STAT 7650 - Computational Statistics

Dataset list and descriptions

Generated on January 31, 2026.

This document indexes the datasets packaged with the STAT 7650 course materials. Descriptions and metadata are taken from the accompanying *_desc.txt files in the datasets folder.

Dataset	Chapter / where used	Variables	Data file (KB)	Description
sanfrancisco	Ch. 1; Example 1.3	Month, Day, Year, Amount	sanfrancisco.dat (27.0 KB)	Rainfall amounts (inches) for San Francisco for Novermber through March from November 1990 until March 2002. During these months, San Francisco typically receives more than 80% of its precipitation. A zero entry indicates that less than 0.01 inches of preci...
facerecognition	Ch. 2, 12; Example 2.5, Problem 12.4	match, eyediff, nosecheekdiff, variabilityratio	facerecognition.dat (37.6 KB)	These data arise from the evaluation of a human face recognition algorithm. Details of the experiment are given in the text. The variables relate to pairs of images of the same person, one of which is the probe image. The response variable is "match", with ...
flourbeetles	Ch. 2; Problem 2.6	days, beetles	flourbeetles.dat (0.1 KB)	Counts of a flour beetle (Tribolium confusum) population at various points in time (days). Beetles in all stages of development were counted, and food supply was carefully controlled.
leukemia	Ch. 2; Problem 2.3	remissiontime, censored, group	leukemia.dat (0.6 KB)	Remission times for patients suffering from acute leukemia and receiving either 6-mercaptopurine (6-MP) or a placebo. One year after the start of the study, the length (weeks) of the remission period for each patient was recorded. Some outcomes were censore...
oilspills	Ch. 2; Problem 2.5	year, spills, importexport, domestic	oilspills.dat (0.5 KB)	Records of crude oil spills of at least 1000 barrels from tankers in US waters during 1974-1999. For each year, the data indicate the number of spills, the estimated amount of oil shipped through US waters as part of US import/export operations (adjusted fo...
baseball	Ch. 3, 8; Examples 3.3, 3.5, 3.6, 3.7, 8.4; Problems 3-3.4	See Table 3.1 for more details. salary ($1000s) average = batting average obp = on base...	baseball.dat (39.3 KB)	Salaries in 1992 and 27 performance statistics for 337 baseball players (no pitchers) in 1991.
geneticmapping2	Ch. 3; Problem 3.7	loc1, loc2, loc3, loc4, loc5, loc6, loc7, loc8, loc9, loc10, loc11, loc12, loc13, loc14...	geneticmapping2.dat (24.0 KB)	400 chromosomes (rows) consisting of 30 loci (columns). The source parent for the allele at each locus is denoted with a zero or one. (These data are simulated, and the permutation of the observed loci that corresponds to the ordering used to generate the d...
geneticmapping	Ch. 3; Problems 3.5, 3.6	loc1, loc2, loc3, loc4, loc5, loc6, loc7, loc8, loc9, loc10, loc11, loc12	geneticmapping.dat (2.4 KB)	100 chromosomes (rows) consisting of 12 loci (columns). The source parent for the allele at each locus is denoted with a zero or one. (These data are simulated, and the permutation of the observed loci that corresponds to the ordering used to generate the d...
wine	Ch. 3; Problem 3.8	region, x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12, x13	wine.dat (12.8 KB)	Thirteen chemical measurements on 178 wines from 3 regions of Italy. All measurements have been standardized. The true regions have been provided to evaluate the results of clustering.
censoreddata	Ch. 4; Example 4.7	Yi, ci, obsvi (=min(Yi,ci)), deltai	censoreddata.dat (0.4 KB)	Terminology for this example is introduced in Example 4.5. Let Yi ~ Exp(lambda) i.i.d. for i=1,...,30. Let ci denote constants for i=1,...,30. The observed data are Xi=(min(Yi,ci),deltai) where deltai=1 if Yi<=ci and deltai=0 if Yi>ci. Thus deltai=0 corresp...

