Package: gpk (via r-universe)

October 9, 2024

Type Package

Title 100 Data Sets for Statistics Education

Version 1.0

Date 2013-07-12

Author Prabhanjan Tattar

Maintainer Prabhanjan Tattar <prabhanjannt@gmail.com>

Description Collection of datasets as prepared by Profs. A.P. Gore, S.A. Paranjape, and M.B. Kulkarni of Department of Statistics, Poona University, India. With their permission, first letter of their names forms the name of this package, the package has been built by me and made available for the benefit of R users. This collection requires a rich class of models and can be a very useful building block for a beginner.

License GPL-2

NeedsCompilation no

Date/Publication 2013-07-14 08:39:22

Repository https://prabhanjan-tattar.r-universe.dev

RemoteUrl https://github.com/cran/gpk

RemoteRef HEAD

RemoteSha e34de5aae3fa9c22e00ff9a54792b5df42748f6b

Contents

gpk-package *100 Data Sets for Statistics Education*

Description

Collection of datasets as prepared by Profs. A.P. Gore, S.A. Paranjape, and M.B. Kulkarni of Department of Statistics, Poona University, India. With their permission, this package has been built by me and made available for the benefit of R users. This collection requires a rich class of models and can be a very useful building block for a beginner.

Details

Package:	gpk
Type:	Package
Version:	1.0
Date:	2013-07-12
License:	GPL-2

Simply, follow the document files at the website of the data sets.

Author(s)

Prabhanjan Tattar

Maintainer: Prabhanjan Tattar <prabhanjannt@gmail.com>

AIDS *AIDS Data Set*

Description

A : Sr. no B : Pre test score of the student C : Post test score of the student D : Subject Specialization in code numbers 1: Chemistry (Special) 2: Botany (Special) 3: Microbiology (SYBSc level) 4: Microbiology (Special level) 5: Zoology (Special) E: Subject name

Usage

```
data(AIDS)
```

Format

A data frame with 72 observations on the following 5 variables.

`SR.NO` Serial Number

`PRE.TEST` Pre-test

`POST.TEST` Post-test

`Sub.Code` Subject code

`Subject` a factor with levels `Bot` `Chem` `Micro` Sy `Micro` TY `Zoology`

Details

In disease management the proverb ’prevention is better than cure’ is very relevant. Awareness is the first step in prevention. Hence any materials prepared to enhance awareness constitute a potent weapon in the hands of public health personnel. Two questions are of interest. Is the post-test score significantly higher than the pretest score? Are differences uniform across subjects and years?

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(AIDS)
head(AIDS)
t.test(AIDS$PRE.TEST,AIDS$POST.TEST,var.equal=TRUE)
t.test(AIDS$PRE.TEST,AIDS$POST.TEST,var.equal=FALSE)
```

`AirPollution` *Air Pollution Data*

Description

The goal is to understand the pollution dispersion as "Determinants of Air pollution"

Usage

```
data(AirPollution)
```

Format

A data frame with 151 observations on the following 11 variables.

`PM10` Particulate matter (size < 10 micorns)

`Pb` lead content in PM10

`Cd` cadmium content in PM10

`Cu` copper content in PM10

Cr chromium content in PM10

Zn zinc content in PM10

NOx Nitrogen oxide content in PM10

SO2 sulphur dioxide content in PM10

Site The sites

Date dates of the event

Season Season of the year

Details

The authors suggest that you try out Time series, ANOVA, and Regression on the data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(AirPollution)
head(AirPollution)
plot.ts(AirPollution[,1:8],plot.type="multiple",main="Air Pollution")
```

AizawlCancer *Sex-wise differences in cancer types*

Description

Analyze if the cancer percentages of male and female depends on the type of cancer.

Usage

```
data(AizawlCancer)
```

Format

A data frame with 19 observations on the following 5 variables.

Site Cancer in different areas

Female Female death to cancer

Male Male death to cancer

Details

Consider the problem as a count data and use statistical methods as in contingency table and grouping of categories.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(AizawlCancer)
head(AizawlCancer)
```

`Allergy` *Allergy Data Set*

Description

Cosmetic products can cause allergenic response in users. Such a development can do significant harm to the name of its producer. Hence it is routine to carry out safety tests. In a study to compare irritability of 4 products, seven individuals were asked to apply each product on forearm. Higher the irritation score worse is the product.

Warning: since observations on the same individual are correlated, use of ANOVA may not be valid.

Usage

```
data(Allergy)
```

Format

A data frame with 7 observations on the following 4 variables.

`ProdA` Irritation score for product A

`ProdB` Irritation score for product B

`ProdC` Irritation score for product C

`ProdD` Irritation score for product D

Details

Cosmetic products can cause allergenic response in users. Such a development can do significant harm to the name of its producer. Hence it is routine to carry out safety tests. In a study to compare irritability of 4 products, seven individuals were asked to apply each product on forearm. Higher the irritation score worse is the product. Warning: since observations on the same individual are correlated, use of ANOVA may not be valid.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Allergy)
friedman.test(as.matrix(Allergy))
```

Asthma1 *Testing Effect of Curcuma Longa*

Description

Histamine induces contraction of goat trachea. This causes difficulty in breathing. Curcuma longa is expected to reduce contraction.

Usage

```
data(Asthma1)
```

Format

A data frame with 12 observations on the following 4 variables.

`Log_Concentration_Histamine` Histamine dose

`Response_Without_Curcuma_Longa` Response without Curcuma longa

`Response_With_Curcuma_Longa` Response with Curcuma longa

`Group` Set identity

Details

Try fitting a regression model and a lack-of-fit test.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Asthma1)
```

Asthma2 *Testing effect of treatment on milk induced Eosinophilia in mice*

Description

Milk increases Eosinophil count. Abnormal increase in blood Eosinophils causes narrowing of airways. Curcuma longa is expected to reduce impact of milk.

Usage

```
data(Asthma2)
```

Format

A data frame with 10 observations on the following 4 variables.

`Animal.code` Animal code

`Before` Response (density of Eosinophils i.e. count per cubic mm blood) before milk treatment

`After` Response (density of Eosinophils i.e. count per cubic mm blood) 24 hours after milk treatment

`Group` Group identity

Details

Two sample t-test and ANOCOVA are suggested for the data on hand.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Asthma2)
```

`Asthma3` *Effect of curcuma longa on de-granulation of mast cells in mice*

Description

Mast cells if de-granulated release histamine causing allergic reaction. It is of interest to reduce de-granulation percentage.

Usage

```
data(Asthma3)
```

Format

A data frame with 15 observations on the following 5 variables.

`Treatment` Treatment types

`Animal.Code` The Animal Code

`Response` Response as a percentage of the de-granulated cells

Details

ANOVA, multiple comparisons, transformation methods are recommended to be performed on the data.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Asthma3)
```

Asthma4 *Testing effect of Curcuma longa on paw inflammation in rats*

Description

Swelling of paw is an indication of inflammation. Curcuma longa is supposed to reduce this. Two questions are of interest. a) Comparison of three treatments at each time point. b) Fitting trend line over time and comparison of slopes across treatments.

Usage

```
data(Asthma4)
```

Format

A data frame with 15 observations on the following 6 variables.

`Treatment` Treatment

`Animal.number` Animal code

`X30min` Response (paw edema in mm) after 30 min

`X1hr` Response (paw edema in mm) after 1 Hr

`X2hr` Response (paw edema in mm) after 2 Hr

`X3hr` Response (paw edema in mm) after 3 Hr

Details

ANOVA and regression models are suggested.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Asthma4)
```

atombomb *Cancer deaths of atomic bomb survivors*

Description

Two atom bombs were dropped on two cities in Japan (Hiroshima and Nagasaki) in World War II. Thousands of people died in the blast. Even more were exposed to radioactive materials and as an after effect developed cancer. Over the years many of these cancer patients also succumbed to the disease. This was of course a fraction of those exposed to radiation. Person-years at risk (100s) is the sum total of all years spent by all persons in the category. Suppose the mean number of deaths in each group is Poisson with mean = risk*rate. Risk is the person-years at risk and rate is the rate of cancer deaths per person per year. This mean is expected to depend on amount of radiation and time since exposure. Effect of exposure may be linear or quadratic and hence rad and rad2 are the suggested independent variables.

Usage

```
data(atombomb)
```

Format

A data frame with 30 observations on the following 14 variables.

`Extent_of_Exposure` Radian levels

`Years_Exposure` Bucketized into intervals

`Death_Count` the death count

`At_Risk_Count` the at-risk cound

Details

Poisson regression is recommended.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(atombomb)
```

Bacteria — *A multi-factorial experiment on the bacteria growth in the packaged foods*

Description

In packaged foods one critical concern is the shelf life. Spoilage of food due to bacterial growth can cause major losses. Hence it is of interest to identify conditions which minimize bacterial growth. It is suspected that salt and lipid concentration, pH and temperature may affect growth. The task is to identify levels of various factors, check significance of main effects and interactions and plot cell means in case of two factor interactions that are significant.

Usage

```
data(Bacteria)
```

Format

A data frame with 300 observations on the following 5 variables.

`Response` Reponse

`Salt` salt concentration in the medium

`Lipid` lipid concentration in the medium

`pH` pH of the medium

`Temp` temperature

Details

ANOVA is recommended here.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Bacteria)
head(Bacteria)
```

BambooGrowth *Data set relating growth of bamboo to geographic location*

Description

Bamboo is a useful plant belonging to the plant group 'grasses'. An individual bamboo plant is called a clump because it is a cluster of many sticks (culms). Individual culms may live for 10 years. The clump may live for 40 years. Every year the plant develops new shoots which later on become old shoots / culms. New shoots have food value. Culms are used for mats, roofs etc. It is of interest to check variation in growth rates of plants. In particular we may want to assess effect of location on growth. The data has 2 responses in columns D and E. Information on location is hierarchical. Compartment is the largest unit. Blocks are parts of compartments. Transects are lines drawn within blocks. We may compare transects within blocks, blocks within compartments and finally compartments. Analysis can be univariate or bivariate.

