agpop.dat   Data from the U.S. 1992 Census of Agriculture. In columns 3-14, the value "-99" denotes missing data.

Column	Name	Value
1	county	county name
2	state	state abbreviation
3	acres92	number of acres devoted to farms, 1992
4	acres87	number of acres devoted to farms, 1987
5	acres82	number of acres devoted to farms, 1982
6	farms92	number of farms, 1992
7	farms87	number of farms, 1987
8	farms82	number of farms, 1982
9	largef92	number of farms with 1,000 acres or more, 1992
10	largef87	number of farms with 1,000 acres or more, 1987
11	largef82	number of farms with 1,000 acres or more, 1982
12	smallf92	number of farms with 9 acres or fewer, 1992
13	smallf87	number of farms with 9 acres or fewer, 1987
14	smallf82	number of farms with 9 acres or fewer, 1982
15	region	region of country (W = West, NC = North Central, S = South, N = Northeast)

agpps.dat Data from a without-replacement pps sample from file agpop.dat.

Column	Name	Value
1	county	county name
2	state	state abbreviation
3	acres92	number of acres devoted to farms, 1992
4	acres87	number of acres devoted to farms, 1987
5	sizemeas	size measure used to select the pps sample
6	SelectionProb	inclusion probability for county, π_i
7	SamplingWeight	sampling weight for county, $w_i = 1/\pi_i$
8–22	jtprob1–jtprob15	columns of joint inclusion probabilities

agsrs.dat   Data from an SRS of size 300 from the U.S. 1992 Census of Agriculture.
In columns 3-14, the value "-99" denotes missing data.

Column	Name	Value
1	county	county name
2	state	state abbreviation
3	acres92	number of acres devoted to farms, 1992
4	acres87	number of acres devoted to farms, 1987
5	acres82	number of acres devoted to farms, 1982
6	farms92	number of farms, 1992
7	farms87	number of farms, 1987
8	farms82	number of farms, 1982
9	largef92	number of farms with 1,000 acres or more, 1992
10	largef87	number of farms with 1,000 acres or more, 1987
11	largef82	number of farms with 1,000 acres or more, 1982
12	smallf92	number of farms with 9 acres or fewer, 1992
13	smallf87	number of farms with 9 acres or fewer, 1987
14	smallf82	number of farms with 9 acres or fewer, 1982
15	region	region of country
		(W = West, NC = North Central, S = South, N = Northeast)

agstrat.dat   Data from a stratified random sample of size 300 from the U.S. 1992 Census of Agriculture. In columns 3-14, the value "-99" denotes missing data.

Column	Name	Value
1	county	county name
2	state	state abbreviation
3	acres92	number of acres devoted to farms, 1992
4	acres87	number of acres devoted to farms, 1987
5	acres82	number of acres devoted to farms, 1982
6	farms92	number of farms, 1992
7	farms87	number of farms, 1987
8	farms82	number of farms, 1982
9	largef92	number of farms with 1,000 acres or more, 1992
10	largef87	number of farms with 1,000 acres or more, 1987
11	largef82	number of farms with 1,000 acres or more, 1982
12	smallf92	number of farms with 9 acres or fewer, 1992
13	smallf87	number of farms with 9 acres or fewer, 1987
14	smallf82	number of farms with 9 acres or fewer, 1982
15	region	S = south; W = west; NC = north central; NE = northeast
16	rn	random numbers used to select sample in each stratum
17	weight	sampling weight for each county in sample

algebra.dat   Artificial data for a one-stage cluster of 12 algebra classes from a population of 187 algebra classes.

Column	Name	Value
1	class	class number (indicates the psus)
2	Mi	gives the class size M_i
3	score	student's score on the test about function knowledge

anthrop.dat   Length of left middle finger and height for 3000 criminals (source: Macdonell, 1901). This data set contains information for the entire population.

Column	Name	Value
1	finger	length of left middle finger (cm)
2	height	height (inches)

anthsrs.dat   Length of left middle finger and height for a SRS of size 200 from anthrop.dat.

Column	Name	Value
1	finger	length of left middle finger (cm)
2	height	height (inches)

anthuneq.dat   Length of left middle finger and height for a with-replacement unequal probability sample of size 200 from anthrop.dat. The probability of selection, ψ_i, was proportional to 24 for for $y < 65$, 12 for $y = 65$, 2 for $y = 66$ or 67, and 1 for $y > 67$.

Column	Name	Value
1	finger	length of left middle finger (cm)
2	height	height (inches)
3	prob	probability of selection
4	wt	sampling weight

artifratio.dat   Values from all possible SRSs for population in Example 4.4.

Column	Name	Value
1	sample	sample number
2	i1	first unit in sample
3	i2	second unit in sample
4	i3	third unit in sample
5	i4	fourth unit in sample
6	xbars	$\bar{x}_\mathcal{S}$
7	ybars	$\bar{y}_\mathcal{S}$
8	bhat	$\hat{B}$
9	tSRS	$\hat{t}_{y,\mathrm{SRS}} = N\bar{y}_\mathcal{S}$
10	thatr	$\hat{t}_{yr}$

auditresult.dat   Data collected for Example 6.14. Part of this table is shown in Table 6.9.

Column	Name	Value
1	account	audit unit
2	bookvalue	book value of account
3	psi	Selection probability ψ_i
4	auditvalue	audit value of account $(= y)$

auditselect.dat   Selection of accounts for audit in Example 6.14. The first few lines of this file are in Table 6.8.

Column	Name	Value
1	account	audit unit
2	bookval	book value of account
3	cumbv	cumulative book value
4	rn1	random number 1 selecting account
5	rn2	random number 2 selecting account
6	rn3	random number 3 selecting account

azcounties.dat   Population and number of housing units for Arizona counties (excluding Maricopa and Pima Counties), from 2000 census.

Column	Name	Value
1	number	County number
2	name	County name
3	population	Population in county in 2000
4	housing	Number of housing units in county in 2000

baseball.dat   Statistics on 797 baseball players, compiled by Jenifer Boshes from the rosters of all major league teams in November, 2004. Source: Forman, S. L. (2004). *Baseball-reference.com—Major league statistics and information.* Retrieved November 2004 from www.baseball-reference.com.

Column	Name	Value
1	team	team played for at beginning of the season
2	leagueID	AL or NL
3	player	a unique identifier for each baseball player
4	salary	player salary in 2004
5	POS	primary position coded as P, C, 1B, 2B, 3B, SS, RF, LF, or CF
6	G	games played
7	GS	games started
8	InnOuts	number of innings
9	PO	Put Outs
10	A	number of assists
11	E	Errors
12	DP	number of double plays
13	PB	number of passed balls (only applies to catchers)
14	GB	number of games that player appeared at bat
15	AB	number of at bats
16	R	number of runs scored
17	H	number of hits
18	SecB	number of doubles
19	ThiB	number of triples
20	HR	number of home runs
21	RBI	number of runs batted in
22	SB	number of stolen bases
23	CS	number of times caught stealing
24	BB	number of times walked
25	SO	number of strikeouts
26	IBB	number of times intentionally walked
27	HBP	number of times hit by pitch
28	SH	number of sacrifice hits
29	SF	number of sacrifice flies
30	GIDP	grounded into double play

books.dat Data from homeowner's survey to estimate total number of books, used in Exercise 8 of Chapter 5.

Column	Name	Value
1	shelf	shelf number
2	Mi	total number of books on that shelf ($= M_i$)
3	number	number of the book selected
4	purchase	purchase cost of book
5	replace	replacement cost of book

certify.dat   Data from the 1994 Survey of ASA Membership on Certification. The full data set is on Statlib (lib.stat.cmu.edu/asacert/certsurvey). For questions 1 through 5, the responses are coded: 0 = No response, 1 = Yes, 2 = Possibly, 3 = No opinion, 4 = Unlikely, 5 = No. Missing values for other questions are represented by blanks.

Column	Value
1	Should the ASA develop some form of certification?
2	Would you approve of a certification program similar to that described in the July 1993 issue of *Amstat News*?
3	Should there be specific certification programs for statistics sub-disciplines?
4	If the ASA developed a Certification program would you attempt to become certified?
5	If the ASA offered certification should recertification be required every several years?
6	Major sub-discipline BA=Bayesian , BE = Business & Economic, BI=Biometrics, BP=Biopharmaceutical, CM=Computing, EN=Environment, EP=Epidemiology, GV=Government, MR=Marketing, PE=Physical & Engineering Sciences, QP=Quality & Productivity SE=Statistical Education, SG=Statistical Graphics, SP=Sports, SR=Survey Research, SS=Social Statistics, TH=Teaching Statistics in Health Sciences, OT=Other
7	Highest collegiate degree: B=BS or BA, M=MS, N =None, P=Ph.D., O=Other
8	Employment status: E=Employed, I=In School, R=Retired, S=Self-Employed, U=Unemployed, O=Other
9	Primary work environment A=Academia, G=Government, I=Industry, O=Other
10	Primary work activity: C=Consultant, E=Educator, P=Practitioner, R=Researcher, S=Student, O=Other
11	For how many years have you been a member of the ASA?

cherry.dat   Measurements of diameter, height, and timber volume for a sample of 31 black cherry trees. Source: Hand et al. (1994).

Column	Name	Value
1	diam	Diameter of tree in inches, measured at 4.5 feet off the ground
2	height	Height of tree (feet)
3	vol	Volume of tree (cubic feet)

classpps.dat Unequal-probability without-replacement sample from population of statistics classes, used in Example 6.11.

Column	Name	Value
1	class	Class number
2	clssize	Number of students in class, M_i
3	finalweight	Sampling weight for the student
4	hours	Number of hours spent studying statistics

classppsjp.dat   Joint inclusion probabilities for psus for the sample in classpps.dat. This file is used in Exercise 6.14 to calculate the without-replacement variance.

Column	Name	Value
1	class	Class number
2	clssize	Number of students in class, M_i
3	SelectionProb	Inclusion probability π_i
4	SamplingWeight	Sampling weight for psu $= 1/\pi_i$
5–9	JtProb_1–JtProb_5	Joint inclusion probabilities matrix, giving π_{ik}

college91.dat   Four independently chosen SRSs from the 1991 Information Please
Almanac, used in Example 9.3.

Column	Name	Value
1	college	College name
2	group	Random group number
3	enrollment	Enrollment at college
4	resident	Resident tuition
5	nonresident	Nonresident tuition

coots.dat   Selected information on egg size, from a larger study by Arnold (1991). Data provided courtesy of Todd Arnold. Not all observations are used for this data set, so results may not agree with those in Arnold (1991).

