calculate the deviations $x - \bar{x}$ and $y - \bar{y}$ from the means. Like the earlier computing formulas, Equation 1.3 for the variance and Equation 3.2 for the slope of the least squares line, this formula is built up from basic sums. Because it is good practice to compute the means and

* Correlation and its relation to regression is another of Francis Galton's contributions, made in 1888.

standard deviations of each variable as part of the overall description, we give a form of the computing equation that uses the standard deviations.

$$r = \frac{\sum xy - \frac{1}{n}(\sum x)(\sum y)}{(n-1)s_x s_y} \tag{3.4}$$

In statistical practice, r is usually computed by a statistical calculator or by computer software rather than from a formula such as Equation 3.4.

The defining formula Equation 3.3 suggests why r is a measure of association between x and y. Suppose that x and y are the height and the weight of a person. The n cases represent measurements on n people. Height and weight are positively associated, so that larger than average values of x tend to occur together with larger than average values of y. Therefore, the deviations $x - \bar{x}$ and $y - \bar{y}$ from the means will tend to either both be positive (for larger people) or both be negative (for smaller people). In either case, the product $(x - \bar{x})(y - \bar{y})$ will be positive. Hence, r will be positive and will be larger as the positive association grows stronger. In the case of negative association, on the other hand, the deviations $x - \bar{x}$ and $y - \bar{y}$ will tend to have opposite signs, so the sign of r will be negative.

standardized deviations

The use in Equation 3.3 of the *standardized deviations* $(x - \bar{x})/s_x$ and $(y - \bar{y})/s_y$ implies that r measures association between x and y when both variables are measured in standard deviation units about the mean as origin. Changing the unit of measurement of either variable—for example, recording weight in kilograms rather than pounds—does not affect the value of r because both variables are in effect reduced to a standard scale before r is calculated. The properties of correlation will be explored in detail after an example that illustrates the calculation of r.

EXAMPLE 3.12

We will compute the correlation between gas consumption y and heating degree days x for the data below, which first appeared in Example 3.6.

Month	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June
x	15.6	26.8	37.8	36.4	35.5	18.6	15.3	7.9	.0
y	5.2	6.1	8.7	8.5	8.8	4.9	4.5	2.5	1.1

There are $n = 9$ cases. Figure 3.12 shows strong positive linear association in this example. As in the regression calculation of Example 3.7, first calculate the building block sums.

$$\sum x = 15.6 + \cdots + .0 = 193.9$$
$$\sum x^2 = (15.6)^2 + \cdots + (.0)^2 = 5618.11$$
$$\sum y = 5.2 + \cdots + 1.1 = 50.3$$
$$\sum y^2 = (5.2)^2 + \cdots + (1.1)^2 = 341.35$$
$$\sum xy = (15.6)(5.2) + \cdots + (.0)(1.1) = 1375.0$$

Second, calculate the means and standard deviations of both variables. The means and standard deviations are useful in their own right as descriptions of the distributions of x and y.

$$\bar{x} = \frac{1}{n}\sum x = \frac{193.9}{9} = 21.54$$

$$\bar{y} = \frac{1}{n}\sum y = \frac{50.3}{9} = 5.59$$

$$s_x^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{1}{n}(\sum x)^2\right]$$

$$= \frac{1}{8}\left[5618.11 - \frac{(193.9)^2}{9}\right]$$

$$= \frac{1}{8}[5618.11 - 4177.468] = 180.080$$

$$s_x = \sqrt{180.080} = 13.419$$

$$s_y^2 = \frac{1}{n-1}\left[\sum y^2 - \frac{1}{n}(\sum y)^2\right]$$

$$= \frac{1}{8}\left[341.35 - \frac{(50.3)^2}{9}\right]$$

$$= \frac{1}{8}[341.35 - 281.121] = 7.529$$

$$s_y = \sqrt{7.529} = 2.744$$

Finally, substitute into the computing formula Equation 3.4 for the correlation coefficient.

$$r = \frac{\sum xy - \frac{1}{n}(\sum x)(\sum y)}{(n-1)s_x s_y}$$

$$= \frac{1375.0 - \frac{1}{9}(193.9)(50.3)}{(8)(13.419)(2.744)}$$

$$= \frac{291.31}{294.57} = .989$$

As in other multistep calculations, the exact answer depends on how many decimal places are carried in the intervening steps. A computer or statistical calculator will generally give more accurate answers. After a few practice runs to ensure that you understand what the formula for r says, you should automate your arithmetic if possible. ■

To interpret the numerical value of the correlation coefficient r, you must understand its behavior. Here are the basic properties of r.

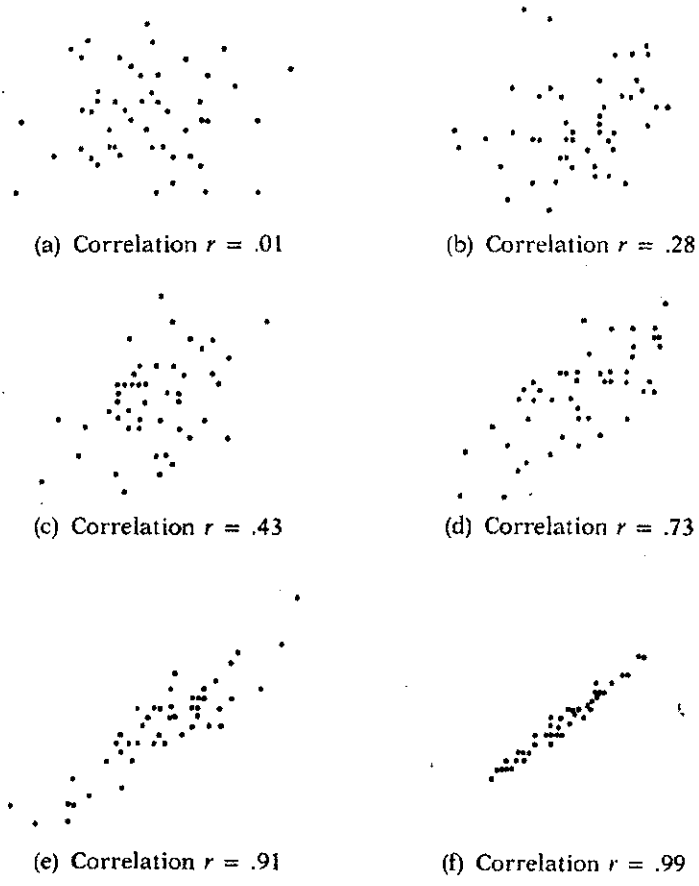
1 The value of r always falls between -1 and 1. Positive r indicates positive association between the variables and negative r indicates negative association.