Usage

```
data(BambooGrowth)
```

Format

A data frame with 595 observations on the following 14 variables.

`Compartment` Compartment (in forest)

`Locality_Block` Locality

`Transect_Number` Transect

`Old_Shoots` number of old shoots in the clump

`New_Shoots` number of new shoots in the clump

Details

Nested ANOVA univarite and bivariate are suggested tools for analyses.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(BambooGrowth)
```

`Bamboolife` *Preparing a life table for the Bamboo plant*

Description

Bamboo is a peculiar plant. Individual plant is a cluster (called clump) of shoots called culms. A clump has a long life of about 40 years and produces seeds in super abundance only once, at the end of life. Further all bamboo clumps in a region develop flowers and seeds together and die together at the same time. There is no variation in age at death. However, there is variation in the lifespan of individual culms. Data is on survival of 439 culms of bamboo (Dendrocalamus strictus).

Usage

```
data(Bamboolife)
```

Format

A data frame with 16 observations on the following 2 variables.

`Age` Age in years

`Survivors` Number of survivors

Details

Life tables is suggested for the current data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Bamboolife)
```

`Bananabats` *The Bat Census data*

Description

This is a study of bats that live in folded banana leaves. The community keeps changing in terms of total number, composition etc. Animals caught are examined for various features such as age, sex, reproductive status etc. and released again. Number in Column E (KTBA) is always greater than that in Column D (Number.observed) which is at least as big as that in column C (Number.caught). Questions of interest are the following: a) Is the percent caught uniform over periods? Is there any time trend in this variable? b) Is the percent observed uniform over periods? Is there any time trend in this variable? c) Does the 'number escaped' (observed - caught) follow a Poisson distribution?

Usage

```
data(Bananabats)
```

Format

A data frame with 16 observations on the following 5 variables.

`Date`	Date
`Period`	Period
`Number.caught`	Number caught
`Number.observed`	Number observed
`KTBA`	Number known to be alive (KTBA)

Details

Contingency tables, regression, goodness of fit chi-square are suggested for the purpose of statistical analyses.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Bananabats)
```

BANK *Bank Churn data set*

Description

Businesses like banks which provide service have to worry about problem of 'Churn' i.e. customers leaving and joining another service provider. It is important to understand which aspects of the service influence a customer's decision in this regard. Management can concentrate efforts on improvement of service, keeping in mind these priorities.

Usage

```
data(BANK)
```

Format

A data frame with 245 observations on the following 20 variables.

`Serial_Number` Serial Number

`Response` Response (1\: deserter, 0\: Loyal)

`Branch` Branch code

`Occupation` Occupation of Customer

`Age` Age in Years

`Sex` Gender

`Pleasant_Ambiance` Pleasant Ambiance ACT1

`Comfortable_seating_arrangement` Comfortable seating arrangement ACT2

`Immediate_attenttion` Immediate attenttion ACT4

`Good_Response_on_Phone` Good Response on Phone ACT5

`Errors_in_Passbook_entries` Errors in Passbook entries ACT10

`Time_to_issue_cheque_book` Time to issue cheque book ACT14

`Time_to_sanction_loan` Time to sanction loan ACT16

`Time_to_clear_outstation_cheques` Time to clear outstation cheques ACT17

`Issue_of_clean_currency_notes` Issue of clean currency notes ACT24

`Facility_to_pay_bills` Facility to pay bills ACT26

`Distance_to_residence` Distance to residence ACT28

`Distance_to_workplace` Distance to workplace ACT30

`Courteous_staff_behaviour` Courteous staff behaviour ACT31

`Enough_parking_place` Enough parking place ACT32

Details

Explore the application of logistic regression and contingency tables for this data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(BANK)
```

`Barleyheight` *Comparison of genotypes and checking time trend*

Description

Data are plant height of barley measured in cm. Averaged over 4 replicates for 15 genotypes. It is of interest if there is any difference in genotypes and also if there is any time trend in heights for the same genotype. Further, one may check if trend is similar across genotypes. It is of interest to explain height using weather information.

Usage

```
data(Barleyheight)
```

Format

A data frame with 9 observations on the following 23 variables.

`Years` Year

`Genotype1` Geno type 1

`Genotype2` Geno type 2

`Genotype3` Geno type 3

`Genotype4` Geno type 4

`Genotype5` Geno type 5

`Genotype6` Geno type 6

`Genotype7` Geno type 7

`Genotype8` Geno type 8

`Genotype9` Geno type 9

`Genotype10` Geno type 10

`Genotype11` Geno type 11

`Genotype12` Geno type 12

`Genotype13` Geno type 13

`Genotype14` Geno type 14

`Genotype15` Geno type 15

`Sowing.day.Number.days.since.April1` Sowing day Number of days since April 1

`Rainfall1` Rainfall per day (mm) averaged for each of growth period 1

`Rainfall2` Rainfall per day (mm) averaged for each of growth period 2

`Rainfall3` Rainfall per day (mm) averaged for each of growth period 3

`Rainfall4` Rainfall per day (mm) averaged for each of growth period 4

`Rainfall5` Rainfall per day (mm) averaged for each of growth period 5

`Rainfall6` Rainfall per day (mm) averaged for each of growth period 6

Details

Try ANOVA, regression, and time series analysis.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Barleyheight)
```

`Batcapture` *Understanding seasonality and species composition of bat population*

Description

Scientists studied Bat community on Barro Colorado island in Panama in late 70's. Bats were captured on many nights regularly throughout the year to understand species composition of the bat community and its dynamics through the year. Following questions are of interest: a) For a given species, do capture proportions change with season? If yes, how? b) Are capture proportions over months similar across species? If not, what are the salient differences? c) Consider the variable, average number caught per night in a month. Does it depend on number of nights/month/ species?

Usage

```
data(Batcapture)
```

Format

A data frame with 8 observations on the following 23 variables.

`Species` The type of species

`Jan.Netting.nights` The January Netting nights

`Jan.Number.caught` The January number count

`Feb.Netting.nights` The February netting nights

`Feb.Number.caught` The February number count

`Mar.Netting.nights` The March netting nights

`Mar.Number.caught` The March number count

`Apr.Netting.nights` The April netting nights

`Apr.Number.caught` The April number count

`May.Netting.nights` The May netting nights

`May.Number.caught` The May number count

`Jun.Netting.nights` The June netting nights

`Jun.Number.caught` The June number caught

`Sep.Netting.nights` The September netting nights

`Sep.Number.caught` The September number count

`Oct.Netting.nights` The October netting nights

`Oct.Number.caught` The October number count

`Nov.Netting.nights` The November netting nights

`Nov.Number.caught` The November number count

`Dec.Netting.nights` The December netting nights

`Dec.Number.caught` The December number caught

`Total.Netting.nights` The total netting nights

`Total.Number.caught` The total number count

Details

Try out Time trends, contingency tables and regression, and comment.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Batcapture)
```

`BatGroup` *Fitting distributions to the bat group size data*

Description

In a study of a bat community, scientists were interested in social behavior. In particular, they wanted to see whether bats are loners or they prefer to be in groups. Research on other taxa suggests that a Poisson or a negative binomial distribution may be appropriate. Of course, value zero of group size is not observable and needs to be truncated.

Usage

```
data(BatGroup)
```

Format

A data frame with 6 observations on the following 9 variables.

`Month` The months

`GS_1` Frequency of occurrence of group size 1

`GS_2` Frequency of occurrence of group size 2

`GS_3` Frequency of occurrence of group size 3

GS_4 Frequency of occurrence of group size 4

GS_5 Frequency of occurrence of group size 5

GS_6 Frequency of occurrence of group size 6

GS_7 Frequency of occurrence of group size 7

GS_GT_7 Frequency of occurrence of group size greater than 7

Details

Try out Goodness of fit and truncated distributions.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(BatGroup)
```

Batrecapture *Fitting a model to bat recapture data*

Description

In a particular study bats were captured regularly for over an year. About 9000 different individuals were captured and released back, some of them more than once. We assume that all bats are equally likely to be captured regardless of how often they have experienced capture earlier. We have to see if such a model fits the data well. If not, we can try a modification in which probability of capture changes depending upon the number of times the individual is captured earlier. The probability may go up (trap attraction) or down (trap shyness).

Usage

```
data(Batrecapture)
```

Format

A data frame with 11 observations on the following 2 variables.

Number.recapture The number of times a bat is recaptured

Number.individuals The frequency of the number of times a bat is caught

Details

Suggested solution: MLE and chi-square goodness of fit test.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Batrecapture)
```

`Biodegradation` *Biodegradation of Dimethoate in Industrial Effluents by Brevundimonas species*

Description

Industrial effluents may contain large amounts of toxic material. This material can cause hazard to birds, fish etc. Scientists make efforts to remove such toxic material from flowing waters. Biodegradation by microorganisms is one of the ways. Parameters likely to promote microbial growth (and in turn cause more degradation of toxic material) include pH, ambient temperature, inoculum size and stirring. A 2^4 experiment was conducted. The data give results of this experiment. Objective is to find conditions, which will cause maximum removal (maximum growth).

Usage

```
data(Biodegradation)
```

Format

A data frame with 16 observations on the following 5 variables.

`pH` pH level

`Temp` Temperature

`Inoculum` Inoculum at two levels

`Aeration` Aeration at two levels `No` `Yes`

`Percent.Removal` Response (% removal of the Dimethoate)

Details

Build a factorial experiment model for the data set and evaluate for the interaction effect.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Biodegradation)
```

birdextinct *Bird extinct at a national park*

Description

One of the major controversies in conservation biology is 'a few small versus many large'. The problem is that of optimum use of resources to conserve species. If we have limited land and we wish to use it to create protected areas for conserving say bird species, should we make one large national park out of it or should we have many small sanctuaries? This depends on extinction rates as a function of area of a park or sanctuary. If the relation is linear then it does not matter. If there are economies of scale, it may be better to have a few large parks. In a study of several islands in Finland, two surveys, one in 1949 and the other in 1959 were used to decide the number of species present and those that went extinct in 10 years. We need to check the relationship between the area and proportion that went extinct.

Usage

```
data(birdextinct)
```

Format

A data frame with 18 observations on the following 4 variables.

`Site` Site Number

`Area` Area in square kilometer

`Species_at_risk` Species at risk

`Number_of_Species_extinct` Species extinct

Details

Experiment a regression model before and after certain transformation.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(birdextinct)
```

BirthDeath *Changes in Human birth and death rates in India over the 20th century*

Description

Birth and death rates are important indices of demographic picture of a country. Given data for last century can we predict the rates and net population growth for next decade?