Column	Name	Value
1	clutch	Clutch number from which eggs were subsampled.
2	csize	Number of eggs in clutch (M_i)
3	length	length of egg (mm)
4	breadth	maximum breadth of egg (mm)
5	volume	calculated as 0.000507*length * breadth^2
6	tmt	= 1 if received supplemental feeding, 0 otherwise

counties.dat   Data from a simple random sample of 100 of the 3141 counties in the United States (source: U.S. Bureau of the Census, 1994.) Missing values are coded as -99.

Column	Name	Value
1	RN	Random number used to select the county
2	State	
3	County	
4	landarea	land area, 1990 (sq. miles)
5	totpop	total persons, 1992
6	physician	active non-Federal physicians on Jan. 1, 1990
7	enroll	school enrollment in elementary or high school, 1990
8	percpub	percent of school enrollment in public schools
9	civlabor	civilian labor force, 1991
10	unemp	number unemployed, 1991
11	farmpop	farm population, 1990
12	numfarm	number of farms, 1987
13	farmacre	acreage in farms, 1987
14	fedgrant	total expenditures in federal funds and grants, 1992 (millions of dollars)
15	fedciv	civilians employed by federal government, 1990
16	milit	military personnel, 1990
17	veterans	number of veterans, 1990
18	percviet	percent of veterans from Vietnam era, 1990

In some cases, the data sets used in this book are a subset of the original data; in others, the information has been modified to protect the confidentiality of the respondents. They are included for instructional purposes only. Anyone wishing to investigate the subject matter further should obtain the original data from the source.

All data sets ending in .csv use commas as a separator between fields.

agpop.dat   Data from the U.S. 1992 Census of Agriculture. In columns 3-14, the value "-99" denotes missing data.

Column	Name	Value
1	county	county name
2	state	state abbreviation
3	acres92	number of acres devoted to farms, 1992
4	acres87	number of acres devoted to farms, 1987
5	acres82	number of acres devoted to farms, 1982
6	farms92	number of farms, 1992
7	farms87	number of farms, 1987
8	farms82	number of farms, 1982
9	largef92	number of farms with 1,000 acres or more, 1992
10	largef87	number of farms with 1,000 acres or more, 1987
11	largef82	number of farms with 1,000 acres or more, 1982
12	smallf92	number of farms with 9 acres or fewer, 1992
13	smallf87	number of farms with 9 acres or fewer, 1987
14	smallf82	number of farms with 9 acres or fewer, 1982
15	region	region of country (W = West, NC = North Central, S = South, N = Northeast)

agpps.dat   Data from a without-replacement pps sample from file agpop.dat.

Column	Name	Value
1	county	county name
2	state	state abbreviation
3	acres92	number of acres devoted to farms, 1992
4	acres87	number of acres devoted to farms, 1987
5	sizemeas	size measure used to select the pps sample
6	SelectionProb	inclusion probability for county, π_i
7	SamplingWeight	sampling weight for county, $w_i = 1/\pi_i$
8–22	jtprob1–jtprob15	columns of joint inclusion probabilities

agsrs.dat   Data from an SRS of size 300 from the U.S. 1992 Census of Agriculture.
In columns 3-14, the value "-99" denotes missing data.

Column	Name	Value
1	county	county name
2	state	state abbreviation
3	acres92	number of acres devoted to farms, 1992
4	acres87	number of acres devoted to farms, 1987
5	acres82	number of acres devoted to farms, 1982
6	farms92	number of farms, 1992
7	farms87	number of farms, 1987
8	farms82	number of farms, 1982
9	largef92	number of farms with 1,000 acres or more, 1992
10	largef87	number of farms with 1,000 acres or more, 1987
11	largef82	number of farms with 1,000 acres or more, 1982
12	smallf92	number of farms with 9 acres or fewer, 1992
13	smallf87	number of farms with 9 acres or fewer, 1987
14	smallf82	number of farms with 9 acres or fewer, 1982
15	region	region of country
		(W = West, NC = North Central, S = South, N = Northeast)

agstrat.dat Data from a stratified random sample of size 300 from the U.S. 1992 Census of Agriculture. In columns 3-14, the value "-99" denotes missing data.

Column	Name	Value
1	county	county name
2	state	state abbreviation
3	acres92	number of acres devoted to farms, 1992
4	acres87	number of acres devoted to farms, 1987
5	acres82	number of acres devoted to farms, 1982
6	farms92	number of farms, 1992
7	farms87	number of farms, 1987
8	farms82	number of farms, 1982
9	largef92	number of farms with 1,000 acres or more, 1992
10	largef87	number of farms with 1,000 acres or more, 1987
11	largef82	number of farms with 1,000 acres or more, 1982
12	smallf92	number of farms with 9 acres or fewer, 1992
13	smallf87	number of farms with 9 acres or fewer, 1987
14	smallf82	number of farms with 9 acres or fewer, 1982
15	region	S = south; W = west; NC = north central; NE = northeast
16	rn	random numbers used to select sample in each stratum
17	weight	sampling weight for each county in sample

algebra.dat   Artificial data for a one-stage cluster of 12 algebra classes from a population of 187 algebra classes.

Column	Name	Value
1	class	class number (indicates the psus)
2	Mi	gives the class size M_i
3	score	student's score on the test about function knowledge

anthrop.dat   Length of left middle finger and height for 3000 criminals (source: Macdonell, 1901). This data set contains information for the entire population.

Column	Name	Value
1	finger	length of left middle finger (cm)
2	height	height (inches)

anthsrs.dat Length of left middle finger and height for a SRS of size 200 from anthrop.dat.

Column	Name	Value
1	finger	length of left middle finger (cm)
2	height	height (inches)

anthuneq.dat   Length of left middle finger and height for a with-replacement unequal probability sample of size 200 from anthrop.dat. The probability of selection, ψ_i, was proportional to 24 for for $y < 65$, 12 for $y = 65$, 2 for $y = 66$ or 67, and 1 for $y > 67$.

Column	Name	Value
1	finger	length of left middle finger (cm)
2	height	height (inches)
3	prob	probability of selection
4	wt	sampling weight

artifratio.dat   Values from all possible SRSs for population in Example 4.4.

Column	Name	Value
1	sample	sample number
2	i1	first unit in sample
3	i2	second unit in sample
4	i3	third unit in sample
5	i4	fourth unit in sample
6	xbars	$\bar{x}_{\mathcal{S}}$
7	ybars	$\bar{y}_{\mathcal{S}}$
8	bhat	$\hat{B}$
9	tSRS	$\hat{t}_{y,\mathrm{SRS}} = N\bar{y}_{\mathcal{S}}$
10	thatr	$\hat{t}_{yr}$

auditresult.dat Data collected for Example 6.14. Part of this table is shown in Table 6.9.

Column	Name	Value
1	account	audit unit
2	bookvalue	book value of account
3	psi	Selection probability ψ_i
4	auditvalue	audit value of account $(= y)$

auditselect.dat   Selection of accounts for audit in Example 6.14. The first few lines of this file are in Table 6.8.

Column	Name	Value
1	account	audit unit
2	bookval	book value of account
3	cumbv	cumulative book value
4	rn1	random number 1 selecting account
5	rn2	random number 2 selecting account
6	rn3	random number 3 selecting account

azcounties.dat Population and number of housing units for Arizona counties (excluding Maricopa and Pima Counties), from 2000 census.

Column	Name	Value
1	number	County number
2	name	County name
3	population	Population in county in 2000
4	housing	Number of housing units in county in 2000

baseball.dat   Statistics on 797 baseball players, compiled by Jenifer Boshes from the rosters of all major league teams in November, 2004. Source: Forman, S. L. (2004). *Baseball-reference.com—Major league statistics and information.* Retrieved November 2004 from www.baseball-reference.com.

Column	Name	Value
1	team	team played for at beginning of the season
2	leagueID	AL or NL
3	player	a unique identifier for each baseball player
4	salary	player salary in 2004
5	POS	primary position coded as P, C, 1B, 2B, 3B, SS, RF, LF, or CF
6	G	games played
7	GS	games started
8	InnOuts	number of innings
9	PO	Put Outs
10	A	number of assists
11	E	Errors
12	DP	number of double plays
13	PB	number of passed balls (only applies to catchers)
14	GB	number of games that player appeared at bat
15	AB	number of at bats
16	R	number of runs scored
17	H	number of hits
18	SecB	number of doubles
19	ThiB	number of triples
20	HR	number of home runs
21	RBI	number of runs batted in
22	SB	number of stolen bases
23	CS	number of times caught stealing
24	BB	number of times walked
25	SO	number of strikeouts
26	IBB	number of times intentionally walked
27	HBP	number of times hit by pitch
28	SH	number of sacrifice hits
29	SF	number of sacrifice flies
30	GIDP	grounded into double play

books.dat   Data from homeowner's survey to estimate total number of books, used in Exercise 8 of Chapter 5.

Column	Name	Value
1	shelf	shelf number
2	Mi	total number of books on that shelf ($= M_i$)
3	number	number of the book selected
4	purchase	purchase cost of book
5	replace	replacement cost of book

certify.dat   Data from the 1994 Survey of ASA Membership on Certification. The full data set is on Statlib (lib.stat.cmu.edu/asacert/certsurvey). For questions 1 through 5, the responses are coded: 0 = No response, 1 = Yes, 2 = Possibly, 3 = No opinion, 4 = Unlikely, 5 = No. Missing values for other questions are represented by blanks.

Column	Value
1	Should the ASA develop some form of certification?
2	Would you approve of a certification program similar to that described in the July 1993 issue of *Amstat News*?
3	Should there be specific certification programs for statistics sub-disciplines?
4	If the ASA developed a Certification program would you attempt to become certified?
5	If the ASA offered certification should recertification be required every several years?
6	Major sub-discipline BA=Bayesian , BE = Business & Economic, BI=Biometrics, BP=Biopharmaceutical, CM=Computing, EN=Environment, EP=Epidemiology, GV=Government, MR=Marketing, PE=Physical & Engineering Sciences, QP=Quality & Productivity SE=Statistical Education, SG=Statistical Graphics, SP=Sports, SR=Survey Research, SS=Social Statistics, TH=Teaching Statistics in Health Sciences, OT=Other
7	Highest collegiate degree: B=BS or BA, M=MS, N =None, P=Ph.D., O=Other
8	Employment status: E=Employed, I=In School, R=Retired, S=Self-Employed, U=Unemployed, O=Other
9	Primary work environment A=Academia, G=Government, I=Industry, O=Other
10	Primary work activity: C=Consultant, E=Educator, P=Practitioner, R=Researcher, S=Student, O=Other
11	For how many years have you been a member of the ASA?

cherry.dat   Measurements of diameter, height, and timber volume for a sample of 31 black cherry trees. Source: Hand et al. (1994).