2 The extreme values $r = -1$ and $r = 1$ occur only in the case of perfect linear association, when the points in a scatterplot lie exactly along a straight line. Values of r close to 1 or -1 indicate that the points lie close to a straight line.

3 The value of r is not changed when the unit of measurement of x, y, or both changes. The correlation r has no unit of measurement; it is a dimensionless number between -1 and 1.

4 Correlation measures only the strength of *linear* association between two variables. Curved relationships between variables, no matter how strong, need not be reflected in the correlation.

The standardization of x and y in Equation 3.3 serves to constrain r to the range -1 to 1. The linear association increases in strength as r moves



(a) Correlation $r = .01$ (b) Correlation $r = .28$

(c) Correlation $r = .43$ (d) Correlation $r = .73$

(e) Correlation $r = .91$ (f) Correlation $r = .99$

Figure 3.27 How the correlation coefficient measures the strength of linear association.

away from 0 toward either -1 or 1. The sign of r indicates only the direction of the association, so that $r = -0.7$ and $r = 0.7$ indicate linear association of the same strength but opposite directions. Understanding that r measures association in a standard scale helps avoid misinterpretation of scatterplots. Stretching or compressing the x or y scale can dramatically alter the appearance of a scatterplot, but does not change the correlation. It is therefore not always easy to guess the value of r from visual inspection of a scatterplot. The scatterplots in Figure 3.27 illustrate how values of r closer to 1 or -1 correspond to stronger linear association. To make the essential meaning of r clear, the standard deviations of both variables in these plots are equal and the horizontal and vertical scales are the same. In general, it is not so easy to guess the value of r from the appearance of a scatterplot.

The nature of the correlation as a measure of linear association is also illustrated by the real data that we have examined. The very linear home heating data in Example 3.12 (Figure 3.12) have a correlation close to 1 ($r = 0.989$). The linear relationship between the percent of votes for Democrats in 1980 and in 1984 (Figure 3.1) is positive but less strong; the correlation is $r = 0.703$. Figure 3.3 shows a quite strong negative association ($r = -0.849$) between mean SAT score and the percent of each state's high school seniors who take the SAT. The association between birth date and draft lottery number in Figure 3.5 is weak and slightly negative; the correlation is $r = -0.226$.

Finally, the correlation coefficient measures the strength of *linear* association only. It is possible to create examples of strong nonlinear association in which the correlation coefficient is small, or even 0 (see Exercise 3.40). For example, the strong nonlinear dependence of corn yield on planting rate was noted in Example 3.3. The correlation between these variables is $r = 0.135$, showing a very small linear association. Correlation is therefore not a general measure of all the types of association that may be visible in a scatterplot.

Correlation and Regression

Although correlation does not presuppose an explanatory-response relationship as regression does, the correlation coefficient r is meaningful for regression as well. In fact, the numerical value of r is most clearly interpreted from the following fact about regression.

r^2 in regression

> The square of the correlation coefficient, r^2, is the fraction of the variation in the values of y that is explained by the least squares regression of y on x. Moreover, the roles of x and y in this interpretation can be interchanged.

To understand this fact intuitively, consider again the scatterplot of household natural gas consumption y versus heating degree days x which appears in Figure 3.28(a). The horizontal dashed line marks the mean gas consumption $\bar{y}$. Gas consumption shows considerable variation from month

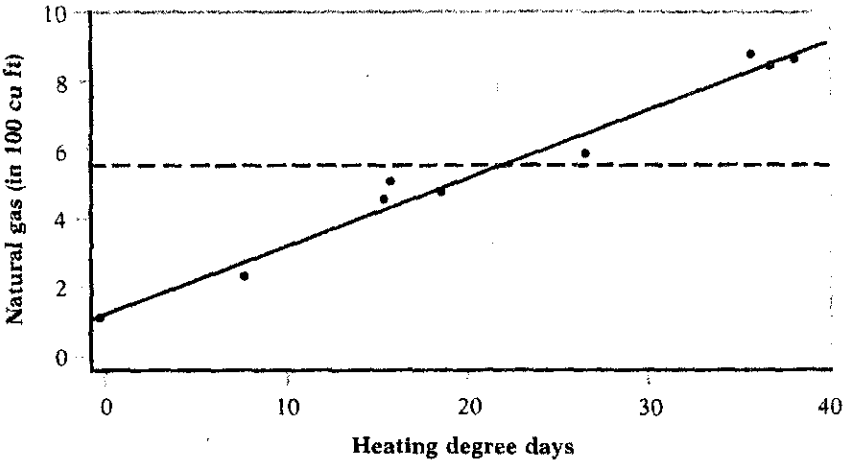


Figure 3.28(a) High r^2: The variation in y about the regression line is much less than the variation in y about the mean $\bar{y}$; most of the variation in y is explained by the linear relationship of y and x.

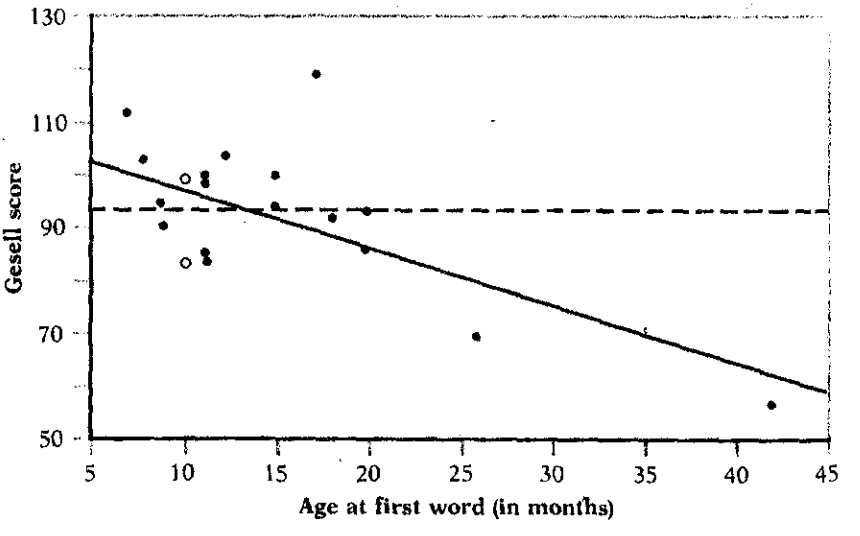


Figure 3.28(b) Low r^2: The variation of y about the regression line remains large; less of the variation in y is explained by the linear relationship between y and x.

to month, as indicated by the vertical deviations of the points from the dashed line. The colored line is the least squares regression line. The vertical deviations of the points from this line are small. That is, when x changes, y changes with it, and this linear relationship accounts for almost all of the variation in y. Since $r = 0.989$ in this example, $r^2 = 0.978$ and we can say that the linear regression explains 97.8% of the observed variation in gas consumption.