Usage

```
data(BirthDeath)
```

Format

A data frame with 27 observations on the following 3 variables.

`Year` The year grip

`Birth.rate` The birth rate

`death.rate` The death rate

Details

Analyze the time trends in birth and death rates. Time trends in net growth rate.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(BirthDeath)
```

BPSYS *Two drug comparison*

Description

Two drugs are to be compared for their effect on blood pressure. Only one aspect of blood pressure is considered here, viz. systolic standing. Response is measured before and after treatment. Principal question of interest is in two parts (a) Is treatment Al (Ay) effective? (b) Is one treatment better than the other? It is advisable to compare treatments after discounting for other differences between patients.

Usage

```
data(BPSYS)
```

Format

A data frame with 35 observations on the following 8 variables.

`Pat_no` Patient Number

`Age` Patient Age

`Sex` Gender of the patient

`Duration_of_hypertension_yrs` Patient history 1 (Duration of hypertension (yrs))

`Duration_of_diabetes_yrs` Patient history 2 (Duration of diabetes (yrs))

`BaselineSystolic_BP` Systolic Blood Pressure: at baseline (before treatment)

`Week_8_Systolic_BP` Systolic Blood Pressure: after 8 weeks of treatment

`Drug` Drug : Al- Alopathic, Ay- Ayurvedic

Details

t-test and ANOCOVA are recommended to carry out the analysis.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(BPSYS)
```

`Butterflies` *Study of distribution of butterfly species count among 5 groups and in different localities in India*

Description

One encounters butterflies in most localities in India. Species richness is the total number of different butterfly species encountered in a locality. This total count can be divided into counts of different families (5 families in the present case). It is expected that distribution among the 5 families should be similar across localities that are geographically closer or ecologically similar. Further, some families may have a nearly constant share in different localities. It may then be enough to count species of that family and guess counts in other families.

Usage

```
data(Butterflies)
```

Format

A data frame with 44 observations on the following 16 variables.

`Serial_Number` Serial Number

`Area` Different areas of the continent

`Locality` The locality of the species

`Total_Species_count` Total species count

`Skippers` Count of skippers

`Swallow_tails` Count of swallow tails

`Whites_Yellows` Count of whites and yellows

`Blues` Count of blues

`Brush_Footed` Count of brush footed species

Details

Try out regression models with principal component analysis.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Butterflies)
```

`Chitalparasite` *Understanding the correlation of occurrence of a parasite*

Description

It is of interest to examine relation between presence/ absence of parasite and other attributes.

Usage

```
data(Chitalparasite)
```

Format

A data frame with 66 observations on the following 10 variables.

`Sarcocystis_Indicator` a numeric vector

`Sanctuary_Indicator` a numeric vector

`Predator_Indicator` a numeric vector

`Tissue_Indicator` a numeric vector

`SEX` Gender

`YEAR` Year

Details

Explore contingency tables and logistic regression model for this data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Chitalparasite)
```

`cloudseed` *Cloud Seeding*

Description

4. It is now believed that we can induce rainfall by seeding clouds i.e. releasing certain chemicals (silver iodide) in clouds. Of course there can be doubts such as whether there is indeed a net increment in rainfall or just a re-distribution over a certain territory etc. Analysis of relevant data can therefore be of considerable interest. In this experiment days for seeding were selected randomly out of a set of 52 days suitable for seeding.

Usage

```
data(cloudseed)
```

Format

A data frame with 52 observations on the following 2 variables.

`Rainfall` Rainfall on a day

`Seeded.Indicator` Treatment

Details

Use box plots to check nature of distributions, and transformation to bring about homoscedasticity followed by testing equality of means.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(cloudseed)
```

Cosmetic1 *Testing efficacy of a cosmetic product*

Description

A manufacturer of cosmetic products is interested in frequent introduction of new formulations in to the market. This may be for one of two reasons. A product may have a short life and may lose popularity after that. Alternatively some new formulation is developed which may be superior to products available in the market. This necessitates comparison of the new product with competitors. In a typical study a product is tried on a group of panelists and some trait is measured before and after use of product. Examples of such traits are skin oiliness or softness or fairness.

Usage

```
data(Cosmetic1)
```

Format

A data frame with 48 observations on the following 3 variables.

`Treatment` Product code

`Initial` Initial Value of trait

`Change` Change in trait

Details

Try ANOCOVA.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Cosmetic1)
```

COWSDATA *Crossbreeding of Cows*

Description

A project on crossbreeding of cows was conducted at multiple centers. A cow when inseminated during heat may or may not conceive. Factors likely to affect success are semen used (fresh or frozen), hormonal status of cow, etc. Cows were brought to the centers when they were found to be in the "heat" state. Time lag between onset of heat and insemination was noted as also success or failure of insemination. Veterinary practitioners believe that maximum "success" is observable if the insemination is practiced within 24-30 hours from onset of heat.

Usage

```
data(COWSDATA)
```

Format

A data frame with 10 observations on the following 7 variables.

`Time` Time since onset of heat

`Sillod_Insemination_C1` Insemination count (center 1)

`Sillod_Conception_C1` Conception count (center 1)

`Sillod_Insemination_C2` Insemination count (center 2)

`Sillod_Conception_C2` Conception count (center 2)

`Sillod_Insemination_C3` Insemination count (center 3)

`Sillod_Conception_C3` Conception count (center 3)

Details

ANOVA-for proportions, arcsine transformation, comparison of slopes, and regression of conception rate on time for each center are some of the suggested methods for the user.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(COWSDATA)
```

Crack *Healing the heel*

Description

People who work bear foot often suffer from cracks in the heel. If the cracks are severe they can cause pain, bleeding, infection etc. Many traditional remedies are in use for this ailment. In a study to test efficacy of an ayurvedic treatment, severity of cracking was recorded and also typical length of a crack. This was done for each heel before and after treatment. It is of interest to check whether the treatment is effective. This broad question can be broken down to many sub-questions, for example, (a) Has the severity grade remained the same for right (left) heel? (b) Has the crack length remained the same for right (left) heel? (c) Is the improvement in severity same for two heels?

Usage

```
data(Crack)
```

Format

A data frame with 17 observations on the following 4 variables.

`Right_Heel_Change_Grade` Change in grade (severity) of cracking (right heel)

`Right_Heel_Change_Length` Change in typical crack length (right heel)

`Left_Heel_Change_Grade` Change in grade (severity) of cracking (left heel)

`Leftt_Heel_Change_Length` Change in typical crack length (left heel)

Details

One sample t-test (univariate), paired t-test (univariate), Hotelling's T2

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Crack)
```

Crime *Relation between crime and intelligence*

Description

It is of interest to know the relationship between intelligence of the criminal and his delinquency (crime) index (from 0 to 50), which is a combination of frequency of crime and seriousness of criminal acts of an individual. This may help in 'managing' the case in jail. So we need to know the general rule and exceptions if any etc. Prepare a report on the nature of relationship between the two variables. It should include essential technical details and should guide a non-statistician who has to use it in his job of jail management.

Usage

```
data(Crime)
```

Format

A data frame with 18 observations on the following 2 variables.

`Delinquency.index` delinquency index

`Intelligence.Quotient` IQ

Details

Regression analysis and study of residuals need to be performed on this data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Crime)
```

DroughtStress *Modeling Genotypic variation in photosynthetic competence of Sorghum bicolor*

Description

Agriculture is the main source of income for nearly half of the Indian population. Most agriculture depends on monsoon rainfall. Hence results are uncertain. In years of drought, crops are often completely wiped out. It is therefore of great interest to identify crop varieties that can give at least some yield, even when faced with drought. That is why study of drought tolerance is very important. It is of course possible to try different varieties under various conditions of water shortage and compare results. Another possibility is to understand the biology of drought tolerance.

Usage

```
data(DroughtStress)
```

Format

A data frame with 33 observations on the following 58 variables.

`Variety` Variety

`RWC_00` Relative Water Content (RWC %) at 0 PEG Concentration

`CO2FIx_00` CO2 Fixation rate at 0 PEG Concentration

`Cond_00` Conductance at 0 PEG Concentration

`IntCO2_00` Intracellular CO2 Concentration At 0 PEG Conc

`RWC_05` Relative Water Content (RWC %) at 5 PEG Concentration

`CO2FIx_05` CO2 Fixation rate at 5 PEG Concentration

`Cond_05` Conductance at 5 PEG Concentration

`IntCO2_05` Intracellular CO2 Concentration At 5 PEG Conc

`RWC_10` Relative Water Content (RWC %) at 10 PEG Concentration

`CO2FIx_10` CO2 Fixation rate at 10 PEG Concentration

`Cond_10` Conductance at 10 PEG Concentration

`IntCO2_10` Intracellular CO2 Concentration At 10 PEG Conc

`RWC_15` Relative Water Content (RWC %) at 15 PEG Concentration

`CO2FIx_15` CO2 Fixation rate at 15 PEG Concentration

`Cond_15` Conductance at 15 PEG Concentration

`IntCO2_15` Intracellular CO2 Concentration At 15 PEG Conc

`RWC_20` Relative Water Content (RWC %) at 20 PEG Concentration

`CO2FIx_20` CO2 Fixation rate at 20 PEG Concentration

`Cond_20` Conductance at 20 PEG Concentration

`IntCO2_20` Intracellular CO2 Concentration At 20 PEG Conc

`RWC_25` Relative Water Content (RWC %) at 25 PEG Concentration

`CO2FIx_25` CO2 Fixation rate at 25 PEG Concentration

`Cond_25` Conductance at 25 PEG Concentration

`IntCO2_25` Intracellular CO2 Concentration At 25 PEG Conc

Details

Drought stress tolerance may be exhibited by plants through their ability to maintain a higher water potential under stress conditions (dehydration avoidance) or by maintaining physiological processes like photosynthesis at lower water potentials (dehydration Tolerance). 11 Sorghum bicolor (jowar) varieties, known to differ in their drought tolerance, were compared for their photosynthetic adaptation. Columns B, C, D and E are all responses (photosynthetic traits). It is of interest to compare these 11 varieties with respect to the responses measured, at a given value of PEG concentration. Also it is of interest to model changes in each response as functions of PEG concentration and comparing these models across varieties. Suggested tools include ANOVA, MANOVA, regression, and Graphical techniques.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(DroughtStress)
```

`Dunglife` *Dung decay data*

Description

In wild life studies it is necessary to estimate number of animals of a species. This can be done in two ways, direct counting of animals or indirect counting using dung piles. The logic behind indirect count is as follows: each animal produces a given number of dung piles per day (P). One dung pile remains observable on the ground for a few days (D) after which it gets mixed up with soil. Estimate of the number of animals is given by total number of dung piles on the ground divided by (D*P). Thus we need to know the average number of days (D) for which a dung pile lasts on the ground. In case of dear, dung is described using the term 'pellet'. The data given refer to a study on dear conducted in Bandipur Tiger Reserve in Karnataka, India.