Column	Name	Value
1	diam	Diameter of tree in inches, measured at 4.5 feet off the ground
2	height	Height of tree (feet)
3	vol	Volume of tree (cubic feet)

classpps.dat Unequal-probability without-replacement sample from population of statistics classes, used in Example 6.11.

Column	Name	Value
1	class	Class number
2	clssize	Number of students in class, M_i
3	finalweight	Sampling weight for the student
4	hours	Number of hours spent studying statistics

classppsjp.dat   Joint inclusion probabilities for psus for the sample in classpps.dat. This file is used in Exercise 6.14 to calculate the without-replacement variance.

Column	Name	Value
1	class	Class number
2	clssize	Number of students in class, M_i
3	SelectionProb	Inclusion probability π_i
4	SamplingWeight	Sampling weight for psu $= 1/\pi_i$
5–9	JtProb_1–JtProb_5	Joint inclusion probabilities matrix, giving π_{ik}

college91.dat   Four independently chosen SRSs from the 1991 Information Please Almanac, used in Example 9.3.

Column	Name	Value
1	college	College name
2	group	Random group number
3	enrollment	Enrollment at college
4	resident	Resident tuition
5	nonresident	Nonresident tuition

coots.dat Selected information on egg size, from a larger study by Arnold (1991). Data provided courtesy of Todd Arnold. Not all observations are used for this data set, so results may not agree with those in Arnold (1991).

Column	Name	Value
1	clutch	Clutch number from which eggs were subsampled.
2	csize	Number of eggs in clutch (M_i)
3	length	length of egg (mm)
4	breadth	maximum breadth of egg (mm)
5	volume	calculated as 0.000507*length * breadth^2
6	tmt	= 1 if received supplemental feeding, 0 otherwise

counties.dat   Data from a simple random sample of 100 of the 3141 counties in the United States (source: U.S. Bureau of the Census, 1994.) Missing values are coded as -99.

Column	Name	Value
1	RN	Random number used to select the county
2	State	
3	County	
4	landarea	land area, 1990 (sq. miles)
5	totpop	total persons, 1992
6	physician	active non-Federal physicians on Jan. 1, 1990
7	enroll	school enrollment in elementary or high school, 1990
8	percpub	percent of school enrollment in public schools
9	civlabor	civilian labor force, 1991
10	unemp	number unemployed, 1991
11	farmpop	farm population, 1990
12	numfarm	number of farms, 1987
13	farmacre	acreage in farms, 1987
14	fedgrant	total expenditures in federal funds and grants, 1992 (millions of dollars)
15	fedciv	civilians employed by federal government, 1990
16	milit	military personnel, 1990
17	veterans	number of veterans, 1990
18	percviet	percent of veterans from Vietnam era, 1990

divorce.dat Data from a sample of divorce records for states in the Divorce Registration Area. (source: Vital Statistics of the United States, 1987.)

Column	Name	Value
1	state	state name
2	abbrev	state abbreviation
3	samprate	sampling rate for state
4	numrecs	number of records sampled in state
5	hsblt20	number of records in sample with husband's age < 20
6	hsb20-24	number of records with $20 \leq$ husband's age ≤ 24
7	hsb25-29	number of records with $25 \leq$ husband's age ≤ 29
8	hsb30-34	number of records with $30 \leq$ husband's age ≤ 34
9	hsb35-39	number of records with $35 \leq$ husband's age ≤ 39
10	hsb40-44	number of records with $40 \leq$ husband's age ≤ 44
11	hsb45-49	number of records with $45 \leq$ husband's age ≤ 49
12	hsbge50	number of records with husband's age ≥ 50
13	wflt20	number of records with wife's age < 20
14	wf20-24	number of records with $20 \leq$ wife's age ≤ 24
15	wf25-29	number of records with $25 \leq$ wife's age ≤ 29
16	wf30-34	number of records with $30 \leq$ wife's age ≤ 34
17	wf35-39	number of records with $35 \leq$ wife's age ≤ 39
18	wf40-44	number of records with $40 \leq$ wife's age ≤ 44
19	wf45-49	number of records with $45 \leq$ wife's age ≤ 49
20	wfge50	number of records with wife's age ≥ 50

forest.dat   Measurements from 581,012 30×30 m cells from Region 2 of the U.S. Forest Service Resource Information System, from kdd.ics.uci.edu/databases/covertype/covertype.data.html.

Column	Name	Value
1	elevation	Elevation in meters
2	Aspect	Aspect in degrees azimuth
3	Slope	Slope in degrees
4	Horiz	Horizontal Distance to nearest surface water features (meters)
5	Vert	Vertical Dist to nearest surface water features (meters)
6	HorizRoad	Horizontal Dist to nearest roadway (meters)
7	Hillshade_9am	Hillshade index at 9am, summer solstice (0 to 255 index)
8	Hillshade_Noon	Hillshade index at noon, summer solstice (0 to 255 index)
9	Hillshade_3pm	Hillshade index at 3pmm, summer solstice (0 to 255 index)
10	HorizFire	Horizontal Distance to nearest wildfire ignition points (meters)
11	Wilderness1	= 1 if Rawah Wilderness Area, 0 otherwise
12	Wilderness2	= 1 if Neota Wilderness Area, 0 otherwise
13	Wilderness3	= 1 if Comanche Peak Wilderness Area, 0 otherwise
14	Wilderness4	= 1 if Cache la Poudre Wilderness Area, 0 otherwise
15	Cover	= Cover Type
		1 – Spruce/Fir
		2 – Lodgepole Pine
		3 – Ponderosa Pine
		4 – Cottonwood/Willow
		5 – Aspen
		6 – Douglas-fir
		7 – Krummholz

golfsrs.dat   A simple random sample of 120 golf courses, taken from the population on the web site ww2.golfcourse.com.

Column	Name	Value
1	RN	random number used to select golf course for sample
2	state	state name
3	holes	number of holes
4	type	type of course: priv(ate), semi(-private), pub(lic), mili(tary), res(ort)
5	yearblt	year course was built
6	wkday18	greens fee for 18 holes during week
7	wkday9	greens fee for 9 holes during week
8	wkend18	greens fee for 18 holes on weekend
9	wkend9	greens fee for 9 holes on weekend
10	backtee	back tee yardage
11	rating	course rating
12	par	par for course
13	cart18	golf cart rental fee for 18 holes
14	cart9	golf cart rental fee for 9 holes
15	caddy	Are caddies available? (y or n)
16	pro	Is a golf pro available? (y or n)

htcdf.dat Empirical cdf of height for artificial population of 2000 persons in ht-pop.dat.

Column	Name	Value
1	height	height value (cm) ($= y$)
2	frequency	number of times value of height appears in population
3	epmf	empirical probability mass function, $f(y)$
4	ecdf	empirical cumulative distribution function value, $F(y)$

htpop.dat   Height and gender of 2000 persons in an artificial population.

Column	Name	Value
1	height	height of person, cm
2	gender	M=male, F=female

htsrs.dat   Height and gender for a SRS of 200 persons, taken from 'htpop.dat'.

Column	Name	Value
1	rn	random number used to select unit
2	height	height of person, cm
3	gender	M=male, F=female

htstrat.dat Height and gender for a stratified random sample of 160 women and 40 men, taken from 'htpop.dat'.

Column	Name	Value
1	rn	random number used to select unit
2	height	height of person, cm
3	gender	M=male, F=female

integerwt.dat   Artificial population of 2000 observations, used in Exercise 6 of Chapter 7.

Column	Name	Value
1	stratum	stratum number
2	y	y value of observation

ipums.dat Data extracted from the 1980 Census Integrated Public Use Microdata Series, from Ruggles, S., Sobek, M., Alexander, T., Fitch, C. A., Goeken, R., Hall, P. K., et al. (2004). Integrated public use microdata series: Version 3.0 [machine-readable database]. Retrieved from www.ipums/org/usa. The stratum and psu variables were constructed for use in the book exercises. Data analyses on this file do NOT give valid results for inference to the 1980 U. S. population.

Column	Name	Value
1	stratum	stratum number (1–9)
2	psu	psu number (1–90)
3	inctot	total personal income (dollars)
		Negative values are possible: Be careful if you take logs!
4	age	age, with range 15–90
5	sex	1 = Male, 2 = Female
6	race	1 = White, 2 = Black, 3 = American Indian or Alaska Native,
		4 = Asian or Pacific Islander, 5 = Other Race
7	hispanic	0 = Not Hispanic, 1 = Hispanic
8	marstat	Marital Status: 1 = Married, 2 = Separated, 3 = Divorced,
		4 = Widowed, 5 = Never married/single
9	ownershg	Ownership of housing unit: 0 = Not Applicable,
		1 = Owned or being bought, 2 = Rents
10	yrsusa	Number of years a foreign-born person has lived in the U.S.:
		0= N/A, 1= 0-5 years, 2= 6-10 years, 3= 11-15 years,
		4= 16-20 years, 5= 21+ years, 6= Missing
11	school	Is person in school?
		0 = N/A, 1 = No, not in school, 2 = Yes, in school
12	educrec	Educational Attainment: 0= N/A; 1=None or preschool
		2= Grade 1, 2, 3, or 4; 3= Grade 5, 6, 7, or 8;
		4= Grade 9; 5= Grade 10; 6= Grade 11; 7= Grade 12;
		8= 1 to 3 years of college; 9= 4+ years of college
13	labforce	In labor force? 0 = Not Applicable (N/A), 1 = No, 2 = Yes
14	occ	Occupation code: see codes in
		www.ipums.org/usa/pwork/occa.html
15	sei	Duncan socioeconomic index: a constructed measure of
		occupational status based on income level and educational
		attainment associated with each occupation.
		(if no occupation is reported, SEI = 0)
16	classwk	class of worker: 0 = Not applicable, 10= Self-employed;
		11= Employer; 12= Working on own account; 13= Self-employed,
		not incorporated; 14= Self-employed, incorporated;
		20= Works for wages; 22= Wage/salary, private;
		23= Wage/salary at non-profit; 24= Wage/salary, government;
		25= Federal govt employee; 26= Armed forces; 27= State govt
		employee; 28= Local govt employee; 29= Unpaid family worker
17	vetstat	Veteran Status
		0 = N/A, 1 = No Service; 2 = Yes; 9 = Not ascertained

journal.dat   Types of sampling used for articles in a sample of journals. Source: Jacoby, J. and Handlin, A. H. (1991). Non-probability sampling designs for litigation surveys. *Trademark Reporter*, 81, 169–179.