On the other hand, the scatterplot of Gesell score y versus age at first word x in Figure 3.28(b) shows a weaker linear relationship. The vertical variation about the colored regression line, although smaller than the variation about the mean $\bar{y}$ (dashed line), is still large. The linear tie between x and y explains a smaller fraction of the observed variation in y. In fact, since $r = -0.640$ and $r^2 = 0.410$, we can say that 41% of the variation *in either variable* is explained by linear regression on the other variable. As Figure 3.19 illustrates, the regression lines of y on x and of x on y are quite different. But there is only a single correlation r between x and y (in either order), and r^2 helps interpret both regressions. The correlation between Gesell score and age at first word is negative, since late talkers tend to have lower aptitude scores. The interpretation of r through r^2 makes it clear that the magnitude of r, not its sign, measures the strength of a linear association.

We have gained not only more insight into interpreting correlation, but also a valuable numerical measure of the usefulness of a least squares regression line. Since the goal of regression is to explain y by linear dependence on x, r^2 is a direct measure of the success of the regression and is almost always reported along with the regression results. This close connection with correlation is specifically a property of least squares regression and is not shared by more resistant methods of fitting a line to a scatterplot.

The use of r^2 to measure how successfully a regression line explains the observed variation in y is not the only connection between correlation and regression. There is a close relationship between r and the slope b of the least squares regression line $\hat{y} = a + bx$. Some algebra based on Equations 3.1 and 3.3 establishes the following fact:

Regression slope

If s_x and s_y are the standard deviations of the observed x_i and y_i, and r is the correlation coefficient, then the slope of the least squares regression line of y on x is

$$b = r\frac{s_y}{s_x} \tag{3.5}$$

That is, the least squares regression line of y on x is the line with slope rs_y/s_x that passes through the point $(\bar{x}, \bar{y})$. Regression can therefore be described entirely in terms of the basic descriptive measures $\bar{x}$, s_x, $\bar{y}$, s_y, and r. If both x and y are standardized variables, so that their means are 0 and

their standard deviations are 1, then the regression line has slope r and passes through the origin.

Equation 3.5 is the formula for the slope b that is easiest to understand and remember. This equation says that along the regression line, *a change of one standard deviation in x corresponds to a change of r standard deviations in y.* When the variables are perfectly correlated ($r = 1$ or $r = -1$), the change in the response y is the same (in standard deviation units) as the change in x. Otherwise, since $-1 \le r \le 1$, the change in y is less than the change in x. As the correlation grows less strong, y moves less in response to changes in x.

EXAMPLE 3.13

In Example 3.12 we found that for heating degree days x and natural gas consumption y,

$$\bar{x} = 21.54 \quad \text{and} \quad s_x = 13.42$$
$$\bar{y} = 5.59 \quad \text{and} \quad s_y = 2.74$$
$$r = .989$$

The slope of the regression line of gas usage on degree days is therefore

$$b = .989 \frac{2.74}{13.42} = .202$$

in agreement with the result of Example 3.7.

The regression line passes through the point $(\bar{x}, \bar{y})$, which is (21.54, 5.59). Along the line, y increases by $b = 0.202$ when x increases by 1. In terms of correlation, y increases by $r = 0.989$ standard deviations when x increases by one standard deviation. ■

Interpreting Correlation and Regression

Limitations of correlation and regression Correlation and regression are powerful tools for measuring the association between two variables and for expressing the dependence of one variable on the other. These tools must be used with an awareness of their limitations, beginning with the fact that they apply to *only linear* association or dependence. Also remember that *neither r nor the least squares regression line is resistant.* One influential observation or incorrectly entered data point can greatly change these measures.

EXAMPLE 3.14

We saw in Example 3.10 that case 18 is an influential observation in the regression of the Gesell score y on age at first word x for young children. In fact, the correlation based on all 21 children is $r = -0.640$. Since $(-0.64)^2 = 0.41$, age at first word appears to explain 41% of the variation in Gesell score among children. But if case 18 is omitted, the correlation for the remaining 20 children is only $r = -0.335$. Only 11% of the variation in aptitude score among these 20 children is explained by the age at which they first spoke. Excluding case 19,

which is an outlier, we have $r = -0.756$ and $r^2 = 0.572$ or 57%. Case 19 is also influential, though not as much as case 18. The least squares line of y on x for the complete data set is

$$\hat{y} = 109.87 - 1.127x$$

but if case 18 is excluded the regression line becomes

$$\hat{y} = 105.63 - .779x$$

These regression lines are graphed in Figure 3.20. This single observation of case 18 dramatically changes both the correlation and the fitted line. A decision to exclude this child as not belonging to the same population as the other children will weaken the study's conclusion that Gesell score can be partially predicted from age at first word. Just as calculation often adds to the information provided by a scatterplot, a plot is essential if calculation is not to be blind. Without a plot to help spot the influential observation, numerical calculations for these data can be seriously misleading. ■

Lurking variables We have seen repeatedly that the effect of variables not included in a study can render a correlation or regression misleading. Examples 3.9 and 3.11 both illustrate the effect of such lurking variables. To give another example, there is a strong positive correlation over time between teachers' salaries and sales of liquor. Both increase with rising price levels and general prosperity, creating a strong association. Such correlations are sometimes called "nonsense correlations," but the correlation is perfectly real. What *is* nonsense is the conclusion that because the correlation exists, teachers must be spending their salary increases on liquor. *Even a strong correlation does not imply any cause and effect relationship.* The question of causation is important enough to merit separate treatment in Section 3.5. For now, just remember that a correlation between two variables x and y can reflect many types of relationship between x, y, and other variables not explicitly recorded.

