Package 'loon.data'

May 13, 2021

Type Package

Title Data Used to Illustrate 'Loon' Functionality

Version 0.1.3

Description Data used as examples in the 'loon' package.

URL https://great-northern-diver.github.io/loon.data/

Depends R (>= 3.5.0)

License GPL-2

RoxygenNote 7.1.1

Encoding UTF-8

NeedsCompilation no

Author R. Wayne Oldford [aut, cre],
Adrian Waddell [aut]

Maintainer R. Wayne Oldford <rwoldford@uwaterloo.ca>

Repository CRAN

Date/Publication 2021-05-13 20:00:02 UTC

R topics documented:

`alaska_forest` *Cooperative Alaska Forest Inventory data*

Description

Cooperative Alaska Forest Inventory data

Format

A data frame with 4848 rows and 8 variables

site A number identifying the sample plot at which measurements were taken.

mass The biomass recorded in metric tonnes per hectare (Mg/ha)

speciesCode Code identifying the ranking of the species at the site. One of { `sp01`, `sp02`, ... , `sp05` } with `sp01` identifying this record as the most prevalent species at the site when measured, `sp02` the second most prevalent, and so on.

year The year in which the site was visited and the measurements taken.

scientific The scientific name of the species measured

common The common name of the species measured.

longitude The longitude locating the site in degrees.

latitude The latitude locating the site in degrees.

Details

Data on the biomass (Mg/ha) of dominant forest species from several sample plots (sites) in Alaska, each sampled over several years. Original data taken from the Cooperative Alaska Forest Inventory website.

From website:

"The Cooperative Alaska Forest Inventory (CAFI) is a comprehensive database of boreal forest conditions and dynamics in Alaska. The CAFI consists of field-gathered information from numerous permanent sample plots distributed across interior and south-central Alaska including the Kenai Peninsula. The CAFI currently has 570 permanent sample plots on 190 sites representing a wide variety of growing conditions. New plots are being added to the inventory annually. (...) Repeated periodic inventories on CAFI permanent sample plots provide valuable long-term information for modeling of forest dynamics such as growth and yield. Periodic remeasurements can also be used to test and monitor large-scale environmental and climate change."

Author(s)

R.W. Oldford.

Source

Raw data from the Forest Service of the U.S. Department of Agriculture fs.usda.gov publication (No. 32894) website and provided by

Sol Cooperdock and Brendan Rogers, Woods Hole Research Center, 149 Woods Hole Road, Falmouth MA 02540, USA and

Scott Goetz, School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff AZ 86011, USA.

Processed and provided in current form by R.W. Oldford, Statistics and Actuarial Science, University of Waterloo, Waterloo ON N2L 3G1, Canada.

References

Malone, Thomas; Liang, Jingjing; Packee, Edmond C. 2009. Cooperative Alaska Forest Inventory. Gen. Tech. Rep. PNW-GTR-785. Portland, OR: U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station. 42 p

---

`binaryalphadigits` *Binary Alphadigits*

Description

Binary 20x16 digits of "0" through "9" and capital "A" through "Z". 39 examples of each class.

Format

A data frame with 1,404 rows each representing an image (39 examples of 36 classes) and 320 variables taking binary values 1 or 0 (black =1 and white = 0 for each pixel value of a 20x16 alpha-numeric image).

Source

Sam Roweis's data page at cs.nyu.edu.

References

From Simon Lucas' (sml@essex.ac.uk), Algoval system.

See Also

ordalphadigits

`blocks` *A set of 100 plastic "blocks"*

Description

A population of 100 "blocks" of different shapes and sizes cut from a single sheet of opaque plastic a few millimetres thick.

The blocks are of uniform thickness and density (all blocks were cut from the same opaque plastic sheet of about 5 mm thickness), but have different shapes. Each block has a number from 1 to 100 etched on one side. Ambiguous block numbers (e.g. 6 and 9) were disambiguated by placing a decimal place at the bottom of the number (e.g. 6. and 9.).

The blocks are treated as an entire population of 100 and presented in their entirety, spatially mixed, with the number side up. Students are asked to visually judge the blocks and to select 10 blocks whose average weight matches the average weight of all 100. Each student independently selects 10 block ids to serve as their judgment sample (see judgment).

The group variable identifies two strata ("A" or "B") roughly determined by the apparent area/volume of the blocks.

The data can be used to illustrate different sampling methods (e.g. simple random sampling, stratified sampling, and regression estimates based on modelling weight as a function of perimeter). Histograms of resulting estimates (over several samples) can be revealing.

Format

A data frame with 100 rows and 4 variates

id The id number etched on the block.

weight The block's weight in grams.

perimeter The perimeter length of the block in centimetres.

group Group identification for the block: A are smaller blocks, B are larger.

Author(s)

R. Jock Mackay and R. Wayne Oldford

See Also

judgment

---

bone *Relative Spinal Bone Mineral Density Data*

Description

From the web source: "Relative spinal bone mineral density measurements on 261 North American adolescents. Each value is the difference in spnbmd taken on two consecutive visits, divided by the average. The age is the average age over the two visits."

The data are a repackaging and extension of the data of the same name from the now archived (in 2020) of the 2015 'ElemStatLearn' package of Kjetil B. Halvorsen.

The principal changes here is the renaming of the variables `gender` and `spnbmd` AND in the addition of a new variable `ethnic` derived from the `bone_ext` data set.

The variables `gender` and `spnbmd` were renamed to `sex` and `rspnbmd`, respectively, to better agree with names in the larger and more complete data set `bone_ext` taken from the same study and source webpage. The variable `ethnic` was extracted from `bone_ext` by matching 'idnum' in the two datasets.

Format

A data frame with 485 rows and 5 variables

idnum Identifies the subject, and hence the repeat measurements

age Age of subject averaged over the two times measurements of `spnbmd` were taken to determine the relative change `rspnbmd`.

sex Sex of the subject. A factor with levels "female" and "male".

rspnbmd Relative spinal bone mineral density measurement. Each value is the (unitless) difference in `spnbmd`' taken on two consecutive visits, divided by the average of the two measurements.

ethnic The "ethnicity/race" of the subject. A factor with levels "Asian", "Black", "Hispanic", and "White".

The row order of the values follow their order of appearence in the source webpage.

Details

The purpose of the study was to examine ethnic and sex differences in bone mineral acquisition over time for young (aged 9-25 years) healthy Asian, black, Hispanic, and white males and females. The study recorded areal bone mineral density (BMD) in grams per square centimetre in the lumbar spine.

These data are a subset of 261 subjects taken from a convenience sample of 423 healthy young people of various "ethnicities."

The source website does not describe how this subset was chosen.

Rather than the spnbmd measurement at each visit of a subject (as with 'bone_ext'), the response rspnbmd denotes the relative change in spnbmd between visits and is calculated as the difference between the later and early visit values of spnbmd divided by the average of these two values. The 'age' variable is similarily taken to be the average of the subject's ages at the two visits.