Usage

```
data(Dunglife)
```

Format

A data frame with 55 observations on the following variable.

`Decay` Days to decay (life in days)

Details

Fitting exponential, Weibull, gamma distributions, and fitting quadratic hazard function may be attempted on the data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Dunglife)
```

Earthquake *Modeling earthquake aftershocks*

Description

5 Seismology is the study of earthquakes. An earthquake creates shock waves that travel from epicenter outwards. Like any waveform these waves have two main features. One is amplitude and the other is period. Amplitude is the maximum height from the x axis (or alternatively half of the distance between peak and trough) while period is the distance along x axis between two successive peaks. For more details, go to the web and fetch details from the file "EarthQuake.doc".

Usage

```
data(Earthquake)
```

Format

A data frame with 66 observations on the following 13 variables.

- `Date` Date
- `Hours` Hours
- `Minutes` Minute
- `Magnitude_IMD` Magnitude of earth quake as reported by India meteorology department (IMD)
- `Magnitued_USGS` Magnitude of earth quake at epicentral distance >= 50 as reported by United States Geological Survey (USGS) (mb)
- `Magnitude_NGRI` Magnitude of earth quake at epicentral distance >= 150 as reported by National Geophysical Research Institute (NGRI) (Ms)
- `Coda_duration_1_mm` Coda duration (seconds) at 1mm of background noise level
- `Coda_duration_2_mm` Coda duration (seconds) at 2mm of background noise level
- `Coda_duration_6_mm` Coda duration (seconds) at 6mm of background noise level
- `Coda_duration_10_mm` Coda duration (seconds) at 10mm of background noise level

Details

Regression, data transformation may be considered for analysis.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Earthquake)
```

Earthwormbiomass *Earthworms in cultivated soils*

Description

Earthworms are known to play an important role in farming by improving soil fertility. They enhance physical, chemical and biological aspects of soil fertility. Hence understanding their population dynamics is important. In one of the studies on earthworms in cultivated soils, over 2000 individual worms belonging to 6 species were collected. Samples were collected from three crops for two consecutive years. Questions of interest are: a) What are the factors affecting density of the earthworms? b) What are the factors affecting biomass of the earthworms?

Usage

```
data(Earthwormbiomass)
```

Format

A data frame with 12 observations on the following 5 variables.

`Density` a numeric vector

`Biomass` a numeric vector

`Crop` a factor with levels `Maize Paddy and Pulses` `Wheat and Mustard`

`Year` a numeric vector

`Soil` a factor with levels `0-10` `10-20`

Details

Consider a generalized linear model!

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Earthwormbiomass)
```

EarthwormSeason *Population dynamics of earthworms*

Description

Earthworms are known to play an important role in farming by improving soil fertility. They enhance physical, chemical and biological aspects of soil fertility. Hence understanding their population dynamics is important. In one of the studies on earthworms in cultivated soils, over 2000 individual worms belonging to 6 species were collected. Samples were collected from three crops for two consecutive years. Questions of interest are: a) How does the worm density change with season? b) How does the worm biomass change with season? c) What is the relationship between density and biomass?

Usage

```
data(EarthwormSeason)
```

Format

A data frame with 46 observations on the following 3 variables.

`Month` Month

`Density` the number of earthworms per square meter

`Biomass` biomass, fresh weight per square meter

Details

Try out time series techniques and regression methods.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(EarthwormSeason)
```

`elephant` *Age and mating success for Elephants*

Description

Elephants reach maturity at about 14 years of age. But they have to compete with all adult males for mating opportunity. Females are generally more receptive to larger males. Size of an elephant increases as age increases. Hence it is expected that generally the number of matings should increase with age. Is there an optimal age after which the success rate does not rise further? Mating is a rare event and hence may follow a Poisson distribution.

Usage

```
data(elephant)
```

Format

A data frame with 41 observations on the following 2 variables.

`Age_in_Years` age of the elephant in yrs

`Number_of_Matings` number of successful matings

Details

Poisson regression may be attempted.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(elephant)
```

`Euphorbiaceae` *Relationship between tree height and girth of Euphorbiaceae*

Description

For various reasons it is of interest to estimate volume of a tree. Obviously direct measurement is quite difficult. The next best thing is to assume that the tree has a conical shape. Even with this assumption we need to measure height and radius at the base for estimating volume. Measurement of height can be cumbersome. However, measurement of girth is easy at about one meter height. Hence the simplest method is to use some relationship between girth and height and estimate height from girth.

Usage

```
data(Euphorbiaceae)
```

Format

A data frame with 106 observations on the following 4 variables.

`Family` Family name

`Species_Name` Species name

`GBH` Girth at breast height (GBH-cm)

`Height` Height (meters)

Details

Linear/ nonlinear regression. Testing the hypothesis: the relationship is same across species. Deciding which species are closer to each other in this regard.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Euphorbiaceae)
```

`Extruder` *Understanding effect of manufacturing conditions on product characteristics*

Description

Extrusion is a process in which dough-like raw material is pushed through a machine and the machine puts out product in desired form, followed by some finishing touches. One important characteristic of an item produced is its weight. If weight is too low, product may be weak. If weight is too high, it may mean wastage of raw material. Hence a manufacturer is keen to know the relation between product weight and various parameters of manufacturing process. In a particular factory 3 parameters likely to affect weight were monitored. These were Extruder RPM, current and conveyer speed.

Usage

```
data(Extruder)
```

Format

A data frame with 49 observations on the following 4 variables.

`WEIGHT` Weight of product

`EXTRUDER_RPM` Extruder speed \[RPM - revolutions per minute\]

`CURRENT` Current

`Conveyer_Speed` Conveyer speed

Details

Fit a multiple regression model and carry out the residual analysis. Also, perform the identification of outliers.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Extruder)
```

`Fairness` *Comparison of formulations and sample size determination of a fairness product*

Description

In India, teenagers are very conscious of their looks, in particular, skin complexion. There is a considerable premium on fairness. Hence one encounters many fairness-inducing products in the market. The present data set shows changes in fairness level after use of a product. There are records on 25 panelists for each product. It is necessary to carry out comparison among three products, assuming that data are continuous. Note that actual values appear to be discrete. Organize the data in the form of a contingency table and check if the conclusion remains the same. This was a pilot study. The main experiment is now to be planned. Calculate the minimum sample size necessary to compare products A and B. Assume level of significance 0.05 and power =0.9 at the alternative that the two means differ by 0.25 (assume known and common variance).

Usage

```
data(Fairness)
```

Format

A data frame with 25 observations on the following 3 variables.

`Prod_A` Response to product A

`Prod_B` Response to product B

`Prod_C` Response to product C

Details

Try out ANOVA and chi-square test for comparison of the average response.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Fairness)
```

FAMILY *Understand relationship between height of parents and child*

Description

In an undergraduate program in statistics, students collected data on height of father, mother and age, sex and height of the children. Are heights of adults normally distributed? Can we predict the height of a child using the data on parents' heights? Does sex of a child matter? Does parity matter? Is the sex of the child related to parity? Study the distribution of time gap between successive births in a family.

Usage

```
data(FAMILY)
```

Format

A data frame with 288 observations on the following 17 variables.

`Serial_Number` Serial Number

`Family_Code` Family code

`FHT` Father's height (in cm.)

`MHT` Mother's height (in cm.)

`Children` Number of children in the family

`SEX_C1` Sex of child 1

`AGE_C1` Age of child 1 (in yrs)

`HT_C1` Height of child 1 (in cm.)

`SEX_C2` Sex of child 2

`AGE_C2` Age of child 2 (in yrs)

`HT_C2` Height of child 2 (in cm.)

`SEX_C3` Sex of child 3

`AGE_C3` Age of child 3 (in yrs)

HT_C3 Height of child 3 (in cm.)

SEX_C4 Sex of child 4

AGE_C4 Age of child 4 (in yrs)

HT_C4 Height of child 4 (in cm.)

Details

Many statistical methods are appropriate here. The following are recommended more (i) Correlation and regression, (ii) Markov chain, (iii) goodness of fit tests - Fitting of Normal distribution to the data on height separately, (iv) Tests for proportions etc..

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(FAMILY)
```

Filariasisage *Infection among Filariasis*

Description

Filariasis is a common infection in tropical and subtropical countries. Several parasites can cause filariasis. In Nigeria a study was conducted to see prevalence of filariasis due to various parasite types. Specific questions of interest are: a) What is the relationship of overall prevalence of filariasis with age? b) What is the relationship of prevalence of filariasis due to Onchocerca volvulus with age?

Usage

```
data(Filariasisage)
```

Format

A data frame with 8 observations on the following 5 variables.

Age_Group Age group

Examined Number Examined

Infected Number infected

Onchocerca_volvulus Number of cases infected by Onchocerca volvulus

Other Number of cases infected by other parasites

Details

Regression and relative risk modeling may be attempted.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Filariasisage)
```

`FilariasisSex` *Sex related prevalence in human filariasis*

Description

Filariasis is a common infection in tropical and subtropical countries. Several parasites can cause filariasis. In Nigeria a study was conducted to see prevalence of filariasis due to various parasite types. Specific question of interest is, whether prevalence is similar in both sexes.