The effect of lurking variables can hide a true relationship between x and y as well as create an apparent relationship, as the following example illustrates.

EXAMPLE 3.15

A study of housing conditions and health in the city of Hull, England measured a large number of variables for each of the wards in the city. (A ward is a small, relatively homogeneous geographic area.) Two of the variables were an index x of overcrowding and an index y of the lack of indoor toilets. Since x and y are both measures of inadequate housing, we expect a high correlation. In fact, the correlation was only $r = 0.08$. How can this be? Investigation disclosed that some poor wards were dominated by public housing (called council housing in England). These wards had high values of x but low values of y because council housing always includes indoor toilets. Other deprived wards lacked council housing, and in these wards high values of x were accompanied by high values of y. Because the relationship between x and y differed in council and noncouncil wards, analyzing all wards together obscured the nature of the relationship.[10] ■

Figure 3.29 shows in simplified form how groups formed by a lurking (categorical) variable as in Example 3.15 can make the correlation r misleading. The groups appear as clusters of points in the scatterplot. There is a strong relationship between x and y within each of the clusters; in fact, $r = 0.85$ and $r = 0.91$ in the two clusters. However, because similar values of x correspond to quite different values of y in the two clusters, x alone is of little value in predicting y. The correlation for all points displayed is therefore low; in fact, $r = 0.14$. This example is another reminder to plot the data rather than to simply calculate numerical measures such as the correlation.

Prediction A regression line is often used to *predict* the response y to a given value x of the explanatory variable. This is clearly valid when the regression reflects a cause and effect relationship and when r^2 is high enough to give us confidence that changes in x explain most of the variation in y. For example, we can predict household natural gas consumption at different outside temperatures (degree days), using the regression line calculated in Example 3.7. However, *successful prediction does not require a causal relationship.* If both x and y respond to the same underlying unmeasured variables, it may be possible to predict y from x even though x has no direct influence on y.

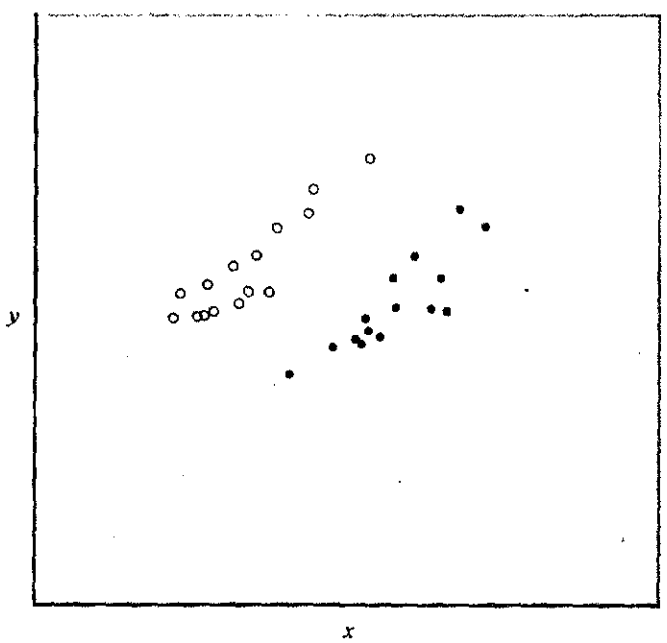


Figure 3.29 This scatterplot has a low r^2 even though there is a strong correlation within each of the two clusters.

EXAMPLE 3.16

Decisions on which applicants to admit to graduate school are based on a number of criteria, such as the applicant's undergraduate grades. Another criterion is scores on the Graduate Record Examinations (GRE), national tests of both aptitude and knowledge administered by the Educational Testing Service. There is no causal relationship between the score x a college senior achieves on the GRE and the student's grade point average (GPA) y as a first-year graduate student the following year. But because both x and y respond to the student's level of ability and knowledge, it is plausible to predict y from x.

The success of this prediction depends on the strength of the association between GRE scores and later GPA in graduate school. A study[11] of a large number of first-year students at many graduate schools in many disciplines showed that for students of economics

- The correlation between the GRE verbal aptitude score and graduate GPA was $r = 0.09$.
- The correlation between the GRE quantitative aptitude score and graduate GPA was $r = 0.34$.
- The correlation between the GRE economics advanced test score and graduate GPA was $r = 0.45$.
- The correlation between undergraduate GPA and graduate GPA was $r = 0.27$.

These results show that the verbal aptitude test bears little relation to success as a graduate student of economics; but they also show that a student's score on the GRE economics advanced test is a better predictor of success than the undergraduate GPA. However, even the GRE economics test accounts for only 20% of the variation in first-year graduate GPA [because $r^2 = (0.45)^2 = 0.20$]. Although prediction from a regression line makes sense in this setting, prediction based on the GRE advanced score alone will be quite unreliable. ■

Notice that for the purpose of assessing whether GRE scores are helpful in making admissions decisions, the data in Example 3.16 are incomplete in a systematic way. We have GPA information only for students who were admitted to graduate school. Because many students with low GRE scores were not admitted, the data refer primarily to students who did well on the GRE. The reported correlations describe the relation between GRE scores and graduate GPAs for students who make it to graduate school. They do not tell us whether students with GRE scores too low to allow them admission would in fact have earned low grades if they had entered graduate school.

Even when prediction is logically justified and r^2 is high, several additional cautions are in order. Chapter 2 noted the danger of *extrapolation*, the use of a regression line for prediction at values of x removed from the range of x-values used to fit the line. Most relationships remain linear only over a restricted range of x, so extrapolation can yield silly results. An investigation of the firing time y of blasting caps used in mining showed a strong linear response to the voltage x applied to the detonator. This is true, however, only over the range of voltages used in practice. The overall relationship between voltage and firing time is similar to that shown in Figure 3.30.