On the subjects (Bachrach et al, 1999):

"A convenience sample of healthy youth was recruited from the community through advertisements and personal contact (21, 22). Individuals with a history of medical conditions or use of medications affecting bone mineral were excluded. Subjects were encouraged to return annually for a total of four visits or until they had reached age 26 yr. Recruitment occurred between May 1992 and February 1996; data collection ended in February 1997. The cohort at entry included 103 non-Hispanic whites, 103 Hispanics, 103 Asians, and 114 non-Hispanic blacks, aged 8.8 –25.9 yr; 230 females and 193 males were enrolled as previously reported (22). For simplicity, ethnicity and race will be used as interchangeable terms, and the groups will be referred to as white, Hispanic, Asian, and black. A total of 280 subjects completed 2 visits; 189 were studied 3 times, and 113 were evaluated 4 times. Subjects who completed fewer than 4 visits included those who refused, relocated, or reached age 26 yr during the study period; in addition, subjects who were recruited late in the study did not complete all visits because funding had terminated.

So-called "ethnicity" can be found in the data set 'bone_ext'.

See references, particularly Bachrach et al (1999), for more details.

Author(s)

R.W. Oldford

Source

Trevor Hastie's "Elements of Statistical Learning" page at Stanford.

References

Laura K. Bachrach, Trevor Hastie, May-Choo Wang, Balasubramanian Narasimhan, and Robert Marcus (1999) "Bone Mineral Acquisition in Healthy Asian, Hispanic, Black and Caucasian Youth. A Longitudinal Study", J Clin Endocrinol Metab, 84, 4702-12.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman (2009) "The Elements of Statistical Learning", 2nd Edition, Springer New York <doi:10.1007/978-0-387-84858-7>

See Also

bone_ext

bone_ext *Spinal Bone Mineral Density (Extended) Data*

Description

These are measures on spinal areal bone mineral density. The data are taken from the "large" bone density dataset on the source website (see source).

Format

A data frame with 1003 rows and 5 variables

idnum Identifies the subject, and hence the repeat measurements

ethnic The "ethnicity/race" of the subject. A factor with levels "Asian", "Black", "Hispanic", and "White".

age The age in years of the subject at the time the measurement `spnbmd` was taken.

sex Sex of the subject. A factor with levels "female" and "male".

spnbmd The spinal areal bone mineral density (BMD) measurement in grams per square centimetre.

The row order of the values follow their order of appearence in the source webpage.

Details

The purpose of the study was to examine ethnic and sex differences in bone mineral acquisition over time for young (aged 9-25 years) healthy Asian, Black, Hispanic, and White males and females. The study recorded areal bone mineral density (BMD) in grams per square centimetre in the lumbar spine.

The sample was a convenience sample of 423 healthy young people of various "ethnicities."

On the subjects (Bachrach et al, 1999):

"A convenience sample of healthy youth was recruited from the community through advertisements and personal contact. Individuals with a history of medical conditions or use of medications affecting bone mineral were excluded. Subjects were encouraged to return annually for a total of four visits or until they had reached age 26 yr. Recruitment occurred between May 1992 and February 1996; data collection ended in February 1997. The cohort at entry included 103 non-Hispanic whites, 103 Hispanics, 103 Asians, and 114 non-Hispanic blacks, aged 8.8 –25.9 yr; 230 females and 193 males were enrolled as previously reported (22). For simplicity, ethnicity and race will be used as interchangeable terms, and the groups will be referred to as white, Hispanic, Asian, and black. A total of 280 subjects completed 2 visits; 189 were studied 3 times, and 113 were evaluated 4 times. Subjects who completed fewer than 4 visits included those who refused, relocated, or reached age 26 yr during the study period; in addition, subjects who were recruited late in the study did not complete all visits because funding had terminated."

See references, particularly Bachrach et al (1999), for more details.

Author(s)

R.W. Oldford

Source

Trevor Hastie's "Elements of Statistical Learning" page at Stanford.

References

Laura K. Bachrach, Trevor Hastie, May-Choo Wang, Balasubramanian Narasimhan, and Robert Marcus (1999) "Bone Mineral Acquisition in Healthy Asian, Hispanic, Black and Caucasian Youth. A Longitudinal Study", J Clin Endocrinol Metab, 84, 4702-12.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman (2009) "The Elements of Statistical Learning", 2nd Edition, Springer New York <doi:10.1007/978-0-387-84858-7>

See Also

bone

---

covidNZ *Covid 19 Case Data New Zealand*

Description

Case data published by the New Zealand Ministry of Health downloaded from the source on May 28, 2020.

The data consist of 1,154 individuals having been confirmed cases of COVID-19. Confirmed cases are people that have had a positive laboratory test.

Format

A data frame with 1154 rows and 9 variables

Case_date The date notified of a potential case. This variable is of class "Date" in day-month-year format.

Sex The sex of the person, a factor with two levels: Male and Female

Age The age group to which the person belongs. This is a factor having levels marked in year groups: "< 1", "01 to 04", "05 to 09", "10 to 14", "15 to 19", "20 to 29", "30 to 39", "40 to 49",

District_Health_Board A character vector giving the name of the DHB or District Health Board where the case occurred.

Overseas_travel A character vector indicating whether the patient recently travelled overseas.

Last_country_visited A character vector giving the name, if known, of the last country visited by the person.

Flight_number A character vector recording the number of the flight flown from the last country.

Flight_departure_date A vector of class "Date" giving the flight's departure date from the last country visited.

Arrival_date A vector of class "Date" giving the flight's arrival date from the last country visited.

Details

From the source: (May 28, 2020: health.gov.nz) "The case definition has been amended to decouple respiratory symptoms from a history of travel. Testing is available to people with respiratory symptoms suggestive of COVID-19 infection (including the acute onset of cough with or without fever). This is regardless of travel history or known contact with a confirmed or probable case of COVID-19. Priority groups for testing have been included in the case definitions."

"Confirmed case:

"A case that has laboratory definitive evidence. Laboratory definitive evidence requires at least one of the following: * detection of SARS-CoV-2 from a clinical specimen using a validated NAAT (PCR) * detection of coronavirus from a clinical specimen using pan-coronavirus NAAT (PCR) and confirmation as SARS-CoV-2 by sequencing * significant rise in IgG antibody level to SARS-CoV-2 between paired sera (when serological testing becomes available)."

Author(s)

R.W. Oldford

Source

New Zealand government health website accessed May 28, 2020.

See Also

`igg1` `medicalRecords` `pandemic`

---

crabSpecies *Colour and Sex of Rock Crabs*

Description

A sample of 200 rock crabs (Leptograpsus variegatus) found in the southern subtropical Pacific Ocean. These are small sea crabs that grow at most to about 50 millimetres of shell width.

The data contain the sex and species of each crab as determined by the researchers (see references).

Format

A data frame with 200 rows and 2 variables

Species A two level factor distinguishing the species by colour: either "blue" or "orange"

Sex A two level factor identifying the crab's sex: either "male" or "female".

Details

The purpose of this data set is to identify which crabs in `lepto` belong to each of the four groups identified here.

There are 50 of each combination of factor levels.