Usage

```
data(FilariasisSex)
```

Format

A data frame with 13 observations on the following 5 variables.

`Community` Community code

`Males_Examined` number of males examined

`Males_Infected` number of males infected

`Females_Examined` number of females examined

`Females_Infected` number of females infected

Details

Test the chi-square technique on the contingency table here.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(FilariasisSex)
```

Filariasistype *Filariasis and different parasites causing it*

Description

Filariasis is a common infection in tropical and subtropical countries. Several parasites can cause filariasis. In Nigeria a study was conducted to see prevalence of filariasis due to various parasite types. Specific questions of interest are: a) is the proportion of infected cases (in total number examined- column B) due to Onchocerca volvulus constant across communities? b) is the proportion of infected cases due to Onchocerca volvulus (out of total infected cases- column C) constant across communities?

Usage

```
data(Filariasistype)
```

Format

A data frame with 13 observations on the following 5 variables.

`Community` Community code

`Examined` Total number of individuals examined

`Infected` Total number infected

`Onchocerca_volvulus` Number of persons infected with parasite Onchocerca volvulus

`Others` Number of persons infected with other parasites

Details

Carry out the tests related to a contingency table.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Filariasistype)
```

`Fish` *Fish species interaction*

Description

Brook trout and cut-throat tout are two species of stream fish. If they compete with each other then high density of one may suggest absence of the other. If they are symbiotic, high density of one may promote the other. These data are from streams. Typical mountain streams are about 2.5m wide. (Kilograms per hectare is a conventional density measurement used in lakes) Can we predict presence/absence of Yellowstone Cutthroat trout as a function of density (kg/ha) of Brook trout?

Usage

```
data(Fish)
```

Format

A data frame with 24 observations on the following 2 variables.

`BKT` density of Brook trout (kg/ha)

`YSC` Presence/ absence of Yellowstone Cutthroat trout

Details

Use the logistic regression model.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Fish)
```

`fishtoxin` *Toxicity effect on fish*

Description

In toxicity studies, different doses of a toxic substance are tried and response is measured. In the present experiment, aflatoxin is mixed with water in the fish tank in five different doses. Response is development of tumor in fish. It is dichotomous. If the toxin has no effect then

Usage

```
data(fishtoxin)
```

Format

A data frame with 10 observations on the following 6 variables.

`Dose` Aflatoxin dose

`Alfatoxin` a factor with levels `total` `count with tumour`

`Tank_1` Count of fish with tumor growth and total for Tank 1

`Tank_2` Count of fish with tumor growth and total for Tank 2

`Tank_3` Count of fish with tumor growth and total for Tank 3

`Tank_4` Count of fish with tumor growth and total for Tank 4

Details

Build ANOVA after after a suitable transformation. Also consider logistic regression model.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(fishtoxin)
```

`Frogfood` *Study of growth and food preference over age in frogs*

Description

When an animal is to be harvested for food it is important to understand its growth pattern and energy requirement. In case of frogs it is of interest to study the relation age X body weight and fit a suitable curve. Further it appears that as age increases, the dietary pattern undergoes change. Hence it is also relevant to study relationship of age with each of 4 components of diet in variables of intake from crabs to total.

Usage

```
data(Frogfood)
```

Format

A data frame with 7 observations on the following 6 variables.

`Age` Age in year group

`Body_Weight` Body weight in grams

`Intake_Crabs` Intake of crabs in grams

`Intake_Insects` Intake of insects in grams

`Intake_Larvae` Intake of larvae in grams

`Total_intake` Total intake in grams

Details

Explore the use of linear regression, curvilinear regression, and examine their residuals.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Frogfood)
```

Frogmating *Relation between body size and number of mates for the frogs*

Description

In case of many animal species, males compete for females and there is uneven distribution of females among males. This can happen for two possible reasons. One is aggressive competition. Males may attack other males and force themselves on females in a locality. Walruses are seen to do this. The other possible reason is female choice. Females may have preference for some males over others and may be able to exercise option. In either case it is generally expected that large body size may help males in mating success. Is this true in case of bullfrogs? If so, what is the relationship between body size and the number of females attracted or dominated?

Usage

```
data(Frogmating)
```

Format

A data frame with 38 observations on the following 2 variables.

`Bode_Size` body size (mm) of male bull frog

`Mates` number of mates

Details

Poisson regression and logistic regression with ordinal response may be explored for this data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Frogmating)
```

Frog_survival *Fitting Ricker curve to frog survival data*

Description

In natural populations, due to mortality, the number of individuals decreases with age. One popular model for describing this phenomenon is the Ricker curve. The model may or may not give a good fit.

Usage

```
data(Frog_survival)
```

Format

A data frame with 8 observations on the following 2 variables.

`Age` Age in years

`Individuals` Number of individuals

Details

Logarithmic transformation and regression may be attempted on this data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Frog_survival)
```

GDS *Modeling Trends in Gross Domestic Savings*

Description

Savings are an important part of any economy. Savings can be invested as capital and that helps economic growth. It is known that Asian families are more inclined to saving compared to West European and North American families, which tend to borrow and spend. A low rate of saving would be below 10

Usage

```
data(GDS)
```

Format

A data frame with 57 observations on the following 5 variables.

`Year` Year

`Household_Sector` Savings in Household sector (Rs. Crores)

`Private_Corporate_Sector` Savings in Private corporate sector (Rs. Crores)

`Public_Sector` Savings in Public sector (Rs. Crores)

`Total` Total GDS (Rs. Crores)

Details

Time series, data transformation, and nonlinear regression may be considered for this data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(GDS)
```

`Geometricbirds` *Rank abundance distribution of bird species*

Description

4. One approach to biodiversity measurement is fitting a distribution to rank abundance data. In the present case data on abundance of different species is recorded at two localities in western-ghats in India. The species are ranked from most abundant (rank 1) to least abundant. Some ecological considerations suggest that a geometric distribution should fit the data. Parameter of the geometric distribution is taken as indicator of diversity. Higher the value of parameter, lower the diversity.

Usage

```
data(Geometricbirds)
```

Format

A data frame with 80 observations on the following 3 variables.

`Location` Location of the species

`Species_Rank` Species rank

`Abundance` number of birds seen of a species

Details

Begin with good ness of fit. Model builder aspirants can also try calculating the Simpson index and Shannon-Wiener index of species diversity.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Examples

```
data(Geometricbirds)
```

Heart *Comparison of Test drug with Placebo for Heart Attack*

Description

Heart attack is a major cause of death in modern society. Aim is to check if test drug reduces primary response. It is also of interest to check whether drug causes improvement in LDL and HDL levels relative to placebo. Possible role of age and sex needs to be taken into account. Similar comments apply to diabetes and hypertension.

Usage

```
data(Heart)
```

Format

A data frame with 205 observations on the following 8 variables.

`AGE` Age of the patient

`SEX` Gender of the patiend

`DIABETES` Diabetes indicator (you need to handle as a group)

`HYPERTENSION` Hypertension indicator

`LDL` Level of Low density lipid

`HDL` Level of high density lipid

`Primary_Response` number of occurrences of events such as death, hospitalization, second attack

`Drug` Drug indicator for placebo or treatment

Details

Contingency tables, t-test, ANOVA, and ANOCOVA may be built for analysis of this data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

`Highjump` *Guessing the gold medal score for 2004 Olympics*

Description

Olympics are being organized every 4 years for over a century. Performance of the gold medallist generally goes on improving with some cases of reversals. It is of interest to anticipate the results for the next tournament. It is good to remember that the observed values represent extreme cases.

Usage

```
data(Highjump)
```

Format

A data frame with 24 observations on the following 2 variables.

`Year` year

`Height` Height of jump in meters for the champion

Details

Regression and time series may be used for the prediction purpose.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

`hundredmrun` *Guessing the gold medal score for 2004 Olympics*

Description

Olympic games are organized every 4 years for over a century. Performance of the gold medallist generally goes on improving with some cases of reversals. It is of interest to anticipate the results for the next tournament. It is good to remember that the observed values represent extreme cases.

Usage

```
data(hundredmrun)
```

Format

A data frame with 24 observations on the following 2 variables.

`Year` Year

`Time.sec.` Time in seconds for the champion (male)

Details

Regression and time series are suggested for this problem.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

`IMR` *Changes in Infant mortality over last century across countries*

Description

Infant mortality is an important indicator of health status of a country. It is of interest to see how different countries have progressed in this aspect over last century, in particular if the relative ranks of countries remain roughly the same.

Usage

```
data(IMR)
```

Format

A data frame with 8 observations on the following 16 variables.

`Country` Name of the Country

`IMR_1900` IMR (per 1000 live births) in 1900

`IMR_1950` IMR (per 1000 live births) in 1950

`IMR_1985` IMR (per 1000 live births) in 1985

`IMR_1993` IMR (per 1000 live births) in 1993

Details

Experiment with comparison of trend and rank correlation for the data under consideration.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

`IOCSharePrice` *Modeling share price series of IOC*

Description

Time series is a sequence ordered in time. Its modeling is of considerable interest to users for purposes of forecasting.

Usage

```
data(IOCSharePrice)
```

Format

A data frame with 250 observations on the following 2 variables.

`Date` Date

`Opening_Price` Opening price

Details

Plotting, Trend fitting, Estimating seasonal component, ARIMA modeling, and Residual analysis are some of the powerful tools which may give rich insight into the share prices.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

`IslandSpArea` *Species area relationship*

Description

4. Island biogeography is a branch of ecology, which discusses evolution and spread of species on Islands. One of the theories is that as area of island shrinks the number of species of animals living on the island declines, but probably not linearly. One model suggested is $sp= k*A^z$ where A is area and k, z are constants.