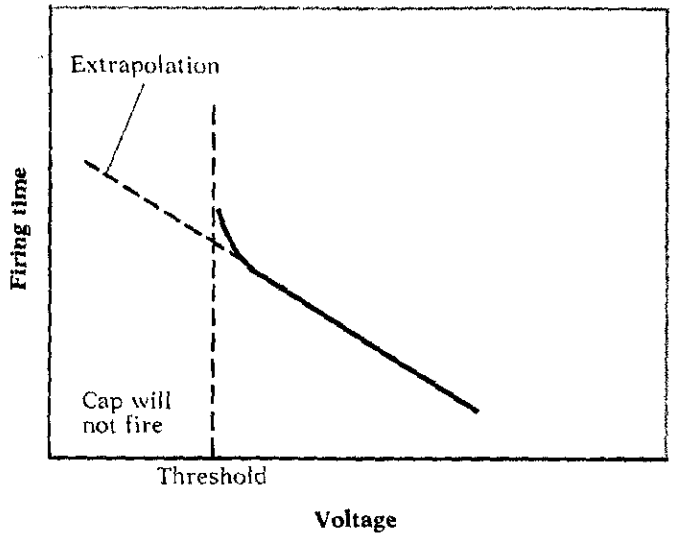


Figure 3.30 Beware of extrapolation: The relationship between the voltage applied and the firing time of a blasting cap is linear only in a restricted range of voltages.

Below a threshold voltage, the blasting cap will not detonate at all. Extrapolation of the regression line to low voltages is meaningless.

Using averaged data Many regression or correlation studies work with averages or other measures that combine information from many individuals. You should note this carefully and resist the temptation to apply the results of such studies to individuals. The regression of natural gas consumption on heating degree days, for example, was based on daily data averaged over each month. The very high correlation observed does not apply to individual days. Prediction of gas usage from outside temperature on a single day would be much less reliable than is suggested by the high r^2 in the regression of Examples 3.7 and 3.12. The reason for this is that averaging over an entire month smooths out much of the day-to-day variation due to doors left open, house guests using more gas to heat water, and so on. Figure 3.31 shows the gas consumption and degree days for several individual days in April (A), May (M), and November (N). There is considerable variation within each month and the correlation for these individual observations would be moderate. But when only the three monthly averages (marked as •) are recorded, the result is three of the points in Figure 3.12, with a correlation near 1. *Correlations based on averages are usually too high when applied to individuals.* This is another reminder that it is important to note exactly what variables were measured in a statistical study.

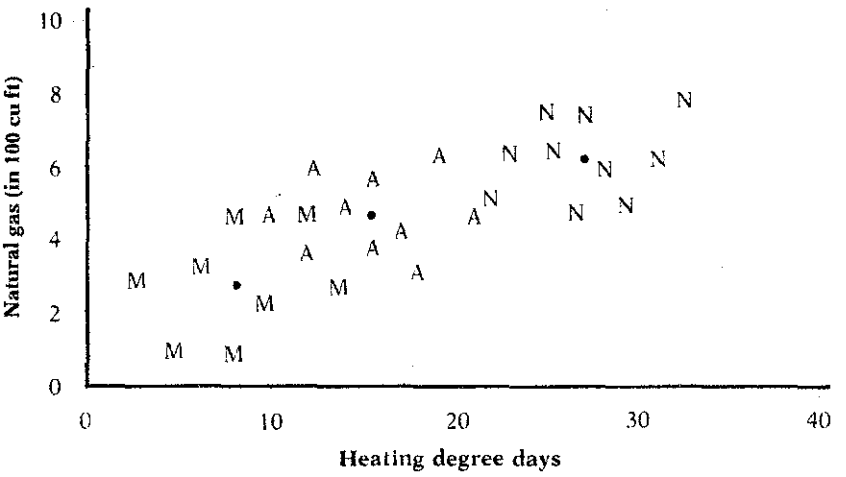


Figure 3.31 Natural gas consumption plotted against degree days for individual days in April (A), May (M), and November (N), and the averages of both variables in the three months (•). The correlation is much higher for the averaged data.

Correlation is not everything Finally, remember that *correlation is not a complete description of two-variable data.* The means and standard deviations of both x and y should be given along with the correlation. (Because correlation and regression make use of means and standard deviations, these nonresistant measures are appropriate to accompany a correlation.) Conclusions based on correlations alone may require rethinking in the light of a more complete description of the data.

EXAMPLE 3.17

Competitive divers are scored on their form by a panel of judges, who use a scale from 1 to 10. The subjective nature of this scoring often results in controversy. We have the scores awarded by two judges, Ivan and George, on a large number of dives. How should we assess their agreement? Some computation shows that the correlation between their scores is $r = 0.9$. But the mean of Ivan's scores is 3 points lower than George's mean.

These facts do not contradict each other. They are simply different kinds of information. Ivan awards much lower scores than George, as the mean scores reveal. But because Ivan gives *every* dive a score about 3 points lower than George, the correlation remains high. Remember that adding or subtracting the same number to all values of either x or y does not change the correlation. If Ivan and George both rate several divers, the contest is consistently scored because, as the high correlation shows, Ivan and George agree on which dives are better than others. But if Ivan scores one diver and George another, we must add 3 points to Ivan's scores to arrive at a fair comparison. ■

Many statistical studies of important issues rely on correlation and regression to describe complex relationships among many variables. The examples in this section demonstrate that even when only two variables are involved, the numerical results of correlation or regression must be interpreted in the light of an understanding both of the behavior of these statistical procedures and of the full background of the issue under study. The numerical work is easily automated, and you should let a computer do it. Interpretation requires an informed and skeptical human mind.

SUMMARY

The **correlation coefficient** r measures the strength and direction of the linear association between two quantitative variables x and y. Correlation always satisfies $-1 \le r \le 1$, and $r = \pm 1$ only in the case of perfect linear association. The value of r is not affected by changes in the unit of measurement of either variable.

Correlation and regression are closely connected. The squared correlation coefficient r^2 is the fraction of the variation of one variable that is explained by least squares regression on the other variable. The regression line of y on x is the line with slope $b = rs_y/s_x$ that passes through the point $(\bar{x}, \bar{y})$.

A correlation or regression should be interpreted with due attention to the following: the possible effects of lurking variables, the lack of resistance of these procedures, the danger of extrapolation, the fact that correlations based on averages are usually too high for individuals, and an understanding that correlation and regression measure only linear relationships to the exclusion of other important aspects of the data.

SECTION 3.3 EXERCISES

3.35 A student wonders if people of similar heights tend to date each other. She measures herself, her dormitory roommate, and the women in the adjoining rooms; then she measures the next man each woman dates. Here are the data (heights in inches).

Women	66	64	66	65	70	65
Men	72	68	70	68	71	65

(a) Make a scatterplot of these data. Based on the scatterplot, do you expect the correlation to be positive or negative? Near ± 1 or not?