Note that columns 2 and 3 do not always sum to column 1; for some articles, the investigators could not determine which type of sampling was used. When working with these data, you may wish to create a fourth column, "indeterminate," which equals column1 - (column2 + column3).

Column	Name	Value
1	numemp	number of articles in 1988 that used sampling
2	prob	number of articles that used probability sampling
3	nonprob	number of articles that used non-probability sampling

measles.dat   Roberts et al. (1995) report on the results of a survey of parents whose children had not been immunized against measles during a recent campaign to immunize all children in the first five years of secondary school. The original data were unavailable; univariate and multivariate summary statistics from these artificial data, however, are consistent with those in the paper. All variables are coded as 1 for yes, 0 for no, and 9 for no answer. A parent who refused consent (variable 4) was asked why, with responses in variables 5 through 10. If a response in variables 5 through 10 was checked, it was assigned value 1; otherwise it was assigned value 0. A parent could give more than one reason for not having the child immunized.

Column	Name	Value
1	school	school attended by child
2	form	Parent received consent form
3	returnf	Parent returned consent form
4	consent	Parent gave consent for measles immunization
5	hadmeas	Child had already had measles
6	previmm	Child had been immunized against measles
7	sideeff	Parent concerned about side effects
8	gp	Parent wanted GP to give vaccine
9	noshot	Child did not want injection
10	notser	Parent thought measles not a serious illness
11	gpadv	GP advised that vaccine was not needed
12	Mitotal	Number of nonimmunized students in school i $(= M_i)$
13	mi	Sample size in school i $(= m_i)$

ncvs2000.dat Selected variables for a subset of records from persons interviewed between January and June in the 2000 National Crime Victimization Survey. Source: U.S. Department of Justice, 2006. Note: some variables are recoded from original data file and other alterations have been made. You should use the original data set (available from the Inter-university Consortium for Political and Social Research, www.umich.icpsr,edu) to study criminal victimization; these data are included only for pedagogical purposes. Missing data are indicated by a period, using the SAS convention.

The full year of data was used to draw Figures 7.22 and 7.23 in the book, so estimates from this file will not agree with those figures.

Column	Name	Value
1	age	
2	married	=1 if married, 0 if not married
3	sex	= 0 if person male, 1 if person female
4	race	1. White
		2. Black
		3. American Indian, Aleut, Native Alaskan
		4. Asian, Pacific Islander
5	hispanic	= 1 if of Hispanic origin, 0 otherwise
6	hhinc	Household income
		01. Less than $5,000
		02. $5,000 to $7,499
		03. $7,500 to $9,999
		04. $10,000 to $12,499
		05. $12,500 to $14,999
		06. $15,000 to $17,499
		07. $17,500 to $19,999
		08. $20,000 to $24,999
		09. $25,000 to $29,999
		10. $30,000 to $34,999
		11. $35,000 to $39,999
		12. $40,000 to $49,999
		13. $50,000 to $74,999
		14. $75,000 and over
7	away	= 1 if away from home at least one evening per week, 0 otherwise
8	employ	= 1 if employed in last six months, 0 otherwise
9	numinc	number of crime incident reports for person
10	violent	number of violent crime reports
11	injury	number of injuries reported by person as a result of crime
12	medtreat	number of times person received medical treatment for injury
13	medexp	amount of medical expenses resulting from crime incidents
14	robbery	number of robbery reports
15	assault	number of assault reports
16	pweight	person weight (use as weight variable for responses involving persons)
17	pstrat	pseudo-stratum (use as stratum variable)
18	ppsu	pseudo-psu (use as clustering variable)

nhanes.dat   Selected variables from the 2003–2004 NHANES data. Source: www.cdc.nchs. The data files merged to create the data set here can be read directly from the SAS transport files demo_c.xpt and bmx_c.xpt on the website. This data set is provided for pedagogical purposes only; anyone wanting to make conclusions about health variables should download and analyze the source data directly.

Column	Name	Value
1	sdmvstra	pseudo-stratum
2	sdmvpsu	pseudo-psu
3	wtmec2yr	MEC Exam weight (use as weight variable)
3	age	Age at examination (years)
4	ridageyr	Age at screening (years)
5	riagendr	= 1 if male, 2 if female
6	ridreth2	Race/ethnicity code
		1 = Non-Hispanic white
		2 = Non-Hispanic black
		3 = Mexican American
		4 = Other race, including multi-racial
		5 = Other Hispanic
7	dmdeduc	Highest level of education completed
		1 = Less than high school
		2 = High school diploma (including GED)
		3 = More than high school
		7 = Refused
		9 = Don't know
8	dmdmartl	Marital status
		1 = married
		2 = widowed
		3 = divorced
		4 = separated
		5 = never married
		6 = living with partner
8	indfminc	Annual family income (dollars)
		1 = $0 to $4,999
		2 = $ 5,000 to $ 9,999
		3 = $10,000 to $14,999
		4 = $15,000 to $19,999
		5 = $20,000 to $24,999
		6 = $25,000 to $34,999
		7 = $35,000 to $44,999
		8 = $45,000 to $54,999
		9 = $55,000 to $64,999
		10 = $65,000 to $74,999
		11 = $75,000 and Over
		12 = Over $20,000
		13 = Under $20,000
		77 = Refused
		99 = Don't know
9	bmxwt	Weight (kg)
10	bmxbmi	Body mass index (kg/m^2)
11	bmxtri	Triceps skinfold measurement (mm)
12	bmxwaist	Waist circumference (cm)
13	bmxthicr	Thigh circumference (cm)
14	bmxarml	Upper arm length (cm)

nybight.dat   Data collected in the New York Bight for June 1974 and June 1975. Source: Wilk et al. (1977). Two of the original strata were combined because of insufficient sample sizes. For variable *catchwt*, weights less than 0.5 were recorded as 0.5 kg.

Column	Name	Value
1	year	1974 or 1975
2	stratum	stratum membership, based on depth
3	catchnum	number of fish caught during trawl
4	catchwt	total weight (kg) of fish caught during trawl
5	numspp	number of species of fish caught during trawl
6	depth	depth of station (m)
7	temp	surface temperature (degrees C)

otters.dat Data on number of holts (dens) in Shetland, U.K., used in Kruuk et al. (1988). Data courtesy of Hans Kruuk.

Column	Name	Value
1	section	
2	habitat	type of habitat (stratum)
3	holts	number of dens

ozone.dat   Hourly ozone readings (ppb) from Eskdalemuir, Scotland for 1994 and 1995. Source: Air Quality Information Centre, www.aeat.co.uk.

Column	Value
1	date (day/month/year)
2	ozone reading at 1:00, GMT
3	ozone reading at 2:00, GMT
⋮	⋮
25	ozone reading at 24:00, GMT

radon.dat   Radon readings for a stratified sample of 1003 homes in Minnesota. Source: www.stat.berkeley.edu/users/statlabs/labs.html, cited in Nolan, D. and Speed, T. (2000). *Statlabs: Mathematical statistics through applications.* New York: Springer.

Column	Name	Value
1	countyum	County Number
2	countyname	County Name
3	sampsize	Sample size in County
4	popsize	Population size in County
5	radon	Radon concentration (pCi/L)

rectlength.dat   Lengths of rectangles for Exercise 38 of Chapter 6.

Column	Name	Value
1	rectangle	Rectangle number
2	length	Rectangle length

samples.dat All possible simple random samples that can be generated from the population in Example 2.2.

Column	Name	Value
1	sampnum	Sample number
2–5	$u1 - u4$	Sampled units in $\mathcal{S}$
6–9	$y1 - y4$	$\{y_i, i \in \mathcal{S}\}$
10	total	$\hat{t}_{\mathcal{S}}$

seals.dat Data on number of breathing holes found in sampled areas of Svalbard fjords, reconstructed from summary statistics given in Lydersen and Ryg (1991).

Column	Name	Value
1	Zone	zone number for sampled area
2	holes	number of breathing holes Imjak found in area

selectrs.dat   Steps used in selecting the simple random sample in Example 2.5.

Column	Value
1	Random number generated between 0 and 1
2	ceiling(3078*RN)
3	Distinct values in column 2
4	New values generated to replace duplicates
5	Set of 300 distinct values to be used in sample

shorebirds.dat   Two-phase sample of shorebird nests, used in Exercise 12.3. These are artificial data constructed from summary statistics given in
Bart, J., and Earnst, S. (2002). Double-sampling to estimate density and population trends in birds. *The Auk*, 119, 36-45.

Column	Name	Value
1	plot	Plot number
2	rapid	Rapid-method count of number of birds in plot
3	intense	Intensive-method count of number of nests in plot
		$= -9$ if the plot is not in the phase II sample

spanish.dat   Cluster sample of introductory Spanish students, used in Exercise 5.5.

Column	Name	Value
1	class	Class number
2	score	Score on vocabulary test (out of 100)
3	trip	= 1 if plan a trip to a Spanish-speaking country, 0 otherwise

srs30.dat   SRS of size 30 from artificial population of size 100.

Column	Name	Value
1	y	observation value

ssc.dat   SRS of 150 members of the Statistical Society of Canada

Column	Value
1	Sex (m or f)
2	Occupation (a = academic, g = government, i = industry, n = not determined)
3	ASA (= 1 if person is member of American Statistical Association, 0 otherwise)

statepop.dat   Unequal probability sample of counties in the United States; counties selected with probability proportional to 1992 population.

Column	Name	Value
1	state	
2	county	
3	landarea	land area of county, 1990 (square miles)
4	popn	population of county, 1992
5	phys	number of physicians, 1990
6	farmpop	farm population, 1990
7	numfarm	number of farms, 1987
8	farmacre	number of acres devoted to farming, 1987
9	veterans	number of veterans, 1990
10	percviet	percent of veterans from Vietnam era, 1990

statepps.dat   Number of counties, land area, and population for the 50 states plus the District of Columbia.

Col.	Name	Value
1	state	state name
2	counties	number of counties in state
3	cumcount	cumulative number of counties
4	landarea	land area of state, 1990 (square miles)
5	cumland	cumulative land area
6	popn	population of state, 1992
7	cumpopn	cumulative population

syc.dat Selected variables from the Survey of Youth in Custody. Source: Inter-University Consortium on Political and Social Research, NCJ-130915 (U.S. Department of Justice, 1989).