Usage

```
data(IslandSpArea)
```

Format

A data frame with 16 observations on the following 2 variables.

`Area` Area of the island (Sq. Km)

`Species` Number of species seen

Details

Nonlinear regression and transformation of variables may improve the linear regression model.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

---

`Ivoryweight` *Trends in illegal ivory trade*

Description

Some people buy products made of ivory (elephant tusks). Hence there is temptation to kill tusker elephants illegally (poaching). There is international agreement that police should arrest such poachers as well as traders in ivory or products made from it. The intention is to reduce illegal killing of elephants so that their populations are preserved. The data give number of pieces and weight of ivory confiscated. It is of interest to see if there is a change in the ivory trade over the time, both in terms of pieces and in terms of weight. It is also of interest to check the relationship between number of pieces and weight. The analysis may be done for three stages raw, semi worked and worked separately or for total weight.

Usage

```
data(Ivoryweight)
```

Format

A data frame with 42 observations on the following 4 variables.

`Year` Year

`Ivory` The type of ivory

`Pieces` The number of ivory pieces recovered

`Weight` The total weight of the ivory pieces

Details

Time series, trend analysis, regression analysis need to be attempted for explaining the variation in this data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

`Lognormalbirds` *Species abundance distribution*

Description

One approach to biodiversity measurement is fitting a distribution to species abundance data. In the present case data on abundance of different species is recorded at two Indian localities in western-ghats (Arvind and Dhoni). These data are said to be 'Frequency of frequency'. Some ecological considerations suggest that a log-normal distribution should fit the data. S - the number of species observed and area under the curve to the left of 1/2 are used together to estimate the number of 'unseen' species.

Usage

```
data(Lognormalbirds)
```

Format

A data frame with 305 observations on the following 3 variables.

`Site` The sites

`Abundance` The number of species with abundance r each

`Species` The numberof individual birds sighted

Details

Report your findings with fitting log-normal distribution and Chi-square test for Goodness of fit.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

`Logseriesbirds` *Species abundance distribution*

Description

4. One approach to biodiversity measurement is fitting a distribution to species abundance data. In the present case data on abundance of different species is recorded at two Indian localities in western-ghats (Arvind and Dhoni). These data are said to be 'Frequency of frequency'. Here 14 species were encountered only once. There were 9 species such that exactly two individuals of each species were seen etc. Some ecological considerations suggest that a log-series distribution should fit the data. Parameter of the log-series distribution (called Fisher's $alpha$) is taken as indicator of diversity. Higher the value of parameter, higher is the diversity.

Usage

```
data(Logseriesbirds)
```

Format

A data frame with 179 observations on the following 3 variables.

`Site` The site

`Abundance` The number of individual birds sighted

`Species` The number of species with abundance r each

Details

Fitting Log-series distribution and Chi-square test for Goodness of fit are appropriate tools for this data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

---

`Loops` *Loops of the finger prints*

Description

It is believed that finger print patterns are inherited. The mode of inheritance is not known. One can reduce the data from three-way table to two-way table by collapsing the third dimension. It is expected that patterns on father's thumb are unrelated to those on mother's thumb. On the other hand patterns in child are expected to be related to patterns of either parent. One can also prepare 4*2 contingency tables. This can be used to test independence of child pattern from parent pattern combination. Another possible exploration is conditional independence. For example, is child pattern independent of mother pattern given father's pattern? Lastly, to test complete independence of all three dimensions, one must use the original 2*2*2 table.

Usage

```
data(Loops)
```

Format

A data frame with 8 observations on the following 4 variables.

`Child` Pattern on childs thumb

`Father` Pattern on fathers thumb

`Mother` Pattern on mothers thumb

`Frequency` The count

Details

MLE and chi-square test!

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

---

`Lung` *Smoking and Lung capacity study*

Description

That smoking causes cancer, is widely known. What is not so well known is that even when you escape cancer, chronic smoking can have an adverse effect on ability of lungs. In a clinical trial, a drug for improving lung capacity was administered in two localities to a number of patients. Lung capacity index was measured before and after treatment of one week. The objective is to test efficacy of the drug after discounting for differences between patients and localities.

Usage

```
data(Lung)
```

Format

A data frame with 41 observations on the following 8 variables.

`Serial_Number` Serial number

`Locality` Locality

`LCI_Before` Lung Capacity index (LCI) before treatment

`LCI_After` Lung Capacity index (LCI) after treatment

`Age` Age

`Gender` Gender

`Weight` Weight

`Smoking_Index` Smoking index (based on intensity and duration of smoking)

Details

ANOCOVA!

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

magazine *Time trends in authorship distribution*

Description

A monthly magazine in Marathi, wishes to examine trends in the distribution of total number of pages, into different forms of writing. The desire is that role of editorial writing and material from other members of the editorial board should decline. Share of articles by outsiders should increase as also reader's reactions.

Usage

```
data(magazine)
```

Format

A data frame with 14 observations on the following 8 variables.

`Year` Year

`Cover_Page` Cover page

`Editor` Editor

`Others_Editorial_Board` Others in Editorial board

`Articles` Articles

`Reprints_Marathi` Reprints in marathi

`Reprints_Other_Languages` Reprints in other languages

`Readers_Reaction` Reader's reaction

Details

Time series and regression may be experimented with.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

`Mammals` *Birth weight and brain size of mammals*

Description

Generally larger the animal larger is its brain size. But some species may be exceptional. The question of interest is 'does birth weight predict brain size?' (Data are averages for different mammal species). If there are any exceptional cases, find out which species they are.

Usage

```
data(Mammals)
```

Format

A data frame with 99 observations on the following 2 variables.

`Birth_Weight` Birth weight

`Adult_Brain_Weight` Adult Brain weight

Details

Regression model with a pre-cursor of transformation may be effective for this data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

`mammalsize` *Correlates of brain size for the mammals*

Description

Data are from American Naturalist (1974) p.593-613. Animals have properties that make them better capable of living and multiplying. One expects that larger brain may be generally better. But there can be penalties and limitations. One limitation is need for longer pregnancy and the other is the need to have fewer offsprings. The benefit must outweigh penalties. What are the characteristics associated with large brains? Generally, larger brain should go with larger body. What if we compare brain sizes after taking into consideration the body size? Are there any species that stand out?

Usage

```
data(mammalsize)
```

Format

A data frame with 96 observations on the following 5 variables.

`Species` Name of the species

`Gestation_Period` Gestation period (days)

`Brain` Brain weight (gms)

`Body` Body weight (kg)

`Litter_Size` Litter size

Details

Fit a regression analysis and perform the analysis of residuals to validate the model assumptions.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

---

`Mice` *Protein intake and lifespan of mice*

Description

Proteins play a very important role in metabolism of animals. Proteins are needed to build body tissue and to facilitate various physiological processes. Indeed animals deprived of proteins must eventually die. In this problem we try to relate level of protein intake with survival in a very simple way. If proteins are as essential as we state, then life span should be reduced if intake is too low. Will the lifespan of mice be different depending on whether they were on a very low protein diet or just low? Notice that unit for 'life span' is missing. What can it be? Minutes, Hours, days, weeks, months, years?

Usage

```
data(Mice)
```

Format

A data frame with 131 observations on the following 2 variables.

`Life_Span` Life span of mice

`Diet` Diet (1=Low protein, 0= very low protein)

Details

Two sample t-test?

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

`Microgrow` *Fit sigmoidal model to bacterial growth*

Description

It is of interest to develop a model for population growth, which may later be used for various purposes. Present data concerns growth of bacteria in a liquid medium. It is very difficult, if not impossible, to count the population of bacterial cells. Hence the same is measured indirectly through optical density of the medium. As population size increases, the medium becomes more opaque. Common model used to describe population growth is logistic.

Usage

```
data(Microgrow)
```

Format

A data frame with 61 observations on the following 2 variables.

`Time` Time since inoculation

`OD` Optical density (indicator of growth)

Details

Try out the 3 point method and non-linear regression.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

`Mimosaceae` *Relationship between tree height and girth*

Description

For various reasons, it is of interest to estimate volume of a tree. Obviously direct measurement is quite difficult. The next best thing is to assume that the tree has a conical shape. Even with this assumption we need to measure radius at the base and height for estimating volume. Measurement of height can be cumbersome. However, measurement of girth is easy at about one meter height. Hence the simplest method is to use some relationship between girth and height and estimate height from girth.

Usage

```
data(Mimosaceae)
```

Format

A data frame with 129 observations on the following 4 variables.

`Family` Family name

`Species_name` Species name

`GBH` Girth at breast height

`Height` Height (meters)

Details

Linear/ nonlinear regression for each species, testing hypothesis that the relationship is same across species.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

---

`moth` *Natural selection*

Description

The basic principle of Darwin's theory of evolution through natural selection is that as environment changes, ability of an organism to survive also changes. This was experimentally tested in and around Liverpool in United Kingdom. A moth species that comes in two varieties (light and dark) was used. Trees in Liverpool have blackened trunks due to industrial smoke. The darkness reduces as we go farther from the city. Dark moths can blend with the dark trunks and hence rate of predation is lower for this variety in the vicinity of Liverpool. As the distance of a locality from Liverpool increases and tree trunks become lighter, pendulum shifts in favor of the light variety. In the experiment in question, dead moths were left on tree trunks and were revisited after 24 hours. The number of moths removed (presumably by predators) was recorded. Question of interest is whether the proportion removed remains the same at all distances and if the null hypothesis is rejected, whether the removal rate increases (decreases) for dark (light) moths as distance of the site from Liverpool increases.

Usage

```
data(moth)
```

Format

A data frame with 14 observations on the following 5 variables.

`Site` Ssite number

`Distance` Distance from city

`Moth_Type` Moth type (light / dark)

`Numbers_of_Moths` Number of moths placed

`Removed_by_Predators` Number of moths removed by predators

Details

ANOVA with transformation and regression for proportions are the suggested tools for this data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

---

`OralCancer` *Comparison of two chemotherapy treatments for oral cancer*

Description

Oral cancer is one common type of cancer in India. Habits like tobacco chewing are supposed to increase the chance of this disease. Our aim is to assess the role of two treatments (A and B). Assessment must be done after discounting for covariates.