Data are a subset of the `crabs` data set from the `MASS` package. Only the species and sex variables appear here and their row order here now match (row for row) the order of the rows from the physical size measurements on the purple rock crab given in `lepto`.

This separation now allows clustering methods to be explored on the data set `lepto` and compared to the classification by colour (Species) and sex given here.

Author(s)

R.W. Oldford

References

N.A. Campbell and R.J. Mahon (1974). "A multivariate study of variation in two species of rock crab of the genus Leptograpsus." Australian Journal of Zoology 22, 417-425. <doi:10.1071/ZO9740417>

Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth edition. Springer. <doi:10.1007/978-0-387-21706-2>

See Also

`lepto`

---

diabetes *Diabetes Data from Andrews and Herzberg*

Description

Data on 145 non-obese patients collected at the Stanford Clinical Reseach Center to investigate "the relationship between chemical subclinical and overt nonketotic diabetes. The three primary variables used in the analysis and presented [here] are glucose intolerance, insulin response to oral glucose, and insulin resistance."

From Andrews and Herzberg (1985) book called "Data". See reference.

Format

A data frame with 145 rows and 6 variables

RelativeWeight The relative weight of the patient.

FastingPlasmaGlucose The fasting plasma glucose level.

GlucoseArea This is a measurement of the glucose intolerance as measured by the area under the straightline connecting glucose levels determined from blood samples drawn during a three hour glucose tolerance test following an oral administration of a glucose load.

InsulinArea This is a measurement of the insulin response to oral glucose as measured by the area under the straightline connecting insulin levels, again determined from blood samples drawn during a three hour glucose tolerance test following an oral administration of a glucose load.

SSPG The steady state plasma glucose (SSPG) determined after chemical suppression of endogenous insulin secretion. This is a measure of insulin resistance.

ClinClass Clinical classification of each patient, by the contemporary (1979) medical criteria, into one of three groups: "Overt" diabetic, "Chemical" diabetic, or "Normal".

The row order of the values match the "Patient Number" given in the source table.

Details

This is a dataset from the "Data" book by Andrews and Herzberg (1985) Chapter 36, pp. 215-220 Table 36.1

A more complete description can be found there. An extract from the source follows.

The purpose of the data was to investigate

"the relationship between chemical subclinical and overt nonketotic diabetes in 145 non-obese adult subjects. The three primary variables used in the analysis and presented [here] are glucose intolerance, insulin response to oral glucose, and insulin resistance. The first two [of these] variables are the areas under the straightline connecting glucose and insulin levels, respectively, determined from blood samples drawn during a three hour glucose tolerance test following an oral administration of a glucose load. [These are variables GlucoseArea and InsulinArea, respectively.] Insulin resistance is measured by the steady state plasma glucose (SSPG) determined afterchemical suppression of endogenous insulin secretion. In addition, the relative weight and fasting plasma glucose were measured for each individual at the Stanford Clinical Reseach Center and are included [here]"

Each row of diabetes is a patient, and the row number is the "Patient Number" from Andrews and Herzberg.

Author(s)

R.W. Oldford

References

David F. Andrews and Agnes M. Herzberg (1985) "Data: A Collection of Problems from Many Fields for the Student and Research Worker", Springer, New York. <doi:10.1007/978-1-4612-5098-2>

`digits` *USPS Handwritten Digits*

Description

U.S. Postal Service handwritten digit "0" through "9".

Each is an 8-bit 16x16 grayscale image of a single digit;

1100 examples of each class.

Format

A data frame with 256 rows and 11000 variables.

Each row corresponds to an 8-bit value of one of the 256 cells contained in a 16 x 16 image.

Each column is the 16 x 16 image of a single handwritten digit. There are 1100 different handwritten images of each digit appearing in blocks of 1100 columns in the order "1", "2", "3", "4", "5", "6", "7", "8", "9", "0"

Source

Sam Roweis's data page at cs.nyu.edu.

See Also

binaryalphadigits

`elements` *The first 118 chemical elements*

Description

The first 118 elements, ordered by atomic number.

Some values of density, melting point,and boiling point are predictions rather than measurements. For those elements which do not have any stable nuclides, the mass number given is that of the longest-lived isotope (exceptions here are bismuth, thorium, protactinium, and uranium standard atomic weights are available). See the source for more detail.

Format

A data frame with 118 rows and 17 variates:

Number The atomic number. The number of protons found the nucleus of every atom of the element. It is the (positive) charge number of the element's nucleus.

Symbol One or two letter atomic symbol for the element.

Name Name of the element.

Group Identifies elements having similar chemical behaviours. For most elements, the column in the periodic table is identical to the group.

Period Period identify a collection of elements of sequential mass typically from metals to non-metals. For most elements, period identifies its row in a periodic table.

Mass Relative atomic mass or atomic weight. It is a dimensionless physical quantity defined as the ratio of the average mass of atoms of a chemical element in a given sample to the atomic mass constant. Measurements are in unified atomic mass units or Daltons. Expressed in these units, it is within 1 percent of the mass number.

Mass_number Total number of protons and neutrons in the atomic nucleus.

Density Element density in grams per cubic centimetre.

Melting_point Temperature in degrees Kelvin at which the element changes state from solid to liquid. Marked as approximate for Flerovium and Oganesson.

Boiling_point Temperature in degrees Kelvin at which the element changes state from liquid to vapour.

Specific_heat_capacity The amount of heat energy required to increase the temperature by one Kelvin degree measured in Joules per gram and degree Kelvin.

Electro_negativity The tendency of an atom to attract a shared pair of electrons (or electron density) towards itself. The higher the number the greater the attraction.

Abundance The estimated abundance of the element in the Earth's crust in millograns per kilogram. For Technetium and Francium the value is marked as approximate; for Neptunium and Plutonium the value is an upper bound.

Category Identifies where elements lie on the metal - metalloid - nonmetal categorization.

Subcategory Identifies the subcategory of the element on the metal - metalloid - nonmetal categorization.

x The geometric horizontal position in the periodic table where this element appears.

y The geometric vertical position in the periodic table where this element appears.

Author(s)

R.W. Oldford

Source

Data extracted from Wikipedia's "List of chemical elements" https://en.wikipedia.org/wiki/List_of_chemical_elements (April 17, 2020).

`faces` *Olivetti Faces*

Description

Grayscale faces 8 bit [0-255], a few images of several different people.

Format

Data frame with 400 variables (one image per variable) and 4,096 rows (the greyscale values of a 64x64 image).

Details

400 total images, 64x64 size.

From the Olivetti database at ATT.

Source

Sam Roweis's data page at cs.nyu.edu.

See Also

frey, ordfrey

---

frey *Frey Faces*

Description

1,965 images of Brendan Frey's face, taken from sequential frames of a small video. Image size: 20x28.

Format

Data frame with 1,965 variables (one image per column) and 560 rows (the 560 greyscale values of a 20x28 image).

Source

Sam Roweis's data page at cs.nyu.edu.