Dataset	Chapter / where used	Variables	Data file (KB)	Description
coin	Ch. 4; Problem 4.5	outcome	coin.dat (0.6 KB)	Sequence of 200 coin flip outcomes from the experiment described in Problem 4.5 regarding the Baum-Welch algorithm. Outcomes are heads=1, tails=0.
gearcouplings	Ch. 4; Problem 4.4	failuretime, censored	gearcouplings.dat (0.1 KB)	Failure times for 14 mining equipment gear couplings. Some observations are right-censored (censored=1). In these cases, we know only that the lifetime was at least as long as the given value.
hivrisk	Ch. 4; Problem 4.2	encounters, frequency	hivrisk.dat (0.1 KB)	Results from a survey of 1500 homosexual men. Each subject reported how many risky sexual encounters he had in the previous 30 days. The data are tabled frequency counts.
trivariatenormal	Ch. 4; Problem 4.3	x1, x2, x3	trivariatenormal.dat (0.7 KB)	Fifty points X=(x1,x2,x3) drawn from a trivariate normal distribution. Missing coordinate values are indicated with NA.
alzheimers	Ch. 5; Examples 5.1, 5.2, 5.3, 5.4, 5.5	subject, month, words	alzheimers.dat (0.9 KB)	Words recalled on five consecutive monthly tests for 22 Alzheimer's patients receiving lecithin.
coal	Ch. 6,7; Problems 6.4, 7.6	year, disasters	coal.dat (0.9 KB)	Time series of the numbers of disasters at coal mines annually between 1851 and 1962.
colorado	Ch. 6; Terrain navigation examples 6.7 - 6.8	These data have a special format. Note below: ####### ####################### ###########...	colorado.dat (166.1 KB)	Topographical data for a region of Colorado
confidenceband	Ch. 6; Terrain navigation examples 6.7 - 6.8	These data have a special format. Note below: ####### ####################### ###########...	colorado.dat (166.1 KB)	Topographical data for a region of Colorado
breastcancer	Ch. 7; Problem 7.5	recurtime, treatment, censored	breastcancer.dat (0.8 KB)	Times to recurrence for breast cancer patients receiving a hormone treatment. Women who did not have a recurrence before the end of the clinical trial have right censored data (censored=1) because their time to recurrence is known to exceed the given value ...
fursealpups	Ch. 7; Examples 7.2,7.3,7.8, Exercise 7.2	censusattempt, captures, newlycaught	fursealpups.dat (0.1 KB)	Data from a capture-recapture study conducted on the Otago Penninsula, South Island, New Zealand. Fur seal pups were marked and released during 7 census attempts in one season. The population is assumed closed. For each census attempt, the number of pups ca...
mixture	Ch. 7; Examples 7.2,7.3,7.8, Problem 7.2	y	mixture.dat (1.7 KB)	Simulated from mixture of normals, distribution given in equation (7.6).
pigment	Ch. 7; Problem 7.8	Batch, Sample, Moisture	pigment.dat (0.3 KB)	Moisture contents from 15 batches of pigment. Units: one tenth of 1%. Fifteen batches of the pigment were sampled twice independently.
image1d	Ch. 8; Problem 8.4	-	image1d.dat (0.1 KB)	Synthetic 1-dimensional binary 'image' data described in Problem 8.4. Thirty-five pixels are given in sequence.
serviceberry	Ch. 8; Example 8.6, 8.7, and 8.9	none	serviceberry.dat (7.3 KB)	The data indicate presence (1) and absence (0) of the Utah serviceberry in a region of western Colorado (west of approximately 104 degrees W longitude). We binned the GIS presence/absence information into pixels that are approximately 8 km by 8 km. This gri...
whalecatch	Ch. 8; Example 8.1	year = years for which catch data were collected C = number of whales harvested	whalecatch.dat (0.7 KB)	Number of whales harvested (caught) for 101 years for a hypothetical population of whales.
whalesurvey	Ch. 8; Example 8.1	datyears = years for which abundance data were collected Xt = survey abundance estimate...	whalesurvey.dat (0.1 KB)	Six observed population abundance estimates for a hypothetical population of whales.
ximage	Ch. 8; Problem 8.5	-	ximage.dat (2.3 KB)	Synthetic data for a 20x20 pixel grey-scale image, as described in Problem 8.5 and Figure 8.12. The value for the (i,j)th pixel is given in the (i,j)th entry in ximage.dat.