Column	Name	Value
1	stratum	stratum number
2	psu	psu (facility) number
3	psusize	number of eligible residents in psu
4	initwt	initial weight
5	finalwt	final weight
6	randgrp	random group number
7	age	age of resident (99=missing)
8	race	race of resident
		1 = white; 2 = black; 3 = Asian/Pacific Islander;
		4 = American Indian, Aleut, Eskimo; 5 = Other; 9 = Missing
9	ethnicty	1 = Hispanic, 0 = not Hispanic, 9=missing
10	educ	highest grade attended before sent to correctional institution
		00 = Never attended school; 01 - 12 = highest grade attended;
		13 = GED; 14 = Other; 99=missing
11	sex	1 = male, 2 = female, 9 = missing
12	livewith	Who did you live with most of the time you were growing up?
		1 = Mother only, 2 = Father only 3 = Both mother and father,
		4 = Grandparents, 5 = Other relatives, 6 = Friends, 7 = Foster home,
		8 = Agency or institution, 9 = Someone else, 99 = Blank
13	famtime	Has anyone in your family, such as your mother, father, brother,
		sister, ever served time in jail or prison?
		1 = Yes, 2 = No, 7 = Don't know, 9 = Blank

Column	Name	Value
14	crimtype	most serious crime in current offense
		1 = violent (e.g., murder, rape, robbery, assault)
		2 = property (e.g. burglary, larceny, arson, fraud, motor vehicle theft)
		3 = drug (drug possession or trafficking)
		4 = public order (weapons violation, perjury, failure to appear in court)
		5 = juvenile status offense (truancy, running away, incorrigible behavior)
		9 = missing
15	everviol	ever put on probation or sent to correctional inst for violent offense
		1 = yes, 0 = no
16	numarr	number of times arrested (99=missing)
17	probtn	number of times on probation (99=missing)
18	corrinst	number of times previously committed to correctional institution (99=missing)
19	evertime	Prior to being sent here did you ever serve time in a correctional institution?
		1 = yes, 2 = no, 9 = missing
20	prviol	=1 if previously arrested for violent offense
21	prprop	=1 if previously arrested for property offense
22	prdrug	=1 if previously arrested for drug offense
23	prpub	=1 if previously arrested for public order offense
24	prjuv	=1 if previously arrested for juvenile status offense
25	agefirst	age first arrested (99=missing)
26	usewepn	Did you use a weapon … for this incident?
		1 = Yes, 2 = No, 9 = Blank
27	alcuse	Did you drink alcohol at all during the year before being sent here this time?
		1 = Yes; 2 = No, didn't drink during year before;
		3 = No, don't drink at all, 9=missing
28	everdrug	Ever used illegal drugs; 0=no, 1=yes, 9=missing

teachers.dat   Selected variables from a study on elementary school teacher workload in Maricopa County, Arizona. Data courtesy of Rita Gnap (see Gnap, 1995). The psu sizes are given in file teachmi.dat. The large stratum had 245 schools; the small/medium stratum had 66 schools. Missing values are coded as -9. The study is described in Exercise 15 of Chapter 5.

Column	Name	Value
1	dist	School district size: large or med/small
2	school	school identifier
3	hrwork	number of hours required to work at school per week
4	size	class size
5	preprmin	minutes spent per week in school on preparation
6	assist	minutes per week that a teacher's aide works
		with the teacher in the classroom

teachmi.dat   Cluster sizes for data in teachers.dat.

Column	Name	Value
1	dist	School district size: large or med/small
2	school	school identifier
3	popteach	number of teachers in that school
4	ssteach	number of surveys returned from that school

teachnr.dat   Data from a follow-up study of nonrespondents from Gnap (1995).
See teachers.dat for a description.

Column	Name	Value
1	hrwork	number of hours required to work at school per week
2	size	class size
3	preprmin	minutes spent per week in school on preparation
4	assist	minutes per week that a teacher's aide works with the teacher in the classroom

uneqvar.dat   Artificial data used in Exercise 18 of Chapter 11.

Column	Name
1	x
2	y

vius.dat   Selected variables from the U.S. Vehicle Inventory and Use Survey (VIUS). The data are from www.census.gov/svsd/www/vius. Missing values are coded as blanks. This data set has 98,682 records, which may be too large for some software packages to handle; the file viusca.dat is a smaller data set, with the same columns described below, containing only trucks from California. The variable descriptions below are taken from the online VIUS Data Dictionary.

Column	Name	Value
1	stratum	stratum number (contains all 255 strata)
2	adm_state	state number
3	state	state name
4	trucktype	type of truck, used in stratification
		1 pickups
		2 minivans, other light vans, and sport utility vehicles
		3 light single-unit trucks with gross vehicle weight less than 26,000 pounds
		4 heavy single-unit trucks with gross vehicle weight greater than or equal to 2
		5 truck-tractors
5	tabtrucks	column of sampling weights
6	hb_state	home base of vehicle on July 1, 2002
7	bodytype	body type of vehicle
		01. Pickup
		02. Minivan
		03. Light van other than minivan
		04. Sport utility
		05. Armored
		06. Beverage
		07. Concrete mixer
		08. Concrete pumper
		09. Crane
		10. Curtainside
		11. Dump
		12. Flatbed, stake, platform, etc.
		13. Low boy
		14. Pole, logging, pulpwood, or pipe
		15. Service, utility
		16. Service, other
		17. Street sweeper
		18. Tank, dry bulk
		19. Tank, liquids or gases
		20. Tow/Wrecker
		21. Trash, garbage, or recycling
		22. Vacuum
		23. Van, basic enclosed
		24. Van, insulated non-refrigerated
		25. Van, insulated refrigerated
		26. Van, open top
		27. Van, step, walk-in, or multistop
		28. Van, other
		99. Other not elsewhere classified

Column	Name	Value
8	adm_modelyear	model year
		01. 2003, 2002
		02. 2001
		03. 2000
		04. 1999
		05. 1998
		06. 1997
		07. 1996
		08. 1995
		09. 1994
		10. 1993
		11. 1992
		12. 1991
		13. 1990
		14. 1989
		15. 1988
		16. 1987
		17. Pre-1987
9	vius_gvw	Gross vehicle weight based on average reported weight
		01. Less than 6,001 lbs.
		02. 6,001 to 8,500 lbs.
		03. 8,501 to 10,000 lbs.
		04. 10,001 to 14,000 lbs.
		05. 14,001 to 16,000 lbs.
		06. 16,001 to 19,500 lbs.
		07. 19,501 to 26,000 lbs.
		08. 26,001 to 33,000 lbs.
		09. 33,001 to 40,000 lbs.
		10. 40,001 to 50,000 lbs.
		11. 50,001 to 60,000 lbs.
		12. 60,001 to 80,000 lbs.
		13. 80,001 to 100,000 lbs.
		14. 100,001 to 130,000 lbs.
		15. 130,001 lbs. or more
10	miles_annl	Number of Miles Driven During 2002
11	miles_life	Number of Miles Driven Since Manufactured
12	MPG	Miles Per Gallon averaged during 2002

Column	Name	Value
13	OPCLASS	Operator Classification With Highest Percent
		1. Private
		2. Motor carrier
		3. Owner operator
		4. Rental
		5. Personal transportation
		6. Not applicable (Vehicle not in use)
14	OPCLASS_MTR	Percent of Miles Driven as a Motor Carrier
15	OPCLASS_OWN	Percent of Miles Driven as an Owner Operator
16	OPCLASS_PSL	Percent of Miles Driven for Personal Transportation
17	OPCLASS_PVT	Percent of Miles Driven as Private
		(Carry Own Goods or Internal Company Business Only)
18	OPCLASS_RNT	Percent of Miles Driven as Rental
19	TRANSMSSN	Type of Transmission
		1. Automatic
		2. Manual
		3. Semi-Automated Manual
		4. Automated Manual
20	TRIP_PRIMARY	Primary Range of Operation
21	TRIP0_50	Percent of Annual Miles Accounted for with Trips
		50 Miles or Less from the Home Base
22	TRIP051_100	Percent of Annual Miles Accounted for with Trips
		51 to 100 Miles from the Home Base
23	TRIP101_200	Percent of Annual Miles Accounted for with Trips
		101 to 200 Miles from the Home Base
24	TRIP201_500	Percent of Annual Miles Accounted for with Trips
		201 to 500 Miles from the Home Base
25	TRIP500MORE	Percent of Annual Miles Accounted for with Trips
		501 or More Miles from Home Base

Column	Name	Value
26	ADM_MAKE	Make of vehicle
		01. Chevrolet
		02. Chrysler
		03. Dodge
		04. Ford
		05. Freightliner
		06. GMC
		07. Honda
		08. International
		09. Isuzu
		10. Jeep
		11. Kenworth
		12. Mack
		13. Mazda
		14. Mitsubishi
		15. Nissan
		16. Peterbilt
		17. Plymouth
		18. Toyota
		19. Volvo
		20. White
		21. Western Star
		22. White GMC
		23. Other (domestic)
		24. Other (foreign)
27	BUSINESS	Business in which vehicle was most often used during 2002
		01. For-hire transportation or warehousing
		02. Vehicle leasing or rental
		03. Agriculture, forestry, fishing, or hunting
		04. Mining
		05. Utilities
		06. Construction
		07. Manufacturing
		08. Wholesale trade
		09. Retail trade
		10. Information services
		11. Waste management, landscaping, or administrative/support services
		12. Arts, entertainment, or recreation services
		13. Accommodation or food services
		14. Other services
		Blank. Not reported or not applicable

winter.dat Selected variables from the ASU Winter Closure Survey, taken in January 1995 (provided courtesy of the ASU Office of University Evaluation). This survey was taken to investigate the attitudes and opinions of university employees towards the closing of the university between December 25 and January 1. Missing values are coded as 9. For the yes/no questions, the responses are coded as 1 = No, 2 = Yes. The variables *treatsta* and *treatme* were coded as 1=strongly agree, 2=agree, 3=undecided, 4=disagree, 5=strongly disagree. The variables process and satbreak were coded as 1=very satisfied, 2=satisfied, 3=undecided, 4=dissatisfied, 5=very dissatisfied. Variables ownsupp through offclose were coded 1 if the person checked that the statement applied to him/her, and 2 if the statement was not checked.