See Also

ordfrey, faces

---

igg1 *Human immunoglobulin G1 antibody molecule*

Description

Human immunoglobulin G1 is an antibody molecule of 10,401 atoms. The great bulk of these form residues attached to 1,556 alpha carbons (alpha refers to the first carbon that attaches to a functional group). In this protein molecule, the functional groups called residues are either amino acids or carbohydrates (sugar molecules).

On the geometry of this protein, from Padlan (1994, p. 172): "An antibody molecule is composed of three major fragments: the two Fabs, which are identical and each of which contains the light chain and the first two domains of the heavy chain, and the Fc, which contains the C-terminal constant domains of the two heavy chains. The Fabs are linked to the Fc by the hinge region, which varies in length and flexibility in the different antibody classes and isotypes. The antigen binding sites (paratopes) are located at the tips of the Fabs."

Full names for amino acid residues and group characteristics were taken from the commercial website www.tocris.com.

The data records all 1,556 alpha carbons, their residues, which of five separate chains each carbon belongs to, and the geometric location given by coordinates x, y, and z as determined by X-ray

crystallography and as available to Padlan (1994) either from the Protein Data Bank or or from original investigators at the time of publication.

From the source website: "It is a composite model built from F(ab)2 fragments from Brookhaven file 2IG2.PDB, and an Fc fragment from Brookhaven file 1FC2.PDB. Part of the hinge region and other details are theoretically modeled."

Format

A data frame with 1556 rows and 10 variates:

recordType Either 'ATOM' or 'HETATM'. Here 'ATOM' indicates an atom having a standard residue of the protein; 'HETATM' (hetero atom) indicates one either having a non-standard residue of protein, or one in a group of a different kind such as carbohydrates, substrates, ligands, solvent, or metal ions. In the igg1 molecule, these will be a part of some carbohydrate.

name Name of the alpha carbon atom.

residue The three letter abbreviated name of the residue.

chainID Chains H and I (heavy), residues 1-452 each; Chains L, and M (light), residues 501-716; Chain C (carbohydrate), residues 1-9, 10-18.

residueSequenceNum Order in which that carbon atom appears in its chain.

x, y, z Coordinates of the carbon atom in three-dimensonal space.

residueName Full name of the residue.

group A group characteristic for that residue.

Author(s)

R.W. Oldford

References

Eduardo A. Padlan (1994) "Anatomy of the Antibody Molecule", Molecular Immunology, 31, Issue 3, pp. 169 - 217.

Lisa J. Harris, Steven B. Larson, Karl W. Hasel, John Day, Aaron Greenwood and Alexander McPherson (1992) "The three-dimensional structure of an intact monoclonal antibody for canine lymphoma", Nature, 360, pp. 369-372.

See Also

`elements` `SCmolecule`

`judgment` *Judgment samples of plastic blocks*

Description

Thirty-three graduate students were shown a set of 100 "physical blocks" of different shapes and sizes, all cut from a single sheet of opaque plastic a few millimetres thick (see `blocks`).

The blocks were numbered 1 to 100 and each student was asked to choose 10 blocks whose average weight would, in their judgment, equal the average weight of all 100 blocks. The only information available to each student to help them make their choice was the visual shape and size of all 100 blocks. They had a few minutes each to make and record their choices.

The task was presented as a competition with a prize to go to the student whose sample came closest to the population average weight.

Actual block weights and other information are available in the dataset `blocks`.

Format

A data frame with 33 rows and 11 variates

studentID The last four digits of the student's ID number (leading zeros removed).

first The block id number of the first block selected by the student.

second The block id number of the second block selected by the student.

third The block id number of the third block selected by the student.

fourth The block id number of the fourth block selected by the student.

fifth The block id number of the fifth block selected by the student.

sixth The block id number of the sixth block selected by the student.

seventh The block id number of the seventh block selected by the student.

eighth The block id number of the eighth block selected by the student.

ninth The block id number of the ninth block selected by the student.

tenth The block id number of the tenth block selected by the student.

Author(s)

R.W. Oldford

See Also

`blocks`

`lepto` *Rock Crabs (Leptograpsus variegatus) data*

Description

Body measurements on a sample of 200 (purple) rock crabs (Leptograpsus variegatus) found in the southern subtropical Pacific Ocean. These are small sea crabs that grow at most to about 50 millimetres of shell width.

The objective is to determine whether the sample actually represents two different species of rock crab base only on their physical measurements. See reference for details.

Format

A data frame with 200 rows and 5 variables

front The length in millimetres of the crab's carapace frontal lobe region just before the "frontal tubercles".

rear The length in millimetres of the crab's carapace rear.

length This is a measurement (in mm) of the length of the crab's carapace from front to rear along its centre line.

width This is a measurement (in mm) of the width of the carapace of the crab at its widest point.

depth This is a measurement (in mm) of the depth of the crab body.

There are known to be four distinguishable subsets in this data set.

Details

These are physical measurements on the purple rock crab (Leptograpsus variegatus), a large eyed crab found in the southern subtropical Pacific ocean. These are small sea crabs that grow at most to about 50 millimetres of shell width.

Data are a subset of the `crabs` data set from the `MASS` package. Only the five size measurements or the crabs are given here and their order has been randomized.

A good data set for cluster analysis

Author(s)

R.W. Oldford

References

N.A. Campbell and R.J. Mahon (1974). "A multivariate study of variation in two species of rock crab of the genus Leptograpsus." Australian Journal of Zoology 22, 417-425. <doi:10.1071/ZO9740417>

Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth edition. Springer. <doi:10.1007/978-0-387-21706-2>

See Also

crabSpecies

---

lightspeeds *Historical Determinations of the Speed of Light*

Description

Contains historical determinations of the speed of light from many studies from Fizeau's toothed wheel in 1849, to determinations using stabilized lasers in 1983. Methods, year of study, authors, information on mechanisms used, as well as other remarks are recorded. The estimated speed of light is recorded for each study as well as the authors' determination of the error of their measurement.

The data are of particular value since in 1974 the speed of light was defined to be 299,792.458 kilometres per second (in vacuo). The data therefore provide a rare case where the 'true value' is known.

Also, the values might be grouped by the different methods were used over time to estimate the speed of light. In this way, the data provide a useful case study to discuss methods of meta-analysis as well.

Format

A data frame with 81 rows and 11 variables

method The general method used to determine the speed of light (see details for more on this).

year Year in which the determination was made.

researcher_1 First researcher named as conducting the study (surname, or prefixed with initials if surname not unique).

researcher_2 Second researcher named, if any.

researcher_3 Third researcher named, if any.

researcher_4 Fourth researcher named, if any.

mechanism Type of mechanism used (see details below).

mechanism_2 More detail on mechanism used (see details below).

remark Remark on some detail of the study or method used.

speed The determined speed of light in air in kilometres per second.

error The error of the estimate (in km/second) as reported by the researchers (see details below).

The row order of the values follow their order of appearence in the paper given as reference below.

Details

See reference for details.

On the meaning of the error variable, from the reference: "This error is rarely a standard deviation. Nor is it based solely on measurements taken in the study. Instead, it is a number out together bt the researchers from a number of possible sources and is very subjective. It is not uncommon for subsequent researchers to examine in detail the results of a given study and to arrive at a different value of the error. Finally, physicists are accustomed to reporting the probable error which can be interpreted as approximately 0.6745 times the standard deviation of the estimate."