Section 3.3 Exercises

(b) Use Equation 3.4 to compute the correlation r between the heights of the men and women.
(c) How would r change if all the men were 12 inches shorter than the heights given in the table? Does the correlation help answer the question of whether women tend to date men taller than themselves?
(d) If every woman dated a man exactly 3 inches taller than she, what would be the correlation between male and female heights?

3.36 The growth of young children is nearly linear. Here again are data on Sarah's height at several ages. (See Exercise 3.17 for the regression line.)

Age (months)	36	48	51	54	57	60
Height (cm)	86	90	91	93	94	95

(a) From a scatterplot of height versus age, explain why you would expect the correlation to be close to 1.
(b) Compute the correlation coefficient r between height and age, using Equation 3.4.
(c) If Sarah were 4 centimeters taller at every age, how would the value of r change?

3.37 Compute the mean and standard deviation for the heights of the men and women in Exercise 3.35. Use your results and the correlation found in Exercise 3.35 to compute the slope of the regression line of male height on female height. What is the slope of the regression of female height on male height (when male height is on the x axis and female height is on the y axis)? If both lines were drawn on the same graph, with female height on the x axis, at what point would they intersect?

3.38 Compute the mean and the standard deviation of both height and age in Exercise 3.36. Use these values and the correlation from Exercise 3.36 to find the slope of the regression line of height on age. (Compare your result with the slope you found in Exercise 3.17.) What is the slope of the regression line of age on height?

3.39 Exercise 3.6 compares two methods of computing the value of a dollar in foreign currencies: by comparing the cost of a McDonald's Big Mac and the official exchange rate.

(a) Calculate the correlation r between the Big Mac value and the official value of the dollar.
(b) The Big Mac value of the dollar in Japan was 231 Japanese yen, while the official exchange rate was 154 yen to the dollar. Add this observation to the data set and calculate the correlation

again. Explain carefully why adding this one observation causes such a large increase in r.

3.40 The gas mileage of an automobile first increases and then decreases as speed increases. Suppose that this relationship is very regular, as shown by the following data on speed (miles per hour) and mileage (miles per gallon):

Speed	20	30	40	50	60
Mileage	24	28	30	28	24

Make a scatterplot of mileage versus speed. Show that the correlation between speed and mileage is $r = 0$. (Note that to show that $r = 0$, you need only compute the numerator in Equation 3.4.) Explain why the correlation is 0 even though there is a strong association between speed and mileage.

3.41 A college newspaper interviews a psychologist about a proposed system for rating the teaching ability of faculty members. The psychologist says, "The evidence indicates that the correlation between a faculty member's research productivity and teaching rating is close to zero." The paper reports this as "Professor McDaniel said that good researchers tend to be poor teachers, and vice versa." Explain why the paper's report is wrong. Write a statement in plain language (don't use the word "correlation") explaining the psychologist's meaning.

3.42 Each of the following statements contains a blunder. Explain in each case what is wrong.

(a) "There is a high correlation between the sex of American workers and their income."

(b) "We found a high correlation ($r = 1.09$) between students' ratings of faculty teaching and ratings made by other faculty members."

(c) "The correlation between planting rate and yield of corn was found to be $r = 0.23$ bushel."

3.43 A study of class attendance and grades among freshmen at a state university showed that in general students who attended a higher percent of their classes earned higher grades. Class attendance explained 16% of the variation in grade index among the freshmen studied. What is the numerical value of the correlation between percent of classes attended and grade index?

3.44 Suppose that the heights of the men in Exercise 3.35 were measured in centimeters (cm) rather than in inches, but that the heights of the women remained in inches. (There are 2.54 cm to an inch.)

Section 3.3 Exercises

(a) What would now be the correlation between male and female height? (Use information from Exercise 3.35—don't compute the new r directly.)
(b) What would be the slope of the regression line of male height on female height? (Use your calculations from Exercise 3.37—don't compute a new regression line. Hint: Use Equation 3.5 for the slope b.)

3.45 Changing the units of measurement can dramatically alter the appearance of a scatterplot. Consider the following data:

x	−4	−4	−3	3	4	4
y	.5	−.6	−.5	.5	.5	−.6

(a) Draw x and y axes each extending from −6 to 6. Plot the data on these axes. Then plot $x^* = x/10$ against $y^* = 10y$ on the same axes using a different plotting symbol. The two plots are very different in appearance.
(b) The correlation between x and y is about $r = 0.25$. What must be the correlation between x^* and y^*?
(c) Will the regression line of y^* on x^* have the same slope as the regression line of y on x? Explain your answer. (Hint: Look at Equation 3.5 for the slope.)

3.46 Return to the scatterplot in Figure 3.2 showing the percent of presidential votes cast for Democrats in 1980 and 1984, with the south emphasized. The correlation for all 50 states is $r = 0.703$. Would r be higher or lower if the 10 southern states were omitted? Why?

3.47 The full MINITAB computer output for Example 3.9 contains the entry `R-sq = 69.4%`. Explain what this means in this specific example, in language that can be understood by someone who knows no statistics.

3.48 Figure 3.10 shows how the price of four-door sedans varies with their weight.
(a) If only foreign models are considered, the correlation between weight and price is $r = 0.707$. What percent of the variation in the prices of these foreign cars can be explained by the fact that price increases linearly as weight increases?
(b) From examination of Figure 3.10, do you think that the linear relation of price to weight explains a higher or a lower percent of the price variation of the domestic models in the study? Why?

3.49 A study of erosion (see Exercises 3.5 and 3.30) produced the following data on the rate at which water flows across land and the resulting amount of erosion:

Flow rate	.31	.85	1.26	2.47	3.75
Eroded soil	.82	1.95	2.18	3.01	6.07

What percent of the variation in the amount of erosion can be explained by the fact that as the flow rate increases, erosion increases with it in a linear manner?

3.50 The mean height of American women in their early twenties is about 65.5 inches and the standard deviation is about 2.5 inches. The mean height of men the same age is about 68.5 inches, with standard deviation about 2.7 inches. If the correlation between the heights of husbands and wives is about $r = 0.5$, what is the slope of the regression line of the husband's height on the wife's height in young couples? Draw a graph of this regression line. Predict the height of the husband of a woman who is 67 inches tall.

3.51 In a large economics class, the correlation between a student's total score prior to the final examination and the final examination score is $r = 0.6$. The pre-exam totals for all students in the course have mean 280 and standard deviation 30. The final exam scores have mean 75 and standard deviation 8. The professor has lost Julie's final exam but knows that her total before the exam was 300. He decides to predict her final exam score from her pre-exam total.