Col.	Name	Value
1	class	stratum number
		1 = faculty ; 2 = classified staff 3= administrative staff;
		4 = academic professional
2	yearasu	number of years worked at ASU
		1= 1-2 years; 2=3-4 years; 3=5-9 years; 4=10-14 years;
		5 = 15 or more years
3	vacation	In the past, have you *usually* taken vacation days
		the entire period between December 25 and January 1?
4	work	Did you work on campus during Winter Break Closure?
5	havediff	Did the Winter Break Closure cause you
		any difficulty/concerns?
6	negaeffe	Did the Winter Break Closure *negatively*
		affect your work productivity?
7	ownsupp	I was unable to obtain staff support in my
		department/office
8	othersup	I was unable to obtain staff support in other
		departments/offices
9	utility	I was unable to access computers, copy machine, etc.
		in my department/office
10	environ	I was unable to endure environmental conditions,
		e.g., not properly climatized
11	uniserve	I was unable to access university services
		necessary to my work
12	workelse	I was unable to work on my assignments
		because I work in another department/office
13	offclose	I was unable to work on my assignments
		because my office was closed
14	treatsta	Compared to other departments/offices, I feel staff
		in my department/office were treated fairly
15	treatme	Compared to other people working in my
		department/office, I feel I was treated fairly
16	process	How satisfied are you with the process used to
		inform staff about Winter Break Closure?
17	satbreak	How satisfied are you with the fact that
		ASU had a Winter Break Closure this year?
18	breakaga	Would you want to have Winter Break Closure again?

wtshare.dat   Artificial data set for Exercise 22 of Chapter 6. The data set has multiple records for adults with more than one child; if adult 254 has 3 children, adult 254 is listed 3 times in the data set. Note that to obtain L_k, you need to take *numadult* +1.

Column	Name	Value
1	id	Identification number of adult in sample
2	child	= 1 if record is for a child, 0 if adult has no children
3	preschool	=1 if child is in preschool, 0 otherwise
4	numadult	number of *other* adults in population who link to that child

divorce.dat Data from a sample of divorce records for states in the Divorce Registration Area. (source: Vital Statistics of the United States, 1987.)

Column	Name	Value
1	state	state name
2	abbrev	state abbreviation
3	samprate	sampling rate for state
4	numrecs	number of records sampled in state
5	hsblt20	number of records in sample with husband's age < 20
6	hsb20-24	number of records with $20 \leq$ husband's age ≤ 24
7	hsb25-29	number of records with $25 \leq$ husband's age ≤ 29
8	hsb30-34	number of records with $30 \leq$ husband's age ≤ 34
9	hsb35-39	number of records with $35 \leq$ husband's age ≤ 39
10	hsb40-44	number of records with $40 \leq$ husband's age ≤ 44
11	hsb45-49	number of records with $45 \leq$ husband's age ≤ 49
12	hsbge50	number of records with husband's age ≥ 50
13	wflt20	number of records with wife's age < 20
14	wf20-24	number of records with $20 \leq$ wife's age ≤ 24
15	wf25-29	number of records with $25 \leq$ wife's age ≤ 29
16	wf30-34	number of records with $30 \leq$ wife's age ≤ 34
17	wf35-39	number of records with $35 \leq$ wife's age ≤ 39
18	wf40-44	number of records with $40 \leq$ wife's age ≤ 44
19	wf45-49	number of records with $45 \leq$ wife's age ≤ 49
20	wfge50	number of records with wife's age ≥ 50

forest.dat   Measurements from 581,012 30 × 30 m cells from Region 2 of the U.S. Forest Service Resource Information System, from kdd.ics.uci.edu/databases/covertype/covertype.data.html.

Column	Name	Value
1	elevation	Elevation in meters
2	Aspect	Aspect in degrees azimuth
3	Slope	Slope in degrees
4	Horiz	Horizontal Distance to nearest surface water features (meters)
5	Vert	Vertical Dist to nearest surface water features (meters)
6	HorizRoad	Horizontal Dist to nearest roadway (meters)
7	Hillshade_9am	Hillshade index at 9am, summer solstice (0 to 255 index)
8	Hillshade_Noon	Hillshade index at noon, summer solstice (0 to 255 index)
9	Hillshade_3pm	Hillshade index at 3pmm, summer solstice (0 to 255 index)
10	HorizFire	Horizontal Distance to nearest wildfire ignition points (meters)
11	Wilderness1	= 1 if Rawah Wilderness Area, 0 otherwise
12	Wilderness2	= 1 if Neota Wilderness Area, 0 otherwise
13	Wilderness3	= 1 if Comanche Peak Wilderness Area, 0 otherwise
14	Wilderness4	= 1 if Cache la Poudre Wilderness Area, 0 otherwise
15	Cover	= Cover Type
		1 – Spruce/Fir
		2 – Lodgepole Pine
		3 – Ponderosa Pine
		4 – Cottonwood/Willow
		5 – Aspen
		6 – Douglas-fir
		7 – Krummholz

golfsrs.dat   A simple random sample of 120 golf courses, taken from the population on the web site ww2.golfcourse.com.

Column	Name	Value
1	RN	random number used to select golf course for sample
2	state	state name
3	holes	number of holes
4	type	type of course: priv(ate), semi(-private), pub(lic), mili(tary), res(ort)
5	yearblt	year course was built
6	wkday18	greens fee for 18 holes during week
7	wkday9	greens fee for 9 holes during week
8	wkend18	greens fee for 18 holes on weekend
9	wkend9	greens fee for 9 holes on weekend
10	backtee	back tee yardage
11	rating	course rating
12	par	par for course
13	cart18	golf cart rental fee for 18 holes
14	cart9	golf cart rental fee for 9 holes
15	caddy	Are caddies available? (y or n)
16	pro	Is a golf pro available? (y or n)

htcdf.dat   Empirical cdf of height for artificial population of 2000 persons in ht-pop.dat.

Column	Name	Value
1	height	height value (cm) $(= y)$
2	frequency	number of times value of height appears in population
3	epmf	empirical probability mass function, $f(y)$
4	ecdf	empirical cumulative distribution function value, $F(y)$

htpop.dat   Height and gender of 2000 persons in an artificial population.

Column	Name	Value
1	height	height of person, cm
2	gender	M=male, F=female

htsrs.dat   Height and gender for a SRS of 200 persons, taken from 'htpop.dat'.

Column	Name	Value
1	rn	random number used to select unit
2	height	height of person, cm
3	gender	M=male, F=female

htstrat.dat   Height and gender for a stratified random sample of 160 women and 40 men, taken from 'htpop.dat'.

Column	Name	Value
1	rn	random number used to select unit
2	height	height of person, cm
3	gender	M=male, F=female

integerwt.dat   Artificial population of 2000 observations, used in Exercise 6 of Chapter 7.

Column	Name	Value
1	stratum	stratum number
2	y	y value of observation

ipums.dat   Data extracted from the 1980 Census Integrated Public Use Microdata Series, from Ruggles, S., Sobek, M., Alexander, T., Fitch, C. A., Goeken, R., Hall, P. K., et al. (2004). Integrated public use microdata series: Version 3.0 [machine-readable database]. Retrieved from www.ipums/org/usa. The stratum and psu variables were constructed for use in the book exercises. Data analyses on this file do NOT give valid results for inference to the 1980 U. S. population.

Column	Name	Value
1	stratum	stratum number (1–9)
2	psu	psu number (1–90)
3	inctot	total personal income (dollars)
		Negative values are possible: Be careful if you take logs!
4	age	age, with range 15–90
5	sex	1 = Male, 2 = Female
6	race	1 = White, 2 = Black, 3 = American Indian or Alaska Native,
		4 = Asian or Pacific Islander, 5 = Other Race
7	hispanic	0 = Not Hispanic, 1 = Hispanic
8	marstat	Marital Status: 1 = Married, 2 = Separated, 3 = Divorced,
		4 = Widowed, 5 = Never married/single
9	ownershg	Ownership of housing unit: 0 = Not Applicable,
		1 = Owned or being bought, 2 = Rents
10	yrsusa	Number of years a foreign-born person has lived in the U.S.:
		0= N/A, 1= 0-5 years, 2= 6-10 years, 3= 11-15 years,
		4= 16-20 years, 5= 21+ years, 6= Missing
11	school	Is person in school?
		0 = N/A, 1 = No, not in school, 2 = Yes, in school
12	educrec	Educational Attainment: 0= N/A; 1=None or preschool
		2= Grade 1, 2, 3, or 4; 3= Grade 5, 6, 7, or 8;
		4= Grade 9; 5= Grade 10; 6= Grade 11; 7= Grade 12;
		8= 1 to 3 years of college; 9= 4+ years of college
13	labforce	In labor force? 0 = Not Applicable (N/A), 1 = No, 2 = Yes
14	occ	Occupation code: see codes in
		www.ipums.org/usa/pwork/occa.html
15	sei	Duncan socioeconomic index: a constructed measure of
		occupational status based on income level and educational
		attainment associated with each occupation.
		(if no occupation is reported, SEI = 0)
16	classwk	class of worker: 0 = Not applicable, 10= Self-employed;
		11= Employer; 12= Working on own account; 13= Self-employed,
		not incorporated; 14= Self-employed, incorporated;
		20= Works for wages; 22= Wage/salary, private;
		23= Wage/salary at non-profit; 24= Wage/salary, government;
		25= Federal govt employee; 26= Armed forces; 27= State govt
		employee; 28= Local govt employee; 29= Unpaid family worker
	vetstat	Veteran Status
		0 = N/A, 1 = No Service; 2 = Yes; 9 = Not ascertained

journal.dat   Types of sampling used for articles in a sample of journals. Source: Jacoby, J. and Handlin, A. H. (1991). Non-probability sampling designs for litigation surveys. *Trademark Reporter*, 81, 169–179.

Note that columns 2 and 3 do not always sum to column 1; for some articles, the investigators could not determine which type of sampling was used. When working with these data, you may wish to create a fourth column, "indeterminate," which equals column1 - (column2 + column3).

Column	Name	Value
1	numemp	number of articles in 1988 that used sampling
2	prob	number of articles that used probability sampling
3	nonprob	number of articles that used non-probability sampling

measles.dat Roberts et al. (1995) report on the results of a survey of parents whose children had not been immunized against measles during a recent campaign to immunize all children in the first five years of secondary school. The original data were unavailable; univariate and multivariate summary statistics from these artificial data, however, are consistent with those in the paper. All variables are coded as 1 for yes, 0 for no, and 9 for no answer. A parent who refused consent (variable 4) was asked why, with responses in variables 5 through 10. If a response in variables 5 through 10 was checked, it was assigned value 1; otherwise it was assigned value 0. A parent could give more than one reason for not having the child immunized.