Usage

```
data(OralCancer)
```

Format

A data frame with 31 observations on the following 8 variables.

`Response` Response after treatment

`Age` Age (years)

`Gender` Gender

`Tobacco` Tobacco indicator

`Smoking` Smoking indicator

`Alcohol` Alcohol indicator

`History` History of surgical treatment

`Treatment` Treatment for oral cancer (A/B)

Details

First formulate a 2-way contingency table (treatment X response). Compare treatments. Now form three way contingency tables using one covariate at a time. Compare treatments in each case. Verify that a contingency table with 4 or more dimensions tends to become sparse and has many empty cells. Use Logistic regression to handle all covariates simultaneously.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Plaque *Studying effect of toothpaste on plaque accumulation*

Description

Accumulation of plaque on teeth is a major cause of dental ill health. Producers of toothpaste often claim that use of their product will reduce the plaque. Data on three products are to be used to check two things (a) does a particular product reduce plaque? (b) Is reduction achieved the same for all products?

Usage

```
data(Plaque)
```

Format

A data frame with 60 observations on the following 3 variables.

`Product` The product

`Before` Plaque score before treatment

`After` Plaque score AFTER treatment

Details

Paired t-test and one-way ANOVA may be attempted for this data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Plastic *Seasonality in sales of plastic granules*

Description

A manufacturer of plastic granules wants to study sales pattern over different months. Comparing different sales officers as well as two plants is also of interest. Any patterns that can suggest suitable action by management are of interest.

Usage

```
data(Plastic)
```

Format

A data frame with 1000 observations on the following 4 variables.

`Plant_Code` Plant Code 1 or 2

`Month` Month

`Employee_Code` Employee code

`Quantity` Quantity of sales (Metric Ton)

Details

Try out Histogram, ANOVA, and paired comparisons.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

---

`Poliocases` *The number of polio cases*

Description

Poliomyelitis is a crippling disease with dramatically visible impact on the patient. Over the last fifty years the disease has been brought under control by the use of oral vaccine. It is of considerable interest to identify trend, seasonality and other features of data on incidence of polio.

Usage

```
data(Poliocases)
```

Format

A data frame with 180 observations on the following 3 variables.

`Month` Month

`Year` Year

`Polio_Cases` Number of polio cases

Details

Time series analysis, what else?

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Preserve *Predicting fungal growth*

Description

It is customary to add preservatives for enhancing shelf life of processed foods. Common salts, sugar and oil are three preservatives widely used by housewives. In food processing industry, many chemical preservatives are used to prevent growth of fungus. In the present experiment this aspect is studied in a systematic way. Preservative is used in different quantities. pH and water activity level are two other factors that affect chance of fungal growth. It is of interest to delineate a 'safe zone', set of conditions under which probability of fungal growth is very small.

Usage

```
data(Preserve)
```

Format

A data frame with 60 observations on the following 4 variables.

`Preservative_Level` Preservative level

`pH` pH

`Water_Activity` Water activity level

`Response` Response (Growth=1, no growth=0)

Details

Logistic regression and contingency table may be explored for this data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Production *Quality control for examining consistency in weight*

Description

The data gives actual weights of a product (for a target weight of 150gm) in a manufacturing unit. Other relevant details like week, date, hour of the day are also given. It is of interest to check consistency in weights of the product with respect to these factors. The data are not balanced.

Usage

```
data(Production)
```

Format

A data frame with 670 observations on the following 7 variables.

`Week` Week number

`Date` Date (6 different dates)

`Hour` Hour of the day (1-23)

`Line_Number` Production line number (2 lines)

`Operator` Operator (5 operators)

`Actual_Weight` Actual weight (gm)

`Vendor` Vendor supplying raw material (3 vendors)

Details

One-way and multi-way ANOVA are suggested tools for this data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

---

`Pureforsure` *Detection of adulteration*

Description

Adulteration is a widespread problem for consumers in India. We have to examine the possibility of adding a marker in very small quantity to original produce. Adulteration will reduce the proportion of this marker. Suppose there is a machine that can test proportion of marker in any sample. It is to be used in the field to detect malpractice. The machine itself invariably contributes to measurement error. 200 values in column B are results of measurements on known pure samples. Ideal value is 100. Deviations from it in these cases indicate the extent of error that the machine commits. Aim is to decide a cut off value of machine reading such that value below it will be treated as evidence of adulteration. Each such choice has two kinds of errors associated with it. Optimum value is to be arrived at.

Usage

```
data(Pureforsure)
```

Format

A data frame with 200 observations on the following 2 variables.

`Day` Day

`Reading` Machine Reading

Details

Analyze with Histogram, time plot of machine readings, and empirical probabilities of type I and type II errors.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

---

Rabbit *Relating Foot length to Body mass*

Description

Birds of prey such as eagles or falcons catch a prey animal and then take it to a suitable place (nest if young ones are to be fed), remove inedible portion (feathers or bones etc) and then eat the meat. Remaining material is dropped to the ground. If some indigestible material is swallowed, it is formed into a ball and regurgitated. All such residues dropped down by the predator can provide good clues to its dietary patterns. Rabbits are common prey of eagles. One inedible portion of a rabbit's body is a hind foot. An ornithologist studied the relationship between foot length and total body mass in rabbits. The idea was that by observing the foot length, one might be able to estimate the meat intake.

Usage

```
data(Rabbit)
```

Format

A data frame with 141 observations on the following 2 variables.

`Hind_Foot_Length` Rabbit Hind Foot length in inches

`Body_Weight` Body Weight in pounds

Details

Regression model and analysis of residuals need to be performed on this data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Rat *Study of rat burrow architecture*

Description

Bandicoot rats live in underground burrows dug by them. 83 burrows were excavated and measured. However, by accident, only the marginal distributions were retained while the original data on joint distribution was lost. Check whether each marginal distribution is normal. It is of interest to estimate proportion of burrows having length greater than average AND depth greater than average. Use the following formula for generating bivariate distribution from marginals.

Usage

```
data(Rat)
```

Format

A data frame with 6 observations on the following 4 variables.

`Tunnel_Length` Total length of tunnel (cm)

`Frequency` Frequency

`Tunnel_Depth` Depth of tunnel (cm)

`Frequency.1` Frequency of tunnel depth

Details

Use the chi-square test for checking univariate normality.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

RiceWheat *Modeling Rice and Wheat production*

Description

At the time of independence India was faced with food shortages. There was dependence on food imports from USA etc. Later, because of green revolution, production went up and the country became self-sufficient. Was this growth in production linear? Was it due to increase in total area under crop or area under irrigation? Is the pattern of growth comparable in wheat and rice?

Usage

```
data(RiceWheat)
```

Format

A data frame with 106 observations on the following 6 variables.

`Food` The type of food

`Year` The year

`Area` Area (million Hectares)

`Production` Production (Million Tons)

`Yield` Yield (Kilogram/Hectare)

`Irrigated` Percentage area for the food type covered by irrigation

Details

Model this data set with time series analysis and regression.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

---

`salamander` *Habitat preference of salamander*

Description

Salamanders are amphibians, a group of animals of special interest because of their sensitivity to environmental changes. A particular species (about 6cm in length) found in California was studied to check the habitat preferred by the animal.

Usage

```
data(salamander)
```

Format

A data frame with 47 observations on the following 4 variables.

`Site` Site number

`Salamander` Salamander count in a sample plot of 7meter by 7 meter area

`Coverage` Extent of canopy cover in the forest

`Forest_Age` Forest age

Details

Anybody for Poisson regression?

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Sheeplife *Fitting probability distribution to life data of Sheeps*

Description

In biology often it is of interest to describe life span of an organism. Parametric approach to this problem involves fitting a given probability distribution. Common distributions used here are exponential, Weibul and Gamma. Occasionally other distributions such as inverse Gaussian, log-normal distribution etc. are also fitted. The data refer to age at death of a species of sheep. Age was estimated using skulls. In case of the above data usual distributions failed to give a good fit. There is considerable improvement by using, instead, a quadratic hazard function.

Usage

```
data(Sheeplife)
```

Format

A data frame with 11 observations on the following 2 variables.

`Age_at_death` Age at death (years)

`Frequency` Frequency

Details

Estimate parameters by maximum likelihood and test goodness of fit using chi-square test.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

SholapurWeather *Has the weather in Sholapur changed over 3 decades?*

Description

4. Records of maximum and minimum temperature are available with the India Meteorology Department. Data set for Sholapur, Maharashtra used here covers only 4 years: 1945, 1955, 1965 and 1972. Question of interest is "has the Sholapur weather changed?"

Usage

```
data(SholapurWeather)
```

Format

A data frame with 1461 observations on the following 5 variables.

`YEAR` Year

`DATE` Date

`MONTH` Month

`MAXT` Maximum temperature

`MINT` Minimum temperature

Details

Host of options here: Descriptive statistics, Comparison of 4 Time series. Take month as blocks and compare years using Friedman test. Variable can be (say) max temperature of the month.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

---

`Sorghumheight` *Modeling sorghum plant growth*

Description

It is generally believed that a sigmoidal model is good for describing growth of plant height. In the present data set we have growth records for two varieties of sorghum- a cereal crop in western India. It is of interest to check if the model parameters for two varieties are equal.

Usage

```
data(Sorghumheight)
```

Format

A data frame with 22 observations on the following 3 variables.

`Day` Age (days from sowing)

`Ramkel` Plant Height (Variety 1)

`Saoner` Plant Height (Variety 2)

Details

Non-linear regression and LR-test need to performed for this data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

SpaccHerb *Species accumulation curve*

Description

One index of diversity is species richness i.e. total number of species of a taxon present in a locality. To estimate richness, it is convenient to go on counting randomly selected individuals and recording their species names. An accumulation curve has total number of individuals or quadrats on X-axis and accumulated number of species on Y axis.

Usage

```
data(SpaccHerb)
```

Format

A data frame with 922 observations on the following 4 variables.

