Summary Document for the Data Sets

Well done, you clicked on the “readme” file, I like your style.

This document gives you a summary of all the data sets.

**The best use of this document is to look at the “Suggested uses” column and choose a data set suitable for what you wish to teach!**

All the data sets come from here: <https://www.sheffield.ac.uk/mash/statistics/datasets>. I save it here for posterity.

Each folder has (i) A csv file for the data set & (ii) A word document with further information.

|  |  |  |
| --- | --- | --- |
| **Dataset** | **General description** | **Suggested uses** |
| Awards | The number of awards earned by students at one high school. Predictors of the number of awards earned include the type of program in which the student was enrolled (e.g., vocational, general or academic) and the score on their final exam in math. This data set is used to show an example of the Poisson Regression. The predicted variable is the number of awards and the predictors are the program type and the Maths score. | Descriptive statistics  Skewed data  Kruskall-Wallis test  Poisson regression |
| Birthweight | This dataset contains information on new born babies and their parents. It contains mostly continuous variables (although some have only a few values e.g. number of cigarettes smoked per day) and is most useful for correlation and regression. The birthweights of the babies who mothers smoked have been adjusted slightly to exaggerate the differences between mothers who smoked and didn’t smoke so students can see the difference more clearly in a scatterplot with gestational age and scatter colour coded by smoking status | Descriptive statistics  Recoding & creating new variables  T-tests  Chi-squared tests  Scatterplots & correlation  Simple, multiple & logistic regression  Cluster analysis  Factor analysis/Principal components |
| Cholesterol | A study tested whether cholesterol was reduced after using a certain brand of margarine as part of a low fat, low cholesterol diet. The subjects consumed on average 2.31g of the active ingredient, stanol easter, a day. This data set contains information on 18 people using margarine to reduce cholesterol over three time points. The data set can be used to demonstrate paired t-tests, repeated measures ANOVA and a mixed between-within ANOVA using the final variable ‘Margarine’. The dataset is also good for discussion about meaningful differences as the difference between weeks 4 and 8 is very small but significant | Descriptive statistics  Recoding and computing new variables  T-tests  ANOVA: within groups & repeated measures |
| Crime | This data set gives a variety of variables by US state at two time points 10 years apart. A variety of regressions and t-tests can be carried out with the main scale dependent being and Crime Rate (offences per million population) and t-tests with the independent being whether or not the state is in the south. Mostly discrete variables as they measure populations per 1000, there are some continuous variables such as those measuring expenditure | Descriptive statistics  Skewed data  Recoding and computing new variables  T-tests  Mann-Whitney U test  Scatterplots and correlation  Simple linear and multiple regression  Chi-squared test  Cluster analysis |
| Diet | This data set contains information on 78 people using one of three diets. The dataset is primarily used for ANOVA. | Descriptive statistics  Recoding and creating new variables  T-test  ANOVA: one & two-way  ANCOVA |
| Graduate | A study looks at factors that influence the decision of whether to apply to graduate school. College juniors are asked if they are unlikely, somewhat likely, or very likely to apply to graduate school. Hence, our outcome variable has three categories. Data on parental educational status, whether the undergraduate institution is public or private, and current GPA is also collected. The researchers have reason to believe that the “distances” between these three points are not equal. For example, the “distance” between “unlikely” and “somewhat likely” may be shorter than the distance between “somewhat likely” and “very likely”. | Descriptive statistics  Mann-Whitney U test  Ordinal regression |
| Ice-cream | This page shows an example of a multinomial logistic regression analysis with footnotes explaining the output. The data were collected on 200 high school students and are scores on various tests, including a video game and a puzzle. The outcome measure in this analysis is the student’s favorite flavor of ice cream – vanilla, chocolate or strawberry- from which we are going to see what relationships exists with video game scores (video), puzzle scores (puzzle) and gender (female). | Descriptive statistics  Normal & skewed data  Chi-squared test  Mann-Whitney U test  Kruskall-Wallis test  Scatterplots & correlation  Simple linear regression  Multinomial logistic regression |
| Normal | This data set contains 2 continuous variables where one is an example of normally distributed data and the other one is an example of skewed data. | Normal & skewed data  Scatterplots |
| Smoker | A researcher wants to investigate the impact of an intervention on smoking. In this hypothetical study, 50 participants were recruited to take part, consisting of 25 smokers and 25 non-smokers. All participants watched an emotive video showing the impact that deaths from smoking-related cancers had on families. Two weeks after this video intervention, the same participants were asked whether they remained smokers or non-smokers. | Descriptive statistics  McNemar's test |
| Titanic | The ship Titanic sank in 1912 with the loss of most of its passengers. Details can be obtained on 1309 passengers and crew on board the ship Titanic. The main use of this data set is Chi-squared and logistic regression with survival as the key dependent variable. Summary statistics for the categorical variables can be demonstrated and the cost of the ticket (fare) is very skewed so it can be used to demonstrate skewed data and differences between means and medians etc. | Descriptive statistics  Skewed data  Recoding and creating new variables  Chi-squared test  Mann-Whitney U test  Kruskall-Wallis test |
| Video | This dataset was collected by Scott Smith (University of Sheffield) to evaluate the use of best method for informing the public about a certain medical condition. There were three videos (New general video A, new medical profession video B, the old video C and a demonstration using props D). He wanted to see if the new methods were more popular so collected data using mostly Likert style questions about a range of things such as understanding and general impressions. This reduced dataset contains some of those questions and 4 scale scores created from summing 5 ordinal questions to give a scale score. | Descriptive statistics  T-test  Mann-Whitney U test  ANOVA: One -way  Kruskall-Wallis test  Friedman test  Correlation |