Dataset	Chapter / where used	Variables	Data file (KB)	Description
alloy	Ch. 9; Examples 9.3, 9.4, 9.5, 9.6, 9.7, 9.8	ironcontent, corrosionloss	alloy.dat (0.2 KB)	Thirteen measurements of corrosion loss (response) in copper-nickel alloys, each with a specific iron content (predictor).
cancersurvival	Ch. 9; Problem 9.5	survivaltime, disease	cancersurvival.dat (0.2 KB)	Survival times (days) for patients with terminal stomach (disease=1) and breast (disease=2) cancer. Both groups were treated with ascorbate.
earthquake	Ch. 9; Problem 9.6	-	earthquake.dat (0.4 KB)	These data are the numbers of earthquakes exceeding magnitude 7.0 for each year from 1900 to 1998. Variable label: quakes
gdp	Ch. 9; Example 9.9	year, gdpchange	gdp.dat (0.4 KB)	Average percent change in gross domestic product for 16 countries for 40 years, 1871-1910.
salmon	Ch. 9; Problem 9.4	year, recruits, spawners	salmon.dat (0.5 KB)	Recruits and spawners in a hypothetical salmon population over 40 years.
tree	Ch. 9; Example 9.11	year, basal	tree.dat (6.6 KB)	Basal area growth increments for a bristlecone pine tree for 452 years from 1532 to 1983. Note: the basal area growth column has already been detrended and standardized.
2drotation	Ch. 10; Example 10.8	x1, x2	2drotation.dat (6.1 KB)	Five hundred points (one per row) from the bivariate normal-gamma rotation described in Example 10.8.
bimodal	Ch. 10; Examples 10.1, 10.5	-	bimodal.dat (0.7 KB)	100 points drawn from a univariate bimodal mixture distribution.
infrared	Ch. 10; Problems 10.1, 10.2	RAh = hour of right ascension RAm = minute of right ascension RAs = second of right asc...	infrared.dat (46.7 KB)	Infrared emissions and other characteristics of objects beyond our galaxy. We have included some additional data not mentioned in the book, just for fun.
manifold	Ch. 10; Problem 10.5	x1, x2, x3, x4	manifold.dat (150.9 KB)	5000 points (one per row) drawn from a four-dimensional mixture distribution, with a low weighting of a density that lies nearly on a 3-dimensional manifold and a high weighting of a heavy-tailed density that fills 4-dimensional space.
whalemigration	Ch. 10; Examples 10.2, 10.3, 10.4, 10.6	-	whalemigration.dat (1.2 KB)	Sighting times for 121 bowhead whale calves passing Point Barrow, Alaska, during the 2001 spring migration. Times are expressed as hours since midnight, April 5, when the first adult whale was spotted during the visual census effort.
airblast	Ch. 11; Problems 11.6, 11.7	time, pressure	airblast.dat (3.3 KB)	Pressure differences between two sensors on a steel plate exposed to a powerful air blast. There are 161 measurements over a period of time just before and after the blast. The noise in these data is primarily due to inadequate temporal resolution.
bivariatecurve	Ch. 11, 12; Figure 11.18, Example 12.7, Problem 12.8	x1, x2	bivariatecurve.dat (2.7 KB)	Two hundred general bivariate data points (one per row). There is no designation of predictor or response, but there is a clear relationship between these two variables.
easysmooth	Ch. 11; Figure 11.1, Examples 11.1, 11.2, 11.3, 11.4, 11.5, 11.7	X, Y	easysmooth.dat (2.7 KB)	Two hundred predictor-response data points for bivariate smoothing practice. The X values are equally spaced.
marsbig	Ch. 11; Problem 11.4	radius = plantocentric radius (km) latitude = areocentric (north) latitude longitude = ...	marsbig.dat (25.6 KB)	Pressure-temperature profile of the Martian atmosphere, as measured by the Mars Global Surveyor spacefcraft in 2003 using a radio occultation technique during seven orbits.
mars	Ch. 11; Problem 11.4	radius = plantocentric radius (km) temperature (K) sigmatemperature = 1 standard deviat...	mars.dat (1.9 KB)	Temperature profile of the Martian atmosphere, as measured by the Mars Global Surveyor spacefcraft in 2003 using a radio occultation technique. Data include the altitude (measured as from the center of the planet) and atmospheric temperature (K).
memory	Ch. 11; Problem 11.4	timeperiod, retention	memory.dat (0.1 KB)	Average percentage memory retention was measured against passing time, for times lapses from one minute to one week.

Dataset	Chapter / where used	Variables	Data file (KB)	Description
toughsmooth	Ch. 11; Example 11.8	X, Y	toughsmooth.dat (2.5 KB)	Two hundred predictor-response data points used in Example 11.8. These data exhibit a strong pattern, but also severely non-constant variance. A variable-span smoother is necessary.
bodyfat	Ch. 12; Problem 12.2	fat, age, weight, height, neck, chest, abd, hip, thigh, knee, ankle, biceps, forearm, w...	bodyfat.dat (16.7 KB)	Percentage of body fat for 251 men. Body fat was estimated using an underwater weighing technique. In addition, the variables age, weight, height, and ten body circumference measurements were recorded for each subject. Variable label Additional information ...
drugabuse	Ch. 12; Example 12.2	drugfree, numtreatments, age	drugabuse.dat (4.6 KB)	The response is whether a patient remained drug free for one year (yes=1, no=0). The predictors are the patient age and the number of prior drug abuse treatments. Data are given for 575 patients.
norwaypaper	Ch. 12; Examples 12.1, 12.2	negy5, x1, x3	norwaypaper.dat (0.6 KB)	A response (negy5) and two predictors (x1 and x3) related to the manufacture of paper at a Norwegian plant. The response is a measure of imperfections, and the definitions of predictors are not known.
stream	Examples 12.4-12.6 and Problem 12.5	STREAM.ID EPA's identification number IBI.BUG Index of biotic integrity YEAR year the d...	stream.dat (91.7 KB)	These data are a subset of those collected by the Environmental Protection Agency as part of a study of 353 sites in the Mid-Atlantic Highlands region of the eastern United States from 1993 to 1998. In this case, the goal is to predict the response, an inde...

Note: file sizes are approximate. For full details (including citations and any special formatting notes), consult each dataset’s corresponding description file.