On the meaning of variables related to method, also from the reference:

Optical: These methods are based on having a light beam leave a source, strike a rotating mirror or pass through the spaces of a toothed wheel, travel some considerable distance to be reflected back (again striking the rotating mirror or passing through the spaces of a toothed wheel) to near the original soource. The speed of rotation must be just right and is used in the determination of the speed of light. Precision of estimation could ve increased by increasing the distance the light had to travel (from source to stationary mirror and back) or by increasing the speed of rotation.

Electrical: This method is introduced after the electromagnetic theory of light was developed. Light could now be thought of as electromagnetic radiation. As such measures of the speed of any electromagnetic radiation in vacuo would also be legitimate measures of the speed of light (in vacuo). Moreover the ratio of electrostatic to electromagnetic units of measurement of electrical quantities could be taken to be measurements of the speed of light.

Electro-optical: These are based largely on the same principle as the toothed wheel in optical methods which effectively use a mechanical shutter to switch light on and off. With the electro-optical methods, non-mechanical shutters are used to much more rapidly (and in a more finely controlled way) alternate the light and so increase the precision. The Kerr cell consists of two electrodes immersed in a liquid like nitrobenzene. When high voltage is applied to the electrodes, the polarity of light passing through the cell changes from planar to elliptical. Switching between high and low voltage effects the shutter. The quartz modulator passes sound waves through a crystal to change its refractive index. An acoustic frequency can be found to produce a diffraction grating for light passing through the crystal; double that frequency and the diffraction grating dissappears. Switching between the two frequencies produces the shutter effect.

Radio wave: Understanding light to be electromagetic radiation also means that radio waves can be used in place of visible light to make measurements of its speed. Radar sends a short pulse of high frequency radio waves from a source to a distant object and measures the time taken to receive the reflected wave from the distant object. Knowing the actual distance allows a determination of the speed of the radio wave (or light). The cavity resonator sends high frequency radio waves down a hollow cylinder sealed at both ends. The cylinder resonates if its length is a whole number multiple of the half-wavelength of the radio wave. The speed of light (in the medium within the cylinder or cavity) can be determined from the dimensions of the cylinder (cavity).

Geodetic: These are improvements on the Kerr cell technology to make it capable of measuring geodetic distances. The resulting (commercial) instrument was called a Geodimeter. With known distances these instruments could be turned around to be used to provide measures of the speed of light. The Tellurometer was another instrument invented to determine geodetic distances. The principal difference between it and the Geodimeter is that it used microwave radiation to carry the signal.

Spectroscopy: Bombarding molecules with electromagnetic radiation causes them to absorb enough energy to change various states. Bombarding molecules with microwave radiation

changes their rotational state, with infra-red radiation their vibrational state. Quantum theory allows the measurement of these to changed states to be turned into a determination of the speed of light. Different studies bombarded different molecules.

`Ultrasonic modulation`: This method can be regarded as an improvement on the `quartz modulator`. Instead of acoustic waves on a crystal, a diffraction grating is produced with ultrasonic waves in a liquid. Turning the diffraction grating on and off produces the shutter.

`Interferometry`: A single source light bean is split in two. Each travels some distance, is reflected, and returns to the source. The two are made to travel different distances and the amount by which they are out of phase with one another upon return, together with the difference in distances travelled, can be turned into a measure of the speed of light. Instead of visible light, `radio waves` or `micro waves` were used.

`Stabilized lasers`: In `interferometry` there can be some uncertainty in the measure of the wavelengths used. With stabilized lasers using a technique called `sub-Doppler saturated absorption spectroscopy` it became possible to fix the frequency (and hence the wavelength) of some lasers within a very narrow range of the electromagnetic spectrum. Such lasers are called stabilized lasers and have nice short `wavelengths` (micrometres) that allow more precise measurements of the speed of light.

Author(s)

R.W. Oldford

Source

https://www.researchgate.net/publication/275521939_The_speed_of_light_A_case_study_in_empirical_problem_solving

References

R.W. Oldford 1994, 'The speed of light: A case study in empirical problem solving', Unpublished manuscript. <doi:10.13140>

See Also

`michelson_1879`

---

lizards *A fictional data set on lizards perch choices*

Description

An entirely artificially constructed data set and context designed for classroom discussion and analysis.

The fictional context is that 384 different lizards have been observed in nature, from each of 4 different species ("A", "B", "C", or "D"). In addition to their species, the sex of each lizard was identified. At the time of observation, each lizard was perched on the branch of a bush. The perch

height from the ground in feet was measured and the diameter in inches of the branch where the lizard was perched was measured. Lizards are collected and numbered (row.names) in the order they appear in the data set.

We could imagine the study having proceeded as follows.

Problem:

Suppose that we are interested in the habitat of lizards on some other remote island archipelago. Here there are trees of all heights and many different species of lizards. Interest lies in understanding the perch height preference for all lizards.

Plan:

As with the source study, lizards will be observed during daylight hours for two separate summer time expeditions. The height (in feet) at which the lizard is observed will be recorded as is the diameter (in inches) of the branch where the lizard is perched. The species and sex of each lizard will be recorded.

Data: A total of 384 lizards of several species were captured and measured together with the height and diameter of their perch when first spotted.

NOTE: The data are inspired by the reference but are otherwise entirely fictional and constructed primarily for pedagogical purposes. Instructors might choose to invent their own context.

Format

A data frame with 384 rows and 4 variates

perchHeight The height in feet where the lizard was found to be perched.

perchDiameter The diameter in inches of the branch where the lizard was perched.

sex The sex of the lizard: a factor with levels "male" and "female"

species The species of lizard: a factor with four levels "A", "B", "C", and "D".

‘rownames(lizards)‘ labels the order in which the lizards were collected.

Author(s)

R.W. Oldford

References

T.W. Schoener (1968) “The Anolis lizards of Bimini: Resource partitioning in a complex fauna”, Ecology, Vol. 49, pp. 704-726.

`medicalRecords` *Mining medical records (fictional)*

Description

An entirely artificially constructed data set and context designed for classroom discussion and analysis.

A medical data mining context is given in detail below. In light of the context, interesting scientific questions will arise as to the data collection, and how the results should, or should not, be interpreted. It should also raise questions on what might be done in any follow up studies.

Instructors might choose to invent their own context.

Format

A data frame with 16 rows and 5 variables (providing the counts for a 2x2x2x2 contingency table).

Age A two level factor recording one of two age groups: "20-39" or "40-59".

Sex A two level factor recording sex: "Male" or "Female".

Treatment A two level factor recording the treatment received: "A" or "B".

Outcome A two level factor recording patient outcome after treatment: "Recovered" or "Died".

Freq The frequency count of patients having that combination of factors.

Details

One fictional context (constructed in March 2020) for this data set is given below (in the PPDAC style of Mackay and Oldford (2000)).