(a) What is the slope of the regression of final exam scores on pre-exam total scores in this course?

(b) Draw a graph of this regression line and use it to predict Julie's final examination score.

3.52 The British government conducts regular surveys of household spending. The following table shows the average weekly household spending on tobacco products and alcoholic beverages for each of the

Region	Alcohol	Tobacco
North	£6.47	£4.03
Yorkshire	6.13	3.76
Northeast	6.19	3.77
East Midlands	4.89	3.34
West Midlands	5.63	3.47
East Anglia	4.52	2.92
Southeast	5.89	3.20
Southwest	4.79	2.71
Wales	5.27	3.53
Scotland	6.08	4.51
Northern Ireland	4.02	4.56

11 regions of Great Britain. (Data from British official statistics, *Family Expenditure Survey*, Department of Employment, 1981.)

(a) Make a scatterplot of spending on tobacco against spending on alcohol.
(b) Describe the pattern of the plot. Circle the most influential observation.
(c) The correlation is only $r = 0.224$. Compute the correlation for the 10 regions with Northern Ireland omitted. Explain why this r differs so greatly from the r for all 11 cases.

3.53 Example 3.12 illustrates the computation of the correlation coefficient r. Redo that computation, rounding off all intermediate steps to the nearest whole number. Compare your answer with the result in the example. (Use the building block sums found in the example as your starting point, but round them to the nearest whole number before proceeding.) Remember that your numerical answers for b, r, and other descriptive measures that require long calculations will vary as you carry more or fewer significant digits in the intermediate steps.

3.54 There is a strong positive correlation between years of education and income for economists employed by business firms. (In particular, economists with doctorates earn more than economists with only a bachelor's degree.) There is also a strong positive correlation between years of education and income for economists employed by colleges and universities. But when all economists are considered, there is a *negative* correlation between education and income. The explanation for this is that business pays high salaries and employs mostly economists with bachelor's degrees, while colleges pay lower salaries and employ mostly economists with doctorates. Sketch a scatterplot with two groups of cases (business and academic) which illustrates how a strong positive correlation within each group and a negative overall correlation can occur at the same time. (Hint: Begin by studying Figure 3.29.)

3.55 If you have a statistical computing system, enter the data on the relation between Gesell score and age at first word (Table 3.3). Compute the correlation and the least squares regression line for the data with *both* cases 18 and 19 omitted. How do the results, particularly r^2, compare with those given in Example 3.14?

3.56 Refer to the data on wood strength used in Exercises 3.16, 3.31, and 3.32.

(a) Compute the correlation between T1 and T2.
(b) Using the slope of the least squares line, the standard deviations of x and y, and the correlation found in (a), verify that Equation 3.5 holds for this set of data.

3.57 Refer to the data on women's stride rates given in Exercise 3.21.

(a) Compute the correlation between speed and stride rate.

(b) What proportion of the variation in stride rate is explained by speed for this set of data?

(c) Repeat (a) and (b) using only the data for speeds 15.86, 16.88, 17.50, and 18.62. Do the results change appreciably? From the plotting that you have done, would you expect the least squares line to change appreciably? Explain what you have learned from this part of the exercise.

3.58 The data in Exercise 3.21 relating stride rate in strides per second to running speed give the average stride rate of 21 elite female runners at each speed. Suppose that you had data on many individual time periods for all 21 runners. If you plotted each individual stride rate at each speed and computed the correlation for these individual data, would you expect the correlation between stride rate and speed to be lower than, about the same as, or higher than the correlation for the published data? Sketch a scatterplot of stride rate versus speed for individual runners to illustrate your answer.

3.59 The price of seafood varies with species and time. The following table gives the prices in cents per pound received in 1970 and 1980 (PR70 and PR80) by fishermen and vessel owners for several species:

Species	PR70	PR80
Cod	13.1	27.3
Flounder	15.3	42.4
Haddock	25.8	38.7
Menhaden	1.8	4.5
Ocean perch	4.9	23.0
Salmon, chinook	55.4	166.3
Salmon, coho	39.3	109.7
Tuna, albacore	26.7	80.1
Clams, soft-shelled	47.5	150.7
Clams, blue hard-shelled	6.6	20.3
Lobsters, American	94.7	189.7
Oysters, eastern	61.1	131.3
Sea scallops	135.6	404.2
Shrimp	47.6	149.0

(a) Plot the data with PR70 on the x axis and PR80 on the y axis.

(b) Describe the overall pattern. Are there any outliers or points that may be highly influential? If so, label them.

(c) Compute the correlation for the entire set of data.

(d) What proportion of the variation in 1980 prices is explained by the 1970 prices?
(e) Recompute the correlation discarding the cases that you labeled in (b).
(f) To what extent do you think the correlation provides a good measure of the relationship between the 1970 and 1980 prices for this set of data? Explain your answer.

3.4 RELATIONS IN CATEGORICAL DATA

Up to this point our focus has been on relationships between quantitative variables, although categorical variables played an important role in Section 3.1. Now our focus will shift to describing relationships between two or more categorical variables. Some variables—such as sex, race, and occupation—are inherently categorical. In other cases, categorical variables are created by grouping values of a quantitative variable into classes. Published data are often reported in this form to save space. Analysis of categorical data is based on the counts or percents of the cases that fall into various categories.

EXAMPLE 3.18

two-way table

Table 3.4 presents Census Bureau data on the educational attainment of Americans of different ages.[12] Because many persons under 25 years of age have not completed their education, they are not included in the table. Both variables, age and education, have been grouped into categories. The entries in this *two-way table* are the frequencies, or counts, of persons in each age by education class. Although both age and education as presented in this table are categorical variables, both have a natural order from least to most. The order of the rows and the columns in Table 3.4 reflects the order of the categories. ■

Table 3.4 Educational attainment by age, 1984 (thousands of persons)

	Age group					
Education	25–34	35–44	45–54	55–64	≥65	Total
Did not complete high school	5416	5030	5777	7606	13,746	37,577
Completed high school	16,431	11,855	9435	8795	7558	54,073
College, 1–3 years	8555	5576	3124	2524	2503	22,282
College, 4 or more years	9771	7596	3904	3109	2483	26,862
Total	40,173	30,058	22,240	22,033	26,291	140,794

2.49 (a) Apr., May, June, July are high; Sept. and Dec. are low. (b) Yes, the secondary peak is between Oct. and Nov. (c) Months with high incidence of diarrhea are followed by months with low weight. Height is not as sensitive to the short-term effects of illness.