`Serial_Number` Serial number

`Quadrat_Number` Quadrat number

`Species_Name` Species name

`Individuals` Number of individuals

Details

Generating species accumulation data by (a) quadrat, (b) number of individuals. Fitting a nonlinear (saturating) model to these data to arrive at estimate of saturating value.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

SpaccShrubs *Species accumulation curve*

Description

One index of diversity is species richness i.e. total number of species of a taxon present in a locality. To estimate richness, it is convenient to go on counting randomly selected individuals and recording their species names. An accumulation curve has total number of individuals or quadrats on X-axis and accumulated number of species on Y axis.

Usage

```
data(SpaccShrubs)
```

Format

A data frame with 98 observations on the following 4 variables.

`Serial_Number` Serial number

`Quadrat_Number` Quadrat number

`Scientific_Name` Species name

`Individuals` Number of individuals

Details

Generating species accumulation data by (a) quadrat, (b) number of individuals. Fitting a nonlinear (saturating) model to these data to arrive at estimate of saturating value.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

`Spaceshuttle` *Modeling Space shuttle O-ring failure data*

Description

Space shuttle Challenger exploded right at the beginning of its flight on January 28, 1986. This was one of the largest disasters in the American space program. The night before, an engineer had recommended to NASA (National Aeronautics and Space Administration) that the shuttle should not be launched in the cold weather. Forecast of temperature for the launch was 31 degrees Fahrenheit, the coldest launch ever. This suggestion was over-ruled. Inquiry Commission appointed by the President of the United States, wanted to see if enough evidence existed to predict serious trouble due to low temperature at the time of launch. Since the shuttle had, up to that time, not met with any accident, the only evidence available was regarding damage to O-rings. These rubber rings fill the gaps between parts of the giant tube that makes the rocket. If there is even a minor leak, hot gases push through it and in milliseconds, large portion of the rocket fuel can come out to destroy the rocket. Hence damaged O-rings can be treated as signs of major trouble. Such instances had indeed been recorded in previous flights of the shuttle. Data are to be analyzed to check if statistical methods would have given the right guidance.

Usage

```
data(Spaceshuttle)
```

Format

A data frame with 24 observations on the following 2 variables.

`Launch.temperature` Launch temperature

`Rings_damaged` Number of O rings damaged

Details

Plotting techniques with further validation using logistic regression analysis is expected to explain this phenomenon.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

---

`Spareabirds` *Species area curve*

Description

A standard observation in ecological fieldwork is that as the area scanned increases, the number of species of any given taxon (here birds) seen in that area increases but at a decreasing rate. Such empirical relationship can be exploited to estimate total number of species (species richness). For more details refer the web link indicated below.

Usage

```
data(Spareabirds)
```

Format

A data frame with 24 observations on the following 3 variables.

`Region` Region

`Area` Area in Sq. Km.

`Species` Number of bird species counted

Details

Nonlinear regression seems appropriate for the data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

StemDensity *Vegetation types and tree density*

Description

Forests are an important renewable natural resource of our society. They conserve water and soil, moderate temperature and provide fuel, fodder, fiber, fertilizer etc. Forests give us timber and medicinal plants. Some uses of forests can be evaluated in money terms rather easily. One such item is timber. Three things decide market value of timber. Tree species (teak and rosewood are very valuable), volume (market price is per unit volume) and dimension of log (larger planks fetch higher price per unit volume). Hence foresters' inventories of stock include tree count by species and girth class. It is relevant to summarize such information using probability distributions. Ecologists are interested further in monitoring the variety of trees as judged by counts in different vegetation types.

Usage

```
data(StemDensity)
```

Format

A data frame with 11 observations on the following 9 variables.

`Girth_Class` Girth class (cm)

`Evergreen` the number of trees in Evergreen forest belonging to this girth class

`Semi_evergreen` number of trees in Semi Evergreen forest belonging to this girth class

`Moist_Deciduous` number of trees in Moist Deciduous forest belonging to this girth class

`Littoral` number of trees in Littoral forest belonging to this girth class

`Bamboo` number of trees in Bamboo forest belonging to this girth class

`Mangrove` number of trees in Mangrove forest belonging to this girth

`Padauk` number of trees in Padauk forest belonging to this girth class

`Teak` number of trees in Teak forest belonging to this girth class

Details

Fitting of distributions with group data, comparison of parameters need to be performed for this data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

`TeethNormal` *Modeling indicators of dental health*

Description

There are three widely accepted indicators of dental health. They are gingival score, bleeding (gums) score and plaque score. It is always of interest to check effect of toothpaste on these variables. To decide which methods of statistical analysis would be appropriate, it is relevant to test normality of these measurements.

Usage

```
data(TeethNormal)
```

Format

A data frame with 69 observations on the following 3 variables.

`Gingival` Gingival score

`Bleeding` Bleeding (of gums) score

`Plaque` Plaque score

Details

Effective use of histogram, P-P plot, and goodness of fit may provide the answers.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

`Tiger7` *Identification of individual tigers from pugmarks*

Description

Refer Tiger7.doc for precise explaination of the data set.

Usage

```
data(Tiger7)
```

Format

A data frame with 78 observations on the following 7 variables.

`P_TC1_Dist` Distance between Pad center and Toe 1 center

`P_TC2_Dist` Distance between Pad center and Toe 2 center

`P_TC3_Dist` Distance between Pad center and Toe 3 center

`P_TC4_Dist` Distance between Pad center and Toe 4 center

`TC1_TC2_Dist` Distance between Toe1 center and Toe 2 center

`TC2_TC3_Dist` Distance between Toe2 center and Toe 3 center

`TC3_TC4_Dist` Distance between Toe3 center and Toe 4 center

Details

Cluster analysis is recommended for gaining an insight into this problem.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

---

`TigerIdentity` *Tiger census using scat samples*

Description

Tiger census is a long debated issue in wildlife conservation. Since direct count is not possible any indirect method is based on similarities/ discrepancies in sample evidence. Two methods, which have received attention, are (i) pugmark counting and (ii) camera trap method. In pugmark method, if two pugmarks appear similar in shape and size, they are supposed to represent same animal. In Camera trap, if two photographs show similar stripe pattern, they represent same tiger. It now appears that a third method is possible. It is based on parasite composition of scat samples. If the composition is similar, most likely the two samples come from same animal.

Usage

```
data(TigerIdentity)
```

Format

A data frame with 55 observations on the following 33 variables.

`ID` Sample ID

`Type.1` Abundance of various pathogen types, 1 to 32, in the sample

`Type.2`

`Type.3`

`Type.4`

`Type.5`

`Type.6`

`Type.7`

`Type.8`

`Type.9`

`Type.10`

`Type.11`

`Type.12`

`Type.13`

`Type.14`

`Type.15`

`Type.16`

`Type.17`

`Type.18`

`Type.19`

`Type.20`

`Type.21`

`Type.22`

`Type.23`

`Type.24`

`Type.25`

`Type.26`

`Type.27`

`Type.28`

`Type.29`

`Type.30`

`Type.31`

`Type.32`

Details

Cluster analysis is again recommended.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Timber *Genetic and environmental components of tree characteristics*

Description

Refer Timber.doc from the web link given below for a detailed description of the problem.

Usage

```
data(Timber)
```

Format

A data frame with 224 observations on the following 10 variables.

`Locality` Locality

`Year` Year of experiment (two years, 2000 and 2002)

`Replicate` Replicate

`Subculture` Subculture

`Elongation` Percent of Elongation

`Multiples` Percent of Multiples

`Rooting` Percent of rooting

`Germination` Percent of germination

`Seed_Length` Seed length in cm

`Seed_Width` Seed width in cm

Details

ANOVA, Variance component analysis, and PCA are recommended for the data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Valvefailure *Valve characteristics and numbers of failures in a nuclear reactor*

Description

Here the number of failures in a run is a Poisson variable with mean affected by the various factor combinations and also observation times. There are 5 factors. Note that all explanatory variables are of qualitative type and in regression they will have to be converted into a group of indicator columns each. If there are 2 categories a single indicator column suffices. If there are k categories we need k-1 columns.

Usage

```
data(Valvefailure)
```

Format

A data frame with 90 observations on the following 7 variables.

`System` System (1=containment, 2= nuclear, 3=power conversion, 4= safety, 5= process auxiliary)

`Operator` Operator type (1= air, 2= solenoid, 3=motor driven, 4= manual)

`Valve` Valve type (1=ball, 2= butterfly, 3=diaphragm, 4= gate, 5= globe, 6= directional control)

`Size` Head size (1= less than 2inches, 2= 2-10 inches, 3= 10-30 inches)

`Mode` Operational mode (1= normally closed, 2= normally open)

`Failures` Number of failures

`Time` Observation times (multiplied by 43800) hours

Details

Poisson regression is recommended for this data set.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

`Waterquality` *Water quality analysis using clustering*

Description

Water samples were collected from 4 cities. The physico-chemical properties were measured. It is of interest to compare water in different cities. It is also intended to try cluster analysis after ignoring the city label (are original groups identified?). If water quality is similar then water treatment can be similar.

Usage

```
data(Waterquality)
```

Format

A data frame with 63 observations on the following 10 variables.

`City` a factor with levels `City1` `City2` `City3` `City4`

`pH` pH

`Conductivity` Conductivity

`Total_Dissolved_Solid` Total dissolved solid

`Alkalinity` Alkalinity

`Hardness` Hardness of the water

`Calcium_Hardness` Calcium hardness

`Magnesium_Hardness` Magnesium hardness in the water

`Chlorides` Chlorides

`Sulphates` Sulphates

Details

Hotelling's T2, MANOVA, Cluster analysis may be used here.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

`widowbird` *Mate selection by females*

Description

Refer the document titled widowbirdexpt.doc obtainable from the link given below.

Usage

```
data(widowbird)
```

Format

A data frame with 36 observations on the following 5 variables.

`New_Tests` Number of new nests

`Treatment_Group` Treatment group

`Male` ID of the male bird

`Tail` Tail length cm.

`Prev` Previous nest count

Details

ANOCOVA may be tried out here.

Source

http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Index