Problem:

A disease epidemic has broken out in the population of some country. It is thought that adults under the age of 60 appear to be particularly vulnerable. Both men and women contract the disease and need to be treated. Those who go untreated die within 5 days of contracting the disease.

The medical community has tried two quite different approaches to treat patients having the disease – call these 'Treatment A' and 'Treatment B'. For the health of the country, it is important to determine which of these two treatments is more effective.

Plan:

To investigate which is the better treatment, it was decided to mine the medical records from another country of those who had contracted the disease and had been treated with one of the two treatments. Patients treated with either A or B survive the disease and recover fully; some however still die.

Electronic medical records available from several of the more populous districts are accessible. These can be searched to provide records from patients that have received treatment. It was decided that there should be the same number of records drawn for each treatment.

Moreover, concern was raised that the investigation have gender balance (i.e. equal numbers of males and females). So, to make sure that both sexes were equally represented, it was also decided that the number of female patients would be the same as the number of male patients.

Finally, it was desirable to detect even small differences in success rates of the two treatments since small differences could mean many more lives being saved. A sample size of about n = 3,000 was decided on.

Records would be collected until 3,000 were found, 1500 of which were treated with 'A', 1500 with 'B', and there were equal numbers of males and females in the study.

Data:

In this stage, the plan is executed. Instead of 1500 records of treatment 'A' and 'B', 1600 of each were found. The number of males and females was kept equal (now 1600 of each sex).

The process was to search the records in order, selecting those first encountered to get 1600 for each treatment and 1600 of each sex. Many records might be discarded whenever one quota was met and the search continued to meet the other quotas. It was also noticed that the patient's age was available for each record, so that the effect of treatment on younger and older adults might also be considered.

The counts which fell into the various categories were assembled into the data presented here.

Author(s)

R.W. Oldford

References

R.J. MacKay and R.W. Oldford 2000, 'Scientific Method, Statistical Method, and the Speed of Light', Statistical Science, Volume 15, No. 3, pp. 254-278. <doi:10.1214/ss/1009212817>

See Also

pandemic covidNZ

---

michelson_1879 *A.A. Michelson's 1879 Determinations of the Speed of Light*

Description

In 1879, Albert Abraham Michelson conducted an experimental study to determine the speed of light using a rotating mirror apparatus at the U.S. Naval Academy in Annapolis, Maryland in 1879.

Details on the apparatus, the optical theory, and the conduct of the experiment are given in the reference. An abbreviated summary of these follows the variable descriptions.

Format

A data frame with 100 rows and 15 variables

Speed The determined speed of light in air in kilometres per second.

Beat Number of beats per second between tuning forks.

Correction Correction for temperature to a standard fork in beats per second.

Day Day of experiment in progress (June 5 is day 1) on which these measurements were taken.

Difference Difference betweeen the greatest and least values of revolutions.

Quality Subjective measure of the quality of the image 'I'; the more distinct was the image the higher the quality (1 = poor, 3 = good).

Displacement Displacement of image 'I' from slit 'S' in divisions of the micrometer.

Image Micrometer position of the deflected image.

Radius Radius of measurement in feet.

Revolutions Number of times the mirror revolved per second.

Screw Measure of one screw turn in millimetres.

Slit Micrometer position of the slit providing the light source 'S'.

Temperature Air temperature measured in degrees Fahrenheit.

Time.of.day Time of day at which the observation was recorded. 'AM' means one hour after sunrise and 'PM' one hour before sunset.

Remarks Unusual remarks recorded for that observation.

Details

The experiment is conducted within a closed and darkened small building at the U.S. Naval Academy. Light enters the building from one corner passing through a slit 'S' whose location is precisely determined using a micrometer.

The light then proceeds to hit a rotating mirror at the other end of the building's interior from whence it is reflected out of the building through an opening in a corner different from that of the source.

The light beam travels outside to strike another (stationary) mirror which reflects it back into the building through the same corner it exited whereupon it then strikes the rotating mirror.

Depending on the position of the rotating mirror, the returning light will be reflected off it to land at some position 'I' near the original source given by the slit 'S'.

The speed of the rotating mirror is controlled using an adjustable pump to blow air across a surface to rotate it. If the speed of rotation is just right, a crisp image 'I' of the reflected slit will appear near the original source 'S'. The speed is adjusted until this is the case.

The speed of rotation is determined using an electric tuning fork connected to the rotating mirror and whose frequency was measured by comparing it to a second standard tunig fork of known frequency. The electronic fork frequency was compared to the standard fork by determining the number of beats per second difference the two (by counting over 60 seconds).

With a speed of revolution and the displacement measured between 'S' and its returned image 'I', a measurement of the speed of light could be had.

See reference for more details.

Author(s)

R.W. Oldford

References

R.J. MacKay and R.W. Oldford 2000, 'Scientific Method, Statistical Method, and the Speed of Light', Statistical Science, Volume 15, No. 3, pp. 254-278. <doi:10.1214/ss/1009212817>

See Also

lightspeeds

`minority` *Canadian Visible Minority Data 2006*

Description

Population census count of various named visible minority groups in each of 33 major census metropolitan areas of Canada in 2006.

These data are from the 2006 Canadian census, publicly available from Statistics Canada.

Format

A data frame with 33 rows and 18 variates

Arab Number identifying as 'Arab'.

Black Number identifying as 'Black'.

Chinese Number identifying as 'Chinese'.

Filipino Number identifying as 'Filipino'.

Japanese Number identifying as 'Japanese'.

Korean Number identifying as 'Korean'.

Latin.American Number identifying as 'Latin American'.

Multiple.visible.minority Number identifying as being a member of more than one visible minority.

South.Asian Number identifying as 'South Asian'.

Southeast.Asian Number identifying as 'Southeast Asian'.

Total.population Total population of the metropolitan census area.

Visible.minority.not.included.elsewhere Number identifying as a member of a visible minority that was not included elsewhere.

Visible.minority.population Total number identifying as a member of some visible minority.

West.Asian Number identifying as 'West Asian'.

lat, long Latitude and longitude in degrees of the metropolitan census area.

googleLat, googleLong Latitude and longitude in degrees determined using the Google Maps Geocoding API.

'rownames(minority)' are the names of the metropolitan areas or cities.

Author(s)

R.W. Oldford

ordalphadigits *Binary Alphadigits Isomap*

Description

Binary Alphadigits Isomap

Format

Object of class 'isomap'.

Details

Dissimilarity object of class 'isomap'. Returned from:

`isomap(vegdist(binaryalphadigits),k=6)`.

Introduced simply to cache the results of this step so as to speed up demos.

See Also

`binaryalphadigits`

ordfrey *Frey Faces Isomap*

Description

Frey Faces Isomap

Format

Object of class 'isomap'.

Details

Dissimilarity object of class 'isomap' for Frey Faces data, created with:

`isomap(vegdist(t(frey),method="euclidean"),k = 12,ndim=6,fragmentedOK = TRUE)`

Introduced simply to cache the results of this step so as to speed up demos.

See Also

`frey`

`pandemic` *Fictional pandemic data*

Description

An entirely artificially constructed data set and context designed for classroom discussion and analysis.