2.51 (a) Phase One (0 to 6 hours): no increase in growth. Phase Two (9 to 24 hours): exponential growth. Phase Three (36 hours): growth is slower than exponential. (b) Predicted log is .4230. Predicted colony size is 2.65. (c) 1.16.

2.53 (b) This control chart is very similar to the control chart found in Exercise 2.4. (c) They are not sensitive to outliers.

2.55 (b) Exponential, exponential, linear.

2.57 (c) Day 5 is above; day 24 is below. (d) No increase or decrease is evident. (e) Thomas' flea eggs show periods of increase and decrease.

2.59 Overall the pattern is neither linear nor exponential. Over short periods the pattern is approximately linear.

Chapter 3

3.1 (a) Categorical. (b) Quantitative. (c) Quantitative. (d) Categorical. (e) Quantitative. (f) Quantitative.

3.3 (a) Negative. Clearly curved. One observation is high on nitrogen oxides. (b) No. Low nitrogen oxide is associated with high carbon monoxide.

3.5 (b) As the flow rate increases, the amount of eroded soil increases. Yes. Positive.

3.7 (a) Heavier cars cost more. The association is weak and positive. (b) The heaviest cars, greater than 3500 pounds, are all domestic cars. In the range of 2500 to 3500 pounds, the foreign cars generally cost more than domestic cars of similar weight.

3.9 (b) No clear relationship is evident. (c) Those who survive appear to be older and have longer incubation periods. (d) The two survivors with incubation periods greater than 70.

3.11 (a) The number of fleas increases and then decreases. (b) The same pattern is evident. The 3-day median trace gives a slightly better picture of the pattern.

3.13 (a) Means = 10.65, 10.43, 5.60, 5.45. (b) The introduction of 1000 nematodes per pot has no effect on seedling growth. With 5000 nematodes there is a substantial reduction in seedling growth. Introduction of 10,000 nematodes causes essentially the same growth reduction as 5000.

3.15 (a) Means = 1520, 1707, 1540, 1816. (b) Against. Pecking order 1 has the lowest mean weight, and pecking order 4 has the highest mean weight.

3.17 (a) .38, 71.95, .38 cm. (c) 87.91 cm.

3.19 Predicted heights = 85.75, 90.35, 91.50, 92.65, 93.80, 94.95. Residuals = .25, −.35, −.50, .35, .20, .05. There is no clear pattern in the residuals.

3.21 (a) Yes. (b) b = .080, a = 1.766. (c) 3.039, 3.121, 3.171, 3.261, 3.369, 3.457, 3.541. (d) .011, −.001, −.001, −.011, −.009, .003, .009. (e) The residuals are all very small, indicating that the line fits the data well. Positive residuals

ศูนย์บรรณสารและสื่อการศึกษา
มหาวิทยาลัยเทคโนโลยีสุรนารี

are associated with high and low speeds, and negative residuals are associated with intermediate speeds.

3.23 **(b)** $\hat{y} = 1.03 + .90x$. **(c)** Predicted = 5.52, 10.03, 14.53, 19.03. Yes.

3.25 **(a)** Rural = 108, city = 123. No. **(b)** 7.4820. **(c)** Yes. **(d)** 93.65.

3.27 **(a)** The relationship is linear with two clusters and one observation that is low in calories and sodium. **(b)** Black. It is closer to the influential observation. **(c)** 407.05.

3.29 **(a)** Yes. The round one scores of 102 and 105 are influential, and the round one score of 105 appears to be an outlier. **(b)** $\hat{y} = 26.332 + .688x$. **(c)** Predicted = 87.54, 88.23, 86.17, 91.67, 85.48, 82.04, 96.48, 98.55, 83.42, 86.85, 88.92, 80.66. **(d)** Residuals = 6.46, −3.23, 2.83, −2.67, −4.48, −6.04, 10.52, −9.55, 3.58, 4.15, −.92, −.66. **(e)** There is a random pattern with large residuals for the high round one scores.

3.4 RELA

3.31 **(a)** Yes. **(b)** $\hat{y} = -.033 + 1.018x$. **(d)** No clear pattern. The linear fit is adequate. **(e)** The last 15 residuals are all positive. **(f)** The distribution is approximately symmetric, approximately normal.

3.33 **(b)** There is one large outlier. **(c)** No. No.

3.35 **(a)** Positive. Not. **(b)** $r = .56533$. **(c)** It does not change. No. **(d)** $r = 1$.

3.37 Means = 69, 66, Standard deviations = 2.52982, 2.09762. Slopes = .6818, .4687. At the means.

3.39 **(a)** $r = .77440$. **(b)** $r = .99824$. The additional point is very extreme in both x and y.

3.41 The paper suggests a negative relationship. The psychologist is saying there is no linear relationship.

3.43 $r = .4$

EXAMPLE 3.11

3.45 **(b)** $r = .25$. **(c)** No. It will be 100 times as large.

3.47 It is the percentage of variation in the number of students enrolled in 100 level mathematics courses explained by the linear relationship with the number of students in the freshman class.

two-way table

3.49 93.5%.

3.51 **(a)** .16. **(b)** 78.2.

3.53 $r = .9519$.

3.55 $r = -.52$, $r^2 = .27$, $\hat{y} = 107.585 - 1.050x$. There is a substantial decrease in r^2.

3.57 **(a)** .99899. **(b)** .9980. **(c)** $r = .99967$, $r^2 = .9993$. No appreciable change in the correlation or the line.

3.59 **(b)** There is a positive linear relationship. No outliers. Sea scallops is an influential observation. **(c)** .96704. **(d)** .9352. **(e)** .93996. **(f)** The correlation indicates that the linear relationship is strong.

3.61 **(a)** 45.90%, 45.08%, 9.02%. **(b)** 60.66%, 39.34%. **(d)** Yes. The percentages for mild, moderate, and severe are similar for each type of operation.

3.63 **(a)** 23,403. Round-off error. **(b)** 56.19%, 32.32%, 7.75%, 3.73%. **(c)** 4 years of high school: 30.27%, 43.05%, 16.68%, 10.00%. 1 to 3 years of college: 22.38%, 42.36%, 19.48%, 15.78%. 4 or more years of college: 10.22%,