Column	Name	Value
1	school	school attended by child
2	form	Parent received consent form
3	returnf	Parent returned consent form
4	consent	Parent gave consent for measles immunization
5	hadmeas	Child had already had measles
6	previmm	Child had been immunized against measles
7	sideeff	Parent concerned about side effects
8	gp	Parent wanted GP to give vaccine
9	noshot	Child did not want injection
10	notser	Parent thought measles not a serious illness
11	gpadv	GP advised that vaccine was not needed
12	Mitotal	Number of nonimmunized students in school i $(= M_i)$
13	mi	Sample size in school i $(= m_i)$

ncvs2000.dat   Selected variables for a subset of records from persons interviewed between January and June in the 2000 National Crime Victimization Survey. Source: U.S. Department of Justice, 2006. Note: some variables are recoded from original data file and other alterations have been made. You should use the original data set (available from the Inter-university Consortium for Political and Social Research, www.umich.icpsr,edu) to study criminal victimization; these data are included only for pedagogical purposes. Missing data are indicated by a period, using the SAS convention.

The full year of data was used to draw Figures 7.22 and 7.23 in the book, so estimates from this file will not agree with those figures.

Column	Name	Value
1	age	
2	married	=1 if married, 0 if not married
3	sex	= 0 if person male, 1 if person female
4	race	1. White
		2. Black
		3. American Indian, Aleut, Native Alaskan
		4. Asian, Pacific Islander
5	hispanic	= 1 if of Hispanic origin, 0 otherwise
6	hhinc	Household income
		01. Less than $5,000
		02. $5,000 to $7,499
		03. $7,500 to $9,999
		04. $10,000 to $12,499
		05. $12,500 to $14,999
		06. $15,000 to $17,499
		07. $17,500 to $19,999
		08. $20,000 to $24,999
		09. $25,000 to $29,999
		10. $30,000 to $34,999
		11. $35,000 to $39,999
		12. $40,000 to $49,999
		13. $50,000 to $74,999
		14. $75,000 and over
7	away	= 1 if away from home at least one evening per week, 0 otherwise
8	employ	= 1 if employed in last six months, 0 otherwise
9	numinc	number of crime incident reports for person
10	violent	number of violent crime reports
11	injury	number of injuries reported by person as a result of crime
12	medtreat	number of times person received medical treatment for injury
13	medexp	amount of medical expenses resulting from crime incidents
14	robbery	number of robbery reports
15	assault	number of assault reports
16	pweight	person weight (use as weight variable for responses involving persons)
17	pstrat	pseudo-stratum (use as stratum variable)
18	ppsu	pseudo-psu (use as clustering variable)

nhanes.dat   Selected variables from the 2003–2004 NHANES data. Source: www.cdc.nchs. The data files merged to create the data set here can be read directly from the SAS transport files demo_c.xpt and bmx_c.xpt on the website. This data set is provided for pedagogical purposes only; anyone wanting to make conclusions about health variables should download and analyze the source data directly.

Column	Name	Value
1	sdmvstra	pseudo-stratum
2	sdmvpsu	pseudo-psu
3	wtmec2yr	MEC Exam weight (use as weight variable)
3	age	Age at examination (years)
4	ridageyr	Age at screening (years)
5	riagendr	= 1 if male, 2 if female
6	ridreth2	Race/ethnicity code
		1 = Non-Hispanic white
		2 = Non-Hispanic black
		3 = Mexican American
		4 = Other race, including multi-racial
		5 = Other Hispanic
7	dmdeduc	Highest level of education completed
		1 = Less than high school
		2 = High school diploma (including GED)
		3 = More than high school
		7 = Refused
		9 = Don't know
8	dmdmartl	Marital status
		1 = married
		2 = widowed
		3 = divorced
		4 = separated
		5 = never married
		6 = living with partner
8	indfminc	Annual family income (dollars)
		1 = $0 to $4,999
		2 = $ 5,000 to $ 9,999
		3 = $10,000 to $14,999
		4 = $15,000 to $19,999
		5 = $20,000 to $24,999
		6 = $25,000 to $34,999
		7 = $35,000 to $44,999
		8 = $45,000 to $54,999
		9 = $55,000 to $64,999
		10 = $65,000 to $74,999
		11 = $75,000 and Over
		12 = Over $20,000
		13 = Under $20,000
		77 = Refused
		99 = Don't know
9	bmxwt	Weight (kg)
10	bmxbmi	Body mass index (kg/m^2)
11	bmxtri	Triceps skinfold measurement (mm)
12	bmxwaist	Waist circumference (cm)
13	bmxthicr	Thigh circumference (cm)
14	bmxarml	Upper arm length (cm)

nybight.dat   Data collected in the New York Bight for June 1974 and June 1975. Source: Wilk et al. (1977). Two of the original strata were combined because of insufficient sample sizes. For variable *catchwt*, weights less than 0.5 were recorded as 0.5 kg.

Column	Name	Value
1	year	1974 or 1975
2	stratum	stratum membership, based on depth
3	catchnum	number of fish caught during trawl
4	catchwt	total weight (kg) of fish caught during trawl
5	numspp	number of species of fish caught during trawl
6	depth	depth of station (m)
7	temp	surface temperature (degrees C)

otters.dat   Data on number of holts (dens) in Shetland, U.K., used in Kruuk et al. (1988). Data courtesy of Hans Kruuk.

Column	Name	Value
1	section	
2	habitat	type of habitat (stratum)
3	holts	number of dens

ozone.dat   Hourly ozone readings (ppb) from Eskdalemuir, Scotland for 1994 and 1995. Source: Air Quality Information Centre, www.aeat.co.uk.

Column	Value
1	date (day/month/year)
2	ozone reading at 1:00, GMT
3	ozone reading at 2:00, GMT
⋮	⋮
25	ozone reading at 24:00, GMT

radon.dat   Radon readings for a stratified sample of 1003 homes in Minnesota. Source: www.stat.berkeley.edu/users/statlabs/labs.html, cited in Nolan, D. and Speed, T. (2000). *Statlabs: Mathematical statistics through applications.* New York: Springer.

Column	Name	Value
1	countyum	County Number
2	countyname	County Name
3	sampsize	Sample size in County
4	popsize	Population size in County
5	radon	Radon concentration (pCi/L)

rectlength.dat   Lengths of rectangles for Exercise 38 of Chapter 6.

Column	Name	Value
1	rectangle	Rectangle number
2	length	Rectangle length

samples.dat   All possible simple random samples that can be generated from the population in Example 2.2.

Column	Name	Value
1	sampnum	Sample number
2–5	$u1 - u4$	Sampled units in $\mathcal{S}$
6–9	$y1 - y4$	$\{y_i, i \in \mathcal{S}\}$
10	total	$\hat{t}_{\mathcal{S}}$

seals.dat   Data on number of breathing holes found in sampled areas of Svalbard fjords, reconstructed from summary statistics given in Lydersen and Ryg (1991).

Column	Name	Value
1	Zone	zone number for sampled area
2	holes	number of breathing holes Imjak found in area

selectrs.dat   Steps used in selecting the simple random sample in Example 2.5.

Column	Value
1	Random number generated between 0 and 1
2	ceiling(3078*RN)
3	Distinct values in column 2
4	New values generated to replace duplicates
5	Set of 300 distinct values to be used in sample

shorebirds.dat   Two-phase sample of shorebird nests, used in Exercise 12.3. These are artificial data constructed from summary statistics given in
Bart, J., and Earnst, S. (2002). Double-sampling to estimate density and population trends in birds. *The Auk*, 119, 36-45.

Column	Name	Value
1	plot	Plot number
2	rapid	Rapid-method count of number of birds in plot
3	intense	Intensive-method count of number of nests in plot
		$= -9$ if the plot is not in the phase II sample

spanish.dat   Cluster sample of introductory Spanish students, used in Exercise 5.5.

Column	Name	Value
1	class	Class number
2	score	Score on vocabulary test (out of 100)
3	trip	= 1 if plan a trip to a Spanish-speaking country, 0 otherwise

srs30.dat   SRS of size 30 from artificial population of size 100.

Column	Name	Value
1	y	observation value

ssc.dat SRS of 150 members of the Statistical Society of Canada

Column	Value
1	Sex (m or f)
2	Occupation (a = academic, g = government, i = industry, n = not determined)
3	ASA (= 1 if person is member of American Statistical Association, 0 otherwise)

statepop.dat   Unequal probability sample of counties in the United States; counties selected with probability proportional to 1992 population.

Column	Name	Value
1	state	
2	county	
3	landarea	land area of county, 1990 (square miles)
4	popn	population of county, 1992
5	phys	number of physicians, 1990
6	farmpop	farm population, 1990
7	numfarm	number of farms, 1987
8	farmacre	number of acres devoted to farming, 1987
9	veterans	number of veterans, 1990
10	percviet	percent of veterans from Vietnam era, 1990

statepps.dat   Number of counties, land area, and population for the 50 states plus the District of Columbia.

Col.	Name	Value
1	state	state name
2	counties	number of counties in state
3	cumcount	cumulative number of counties
4	landarea	land area of state, 1990 (square miles)
5	cumland	cumulative land area
6	popn	population of state, 1992
7	cumpopn	cumulative population

syc.dat Selected variables from the Survey of Youth in Custody. Source: Inter-University Consortium on Political and Social Research, NCJ-130915 (U.S. Department of Justice, 1989).

Column	Name	Value
1	stratum	stratum number
2	psu	psu (facility) number
3	psusize	number of eligible residents in psu
4	initwt	initial weight
5	finalwt	final weight
6	randgrp	random group number
7	age	age of resident (99=missing)
8	race	race of resident
		1 = white; 2 = black; 3 = Asian/Pacific Islander;
		4 = American Indian, Aleut, Eskimo; 5 = Other; 9 = Missing
9	ethnicty	1 = Hispanic, 0 = not Hispanic, 9=missing
10	educ	highest grade attended before sent to correctional institution
		00 = Never attended school; 01 - 12 = highest grade attended;
		13 = GED; 14 = Other; 99=missing
11	sex	1 = male, 2 = female, 9 = missing
12	livewith	Who did you live with most of the time you were growing up?
		1 = Mother only, 2 = Father only 3 = Both mother and father,
		4 = Grandparents, 5 = Other relatives, 6 = Friends, 7 = Foster home,
		8 = Agency or institution, 9 = Someone else, 99 = Blank
13	famtime	Has anyone in your family, such as your mother, father, brother,
		sister, ever served time in jail or prison?
		1 = Yes, 2 = No, 7 = Don't know, 9 = Blank

Column	Name	Value
14	crimtype	most serious crime in current offense
		1 = violent (e.g., murder, rape, robbery, assault)
		2 = property (e.g. burglary, larceny, arson, fraud, motor vehicle theft)
		3 = drug (drug possession or trafficking)
		4 = public order (weapons violation, perjury, failure to appear in court)
		5 = juvenile status offense (truancy, running away, incorrigible behavior)
		9 = missing
15	everviol	ever put on probation or sent to correctional inst for violent offense
		1 = yes, 0 = no
16	numarr	number of times arrested (99=missing)
17	probtn	number of times on probation (99=missing)
18	corrinst	number of times previously committed to correctional institution (99=missing)
19	evertime	Prior to being sent here did you ever serve time in a correctional institution?
		1 = yes, 2 = no, 9 = missing
20	prviol	=1 if previously arrested for violent offense
21	prprop	=1 if previously arrested for property offense
22	prdrug	=1 if previously arrested for drug offense
23	prpub	=1 if previously arrested for public order offense
24	prjuv	=1 if previously arrested for juvenile status offense
25	agefirst	age first arrested (99=missing)
26	usewepn	Did you use a weapon . . . for this incident?
		1 = Yes, 2 = No, 9 = Blank
27	alcuse	Did you drink alcohol at all during the year before being sent here this time?
		1 = Yes; 2 = No, didn't drink during year before;
		3 = No, don't drink at all, 9=missing
28	everdrug	Ever used illegal drugs; 0=no, 1=yes, 9=missing

teachers.dat   Selected variables from a study on elementary school teacher workload in Maricopa County, Arizona. Data courtesy of Rita Gnap (see Gnap, 1995). The psu sizes are given in file teachmi.dat. The large stratum had 245 schools; the small/medium stratum had 66 schools. Missing values are coded as -9. The study is described in Exercise 15 of Chapter 5.