The data (and the fictional narrative below) are identical to those of `trtPan`; the only difference is the organization of the data. Which organization the instructor might choose depends upon the modelling and/or data manipulation is intended for analysis.

Should raise interesting scientific questions on how the results should, or should not, be interpreted. It should also raise questions on what might be done next.

A "pandemic" context is given in the details since the data were created during the first week of March, 2020.

Instructors might choose to invent their own context.

Format

A data frame with 100 rows and 4 variables

City City for which the outcome data were recorded.

A The percent survival rate for infected persons given medical treatment "A".

B The percent survival rate for infected persons given medical treatment "B".

C The percent survival rate for infected persons given medical treatment "C".

Details

One fictional narrative for this data set is as follows.

A virulent virus has led to a world wide pandemic and that the case fatality rate (proportion of those infected who die) is huge (say 6

Suppose that through a concerted and collaborative effort of health scientists worldwide, three different treatments have been developed for this group. All three treatments have been used at one time or another on numerous patients in this group from 100 different cities worldwide. The data are observational, in that they were simply collected and the treatment given noted. No information is available on why one treatment or another was given in any particular instance.

The recovery rates (as a percent) for the patients treated by each of the three treatments were simply recorded for each of the hundred cities and are available for analysis as the data frame `pandemic`.

Some obvious questions of interest are the comparisons of treatments. For example, is treatment A better than B? Than C? Is B better than C?

Author(s)

R.W. Oldford

See Also

trtPan medicalRecords covidNZ

---

pkg_data *Simple summary of data available in named packages*

Description

Simple summary of data available in named packages

Usage

```
pkg_data(package = NULL)
```

Arguments

package	A character vector giving the package(s) to look in for data sets, or NULL. By default, all packages in the search path are used, then the 'data' subdirectory (if present) of the current working directory.

Details

Syntactic sugar wrapping call to `utils::data(package = package)` to return basic information on datasets in package. No data are loaded by the call.

See Also

data

Examples

```
head(pkg_data("loon.data"))
```

---

placenamesCanada *Canadian place names and their geo-locations.*

Description

The names (in English and in French) of 10,776 places in Canada together with their geographic locations as compiled by the Government of Canada.

The data are part of the Canadian government's open data project.

Format

A data frame with 10,676 rows and 6 variables

ID The identification number of the place name (called "PNuid_NLidu" in the source).

Name A character vector containing the place name in English.

Nom A character vector containing the place name in French.

Province A factor with 13 levels giving the two-letter code for the Canadian province or territory.

Latitude A numeric vector giving the latitude of the place.

Longitude A numeric vector giving the longitude of the place.

Details

Note that English and French names rarely differ in this data set.

Details from the source:

"The collection of geolocated placenames in Canada represents a consistent and comprehensive distribution of named places across Canada. Named places include large and small cities, villages, First Nations Communities, Small Hamlets etc.

"This data draws from public information maintained by Natural Resources Canada as part of the Canadian Geographical Names Database and public information maintained by Crown-Indigenous Relations and Northern Affairs Canada.

"The set of geolocated placenames is currently used for the administration of rural broadband Internet contribution programs, but is equally applicable for other mapping or modelling purposes where a comprehensive set of geolocated placenames across Canada is required."

(downloaded May 29, 2020 from open.canada.ca at the slow loading open.canada.ca/data/en/dataset/fe945388-1dd9-4a4a-9a1e-5c552579a28c)

SAheart *South African Heart Disease Data*

Description

From the web source: "A retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa. There are roughly two controls per case of CHD. Many of the CHD positive men have undergone blood pressure reduction treatment and other programs to reduce their risk factors after their CHD event. In some cases the measurements were made after these treatments."

The data are packaged here from the source (below). With one significant change (making chd a factor) they are also a repackaging of the data of the same name from the now archived (in 2020) of the 2015 'ElemStatLearn' package of Kjetil B. Halvorsen.

Format

A data frame with 462 rows and 10 variables

sbp Systolic blood pressure in millimetres of mercury (mm Hg).

tobacco Cumulative tobacco use in kilograms. Appears to be lifetime cumulative; not annual.

ldl Low density lipoprotein cholesterol.

adiposity Not recorded in source; presumably another measurement of obesity similar to BMI. Possibly a "corrected" version of obesity measure.

famhist Factor indicating presence or absence of a family history of ischaemic heart disease.

typea Type-A coronary prone personality behaviour as measured by a self-administered Bortner Short Rating Scale. Possible total scores can range from 12 to 84. Rossouw et al. (1983) "arbitrarily" classify those with scores of 55 or more "as exhibiting type A behaviour."

obesity A measure of obesity; body mass index (or BMI) is consistent with Rossouw et al. (1983). Having BMI >= 30 scored as "obese" by Rossouw et al. (1983).

alcohol Current alcohol consumption. Units of measurement (quantity/time) are unclear (e.g litres per annum, ounces per month?); alcohol not mentioned in Rossouw et al. (1983).

age Age in years at time of study (Source web page: 'Age at onset'.)

chd The response, a factor identifying whether the subject had been diagnosed as having coronary heart disease or not.

The row order of the values follow their order of appearence in the source webpage.

Details

In the late 1970s, an unusually high incidence of ischaemic heart disease had been observed to exist amongst white Afrikaans-speaking segments of South African society (Wyndham, 1982). Using an intensive postal campaign in 1979, Rossouw et al. (1983) recruited about 82 known target population of inhabitants of three Afrikaner communities in the southwestern Cape Province (3,357 white males and 3,831 white females).

For each subject, the binary response "chd" (originally appearing in the original file as 1 if they had coronary heart disease and 0 otherwise; but now as "Yes" or "No") was determined in the survey together with a variety of known risk factors for heart disease.

The goal was to explore the prevalence and intensity of chd risk factors in these high incidence communities with particular attention to those major risk factors (e.g. hypercholestrolaemia, hypertension, and smoking) which might be considered reversible (Rossouw et al., 1983).

Hastie and Tibshirani (1987) selected a subset of 465 subjects from the 3,357 white males (in these communities, male mortality rates were about two and a half times that of the females; see Rossouw et al., 1983). The 465 subjects consisted of all 162 cases having had coronary heart disease as well as 303 controls sampled from the remaining set of survey subjects.

The same (or similar) data seems to be used again for illustration in Hastie, Tibshirani, and Friedman (2009) and it is that which is now ported here from the book's accompanying website (see source). Curiously, this data set (viz. that recorded here) contains values on only 462 subjects, of which now only 160 are cases and 302 are controls.

In the current data set, rows 1-261 have row numbers matching the source "row.name", thereafter the row number is one less than the source "row.name". It would appear that subject with "row.name"

262 is absent from the source (below) and, speculatively, perhaps also those whose "row.name" could have been 464 and 465.

See references, particularly Rossouw et al (1983), for more details.

Author(s)

R.W. Oldford

Source

Trevor Hastie's "Elements of Statistical Learning" page at Stanford.

