R Tutorials for Applied Statistics

1 Introduction to Data Analysis

The tutorials below get you started with the R language and data and statistical analysis. These tutorials are a good place to start for someone who has little to no familiarity with these concepts. In the tutorials below, you will learn how we think about and organize data and how we make some basic statistical summaries and inferences.

Introduction to Data
Here you get your first introduction to using R to explore a data set. The tutorial applies a real data set to examine data analysis concepts including what are variables, what are observations, what scales we use to measure variables, and how we can summarize data.

Estimating the Sample Mean
Here we look at a commonly understood measure of center, the mean. We learn here how to make statistical inferences on the population mean based on a sample mean.

Median and the Interpolated Median
Here we examine another common measure of center, the median. We look at two measures of the median, the unadjusted median, i.e. the middle observation, and the interpolated median, which is an adjusted median to give more precision and information on where the distribution is truly centered.

Estimating the Sample Median
Here we look at how to make statistical inferences on the population median based on sample evidence.

2 Comparing Means and Medians

Comparing Means for Independent Samples

Here we look at how to make statistical inferences on the differences between means from two independent groups. Independent groups imply there are different sampling units (ex: different people) in each of the two groups.

Comparing Medians for Independent Samples
Here we look at how to make statistical inferences on the differences between medians from two independent groups. Independent groups imply there are different sampling units (ex: different people) in each of the two groups.

Comparing Means for Paired Samples
Here we look at how to make statistical inferences on the differences between means from two paired groups. Paired groups imply each group has the same sampling units (ex: same people are measured twice) .

Comparing Medians for Paired Samples
Here we look at how to make statistical inferences on the differences between medians from two paired groups. Paired groups imply each group has the same sampling units (ex: same people are measured twice) .

3 Bivariate Relationships

Estimating Correlation
Here we look at estimating and visualizing how two numerical variables are positively or negatively related.

Chi-Square Test of Independence
Here we look at how estimate the relationship between two categorical variables.

4 Data Visualization

Grammar of Graphics
In this first tutorial, we are introduced to the grammar of graphics, which is a framework for understanding how to think about and build data visualizations. The package `ggplot` is an R plotting package that uses a programming strategy consistent with this framework.

Bar Plots to Illustrate Means

Here we create bar plots to illustrate and compare means. We also produce error bars to make statistical inferences in our plots. Finally, we also use color and facets to introduce additional categorical variables into our plots.

Visualizing Tables and Cross Tabulations
In this tutorial, we examine methods for illustrating proportions of a sample that fall into multiple categories. Cross tabulations can be used to determine if two categorical variables are related to one another.

Anscombe's Quartet
Frank Anscombe illustrated in his