Column	Name	Value
1	dist	School district size: large or med/small
2	school	school identifier
3	hrwork	number of hours required to work at school per week
4	size	class size
5	preprmin	minutes spent per week in school on preparation
6	assist	minutes per week that a teacher's aide works
		with the teacher in the classroom

teachmi.dat   Cluster sizes for data in teachers.dat.

Column	Name	Value
1	dist	School district size: large or med/small
2	school	school identifier
3	popteach	number of teachers in that school
4	ssteach	number of surveys returned from that school

teachnr.dat Data from a follow-up study of nonrespondents from Gnap (1995). See teachers.dat for a description.

Column	Name	Value
1	hrwork	number of hours required to work at school per week
2	size	class size
3	preprmin	minutes spent per week in school on preparation
4	assist	minutes per week that a teacher's aide works with the teacher in the classroom

uneqvar.dat   Artificial data used in Exercise 18 of Chapter 11.

Column	Name
1	x
2	y

vius.dat   Selected variables from the U.S. Vehicle Inventory and Use Survey (VIUS). The data are from www.census.gov/svsd/www/vius. Missing values are coded as blanks. This data set has 98,682 records, which may be too large for some software packages to handle; the file viusca.dat is a smaller data set, with the same columns described below, containing only trucks from California. The variable descriptions below are taken from the online VIUS Data Dictionary.

Column	Name	Value
1	stratum	stratum number (contains all 255 strata)
2	adm_state	state number
3	state	state name
4	trucktype	type of truck, used in stratification
		1 pickups
		2 minivans, other light vans, and sport utility vehicles
		3 light single-unit trucks with gross vehicle weight less than 26,000 pounds
		4 heavy single-unit trucks with gross vehicle weight greater than or equal to 2
		5 truck-tractors
5	tabtrucks	column of sampling weights
6	hb_state	home base of vehicle on July 1, 2002
7	bodytype	body type of vehicle
		01. Pickup
		02. Minivan
		03. Light van other than minivan
		04. Sport utility
		05. Armored
		06. Beverage
		07. Concrete mixer
		08. Concrete pumper
		09. Crane
		10. Curtainside
		11. Dump
		12. Flatbed, stake, platform, etc.
		13. Low boy
		14. Pole, logging, pulpwood, or pipe
		15. Service, utility
		16. Service, other
		17. Street sweeper
		18. Tank, dry bulk
		19. Tank, liquids or gases
		20. Tow/Wrecker
		21. Trash, garbage, or recycling
		22. Vacuum
		23. Van, basic enclosed
		24. Van, insulated non-refrigerated
		25. Van, insulated refrigerated
		26. Van, open top
		27. Van, step, walk-in, or multistop
		28. Van, other
		99. Other not elsewhere classified

Column	Name	Value
8	adm_modelyear	model year
		01. 2003, 2002
		02. 2001
		03. 2000
		04. 1999
		05. 1998
		06. 1997
		07. 1996
		08. 1995
		09. 1994
		10. 1993
		11. 1992
		12. 1991
		13. 1990
		14. 1989
		15. 1988
		16. 1987
		17. Pre-1987
9	vius_gvw	Gross vehicle weight based on average reported weight
		01. Less than 6,001 lbs.
		02. 6,001 to 8,500 lbs.
		03. 8,501 to 10,000 lbs.
		04. 10,001 to 14,000 lbs.
		05. 14,001 to 16,000 lbs.
		06. 16,001 to 19,500 lbs.
		07. 19,501 to 26,000 lbs.
		08. 26,001 to 33,000 lbs.
		09. 33,001 to 40,000 lbs.
		10. 40,001 to 50,000 lbs.
		11. 50,001 to 60,000 lbs.
		12. 60,001 to 80,000 lbs.
		13. 80,001 to 100,000 lbs.
		14. 100,001 to 130,000 lbs.
		15. 130,001 lbs. or more
10	miles_annl	Number of Miles Driven During 2002
11	miles_life	Number of Miles Driven Since Manufactured
12	MPG	Miles Per Gallon averaged during 2002

Column	Name	Value
13	OPCLASS	Operator Classification With Highest Percent
		1. Private
		2. Motor carrier
		3. Owner operator
		4. Rental
		5. Personal transportation
		6. Not applicable (Vehicle not in use)
14	OPCLASS_MTR	Percent of Miles Driven as a Motor Carrier
15	OPCLASS_OWN	Percent of Miles Driven as an Owner Operator
16	OPCLASS_PSL	Percent of Miles Driven for Personal Transportation
17	OPCLASS_PVT	Percent of Miles Driven as Private
		(Carry Own Goods or Internal Company Business Only)
18	OPCLASS_RNT	Percent of Miles Driven as Rental
19	TRANSMSSN	Type of Transmission
		1. Automatic
		2. Manual
		3. Semi-Automated Manual
		4. Automated Manual
20	TRIP_PRIMARY	Primary Range of Operation
21	TRIP0_50	Percent of Annual Miles Accounted for with Trips
		50 Miles or Less from the Home Base
22	TRIP051_100	Percent of Annual Miles Accounted for with Trips
		51 to 100 Miles from the Home Base
23	TRIP101_200	Percent of Annual Miles Accounted for with Trips
		101 to 200 Miles from the Home Base
24	TRIP201_500	Percent of Annual Miles Accounted for with Trips
		201 to 500 Miles from the Home Base
25	TRIP500MORE	Percent of Annual Miles Accounted for with Trips
		501 or More Miles from Home Base

Column	Name	Value
26	ADM_MAKE	Make of vehicle
		01. Chevrolet
		02. Chrysler
		03. Dodge
		04. Ford
		05. Freightliner
		06. GMC
		07. Honda
		08. International
		09. Isuzu
		10. Jeep
		11. Kenworth
		12. Mack
		13. Mazda
		14. Mitsubishi
		15. Nissan
		16. Peterbilt
		17. Plymouth
		18. Toyota
		19. Volvo
		20. White
		21. Western Star
		22. White GMC
		23. Other (domestic)
		24. Other (foreign)
27	BUSINESS	Business in which vehicle was most often used during 2002
		01. For-hire transportation or warehousing
		02. Vehicle leasing or rental
		03. Agriculture, forestry, fishing, or hunting
		04. Mining
		05. Utilities
		06. Construction
		07. Manufacturing
		08. Wholesale trade
		09. Retail trade
		10. Information services
		11. Waste management, landscaping, or administrative/support services
		12. Arts, entertainment, or recreation services
		13. Accommodation or food services
		14. Other services
		Blank. Not reported or not applicable

winter.dat Selected variables from the ASU Winter Closure Survey, taken in January 1995 (provided courtesy of the ASU Office of University Evaluation). This survey was taken to investigate the attitudes and opinions of university employees towards the closing of the university between December 25 and January 1. Missing values are coded as 9. For the yes/no questions, the responses are coded as 1 = No, 2 = Yes. The variables *treatsta* and *treatme* were coded as 1=strongly agree, 2=agree, 3=undecided, 4=disagree, 5=strongly disagree. The variables process and satbreak were coded as 1=very satisfied, 2=satisfied, 3=undecided, 4=dissatisfied, 5=very dissatisfied. Variables ownsupp through offclose were coded 1 if the person checked that the statement applied to him/her, and 2 if the statement was not checked.

Col.	Name	Value
1	class	stratum number
		1 = faculty ; 2 = classified staff 3= administrative staff;
		4 = academic professional
2	yearasu	number of years worked at ASU
		1= 1-2 years; 2=3-4 years; 3=5-9 years; 4=10-14 years;
		5 = 15 or more years
3	vacation	In the past, have you *usually* taken vacation days
		the entire period between December 25 and January 1?
4	work	Did you work on campus during Winter Break Closure?
5	havediff	Did the Winter Break Closure cause you
		any difficulty/concerns?
6	negaeffe	Did the Winter Break Closure *negatively*
		affect your work productivity?
7	ownsupp	I was unable to obtain staff support in my
		department/office
8	othersup	I was unable to obtain staff support in other
		departments/offices
9	utility	I was unable to access computers, copy machine, etc.
		in my department/office
10	environ	I was unable to endure environmental conditions,
		e.g., not properly climatized
11	uniserve	I was unable to access university services
		necessary to my work
12	workelse	I was unable to work on my assignments
		because I work in another department/office
13	offclose	I was unable to work on my assignments
		because my office was closed
14	treatsta	Compared to other departments/offices, I feel staff
		in my department/office were treated fairly
15	treatme	Compared to other people working in my
		department/office, I feel I was treated fairly
16	process	How satisfied are you with the process used to
		inform staff about Winter Break Closure?
17	satbreak	How satisfied are you with the fact that
		ASU had a Winter Break Closure this year?
18	breakaga	Would you want to have Winter Break Closure again?

wtshare.dat Artificial data set for Exercise 22 of Chapter 6. The data set has multiple records for adults with more than one child; if adult 254 has 3 children, adult 254 is listed 3 times in the data set. Note that to obtain L_k, you need to take *numadult* +1.

Column	Name	Value
1	id	Identification number of adult in sample
2	child	= 1 if record is for a child, 0 if adult has no children
3	preschool	=1 if child is in preschool, 0 otherwise
4	numadult	number of *other* adults in population who link to that child