References

Trevor Hastie and Robert Tibshirani (1987) "Non-parametric logistic and proportional odds regression", JRSS-C (Applied Statistics), 36(3), 260–276.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman (2009) "The Elements of Statistical Learning", 2nd Edition, Springer New York <doi:10.1007/978-0-387-84858-7>

J.E. Rossouw, J.P.D. Plessis, A.J.S. Benad\'e, P.C.J. Jordaan, J.P. Kotz\'e, P.L. Jooste, and J.J. Ferreira (1983) "Coronary risk factor screening in three rural communities: The CORIS baseline study". South African Medical Journal, 64, 430-436.

C. Wyndham (1982) "Trends with time of cardiovascular mortality rates in the populations of the RSA for the period 1968-1977", South African Medical Journal, 61, 987-993.

`SCmolecule` *A protein/DNA complex molecule from Saccharomyces Cerevisiae*

Description

Contains three dimensional structure of the GAL4 protein of Saccharomyces Cerevisiae (or Baker's Yeast) recognizing and binding to a Deoxyribonucleic acid (DNA) sequence.

There are effectively two molecules here (a protein and DNA) binding together. It is a transcription/DNA complex within Saccharomyces Cerevisiae.

From Marmorstein et al (1992):

"The yeast protein GAL4 activates transcription of genes required for catabolism of galactose and melibiose. The DNA sequences recognized by GAL4 are 17 base pairs (bp) in length and each site binds a dimer of the protein."

and

"The protein fragment binds to its DNA site as a symmetrical dimer. Each subunit folds into three distinct modules: a compact, metal-binding domain (residues 8-40), an extended linker (41-49), and an a-helical dimerization element (50-64). Residues 1-7 and 65-66 are disordered. An overall view of the complex shows that a large part of the DNA major groove is not contacted by the protein. The DNA is relatively straight. A metal domain lies in the major groove near each end of the DNA fragment. The paired parallel helices of the dimerization element project away from the DNA along

the 2-fold axis of the complex. The metal-binding domain contacts three DNA base pairs in the major groove, and we therefore refer to it as a 'recognition module.' ...

The recognition module is held together by two metal ions, tetrahedrally coordinated by the six cysteines. Two of the cysteines (11 and 28) ligate both metals, creating a 'binuclear cluster' ... "

See Marmorstein et al (1992) for more on the geometry.

The source of most data here is Protein Data Bank (PDB) entry 1d66.

All coordinates, chains, residues, backbone atom identities, and displacements are taken from the PDB entry. Not included here are those entries from the PDB record which simply identify the terminus of each of the chains D, E, A, and B. Each of these 'TER' entries contain no coordinates since it simply marks the end of its chain.

Values were determined by X-ray crystallography at 2.7 Angstrom resolution.

These values have been supplemented with variable values from a variety of sources so as to help in the identification of components of the molecular structure.

Format

A data frame with 1762 rows and 14 variates:

group One of 'ATOM' or 'HETATM'. Here 'ATOM' indicates an atom having a standard residue of the protein; 'HETATM' (hetero atom) indicates one either having a non-standard residue of protein, or one in a group of a different kind such as carbohydrates, substrates, ligands, solvent, or metal ions. In the 'SCmolecule', these will be either a water molecule 'HOH' or a Cadmium ion 'CD'.

id Identification number of the backbone atom as given in the protein data bank (PDB).

label Atom identifier. These follow a standard used by the PDB. The first character is the element abbreviation of the backbone atom. The remaining characters of the nomenclature identify which of the atoms of that type are being referred to in the structure.

residue A two or three letter abbreviation naming the residue attached to that atom.

chain Identifies a chain of atoms. These are polypeptide or DNA chains.

sequence Order in which that bakbone atom appears in its chain.

x, y, z Coordinates of the bakbone atom in three-dimensonal space.

displacement Equivalent isotropic displacement factor; also sometimes earlier called a temperature factor. It is a measure of the possible coordinate location displacement of an atom from any source. Displacements could arise, for example, from atomic vibrations, such as (large) molecular motion or (smaller) internal vibrations, or any of a variety of sources of disorder. This is recorded as a spherical Gaussian (isotropic) measure of the variability of the location by the average eigen-value of a variance-covariance matrix.

type The element symbol of the backbone atom.

mass Atomic mass of the backbone atom.

residueType Type of the residue.

residueName Full name of the residue.

Author(s)

R.W. Oldford

Source

https://www.rcsb.org/3d-view/1D66/. https://bioinformatics.org/firstglance/fgij/fg.htm?mol=1d66

References

Ronen Marmorstein, Michael Carey, Mark Ptashne, and Stephen C. Harrison (1992) "DNA recognition by GAL4: structure of a protein-DNA complex", Nature, 356, pp. 408-414.

John L. Markley, Ad Bax, Yoji Arata, C. W. Hilbers, Robert Kaptein, Brian D. Syke, Peter E. Wright, and Kurt Wuthrich (1998) "Recommendations for the presentation of NMR structures of proteins and nucleic acids", European Journal of Biochemistry, 256, pp. 1-15.

Reinhard X. Fischer and Ekkehart Tilmanns (1988) "The equivalent isotropic displacement factor", Acta Crystallographica C44, pp. 775-776.

K.N. Truebloof, H.-B. Burgi, H. Burzlaff, J.D. Dunitz, C.M. Gramaccioli, H.H. Schulz, U. Shmueli, and S.C. Abrahams (1996) "Atomic displacement parameter nomenclature" Acta Crystallographica A52, pp. 770-781.

See Also

elements igg1

---

trtPan *Fictional pandemic data (with treatment variable)*

Description

An entirely artificially constructed data set and context designed for classroom discussion and analysis.

The data (and the fictional narrative below) are identical to those of pandemic; the only difference is the organization of the data. Which organization the instructor might choose depends upon the modelling and/or data manipulation is intended for analysis.

Should raise interesting scientific questions on how the results should, or should not, be interpreted. It should also raise questions on what might be done next.

A "pandemic" context is given in the details since the data were created during the first week of March, 2020.

Instructors might choose to invent their own context.

Format

A data frame with 300 rows and 3 variables

City City for which the outcome data were recorded.

Treatment The treatment used (one of "A", "B", or "C").

Recovery The percent survival rate for infected persons in that city when given that medical treatment.

Details

One fictional narrative for this data set is as follows.

A virulent virus has led to a world wide pandemic and that the case fatality rate (proportion of those infected who die) is huge (say 6

Suppose that through a concerted and collaborative effort of health scientists worldwide, three different treatments have been developed for this group. All three treatments have been used at one time or another on numerous patients in this group from 100 different cities worldwide. The data are observational, in that they were simply collected and the treatment given noted. No information is available on why one treatment or another was given in any particular instance.

The recovery rates (as a percent) for the patients treated by each of the three treatments were simply recorded for each of the hundred cities and are available for analysis as the data frame `pandemic`.

Some obvious questions of interest are the comparisons of treatments. For example, is treatment A better than B? Than C? Is B better than C?

Author(s)

R.W. Oldford

See Also

`pandemic` `medicalRecords` `covidNZ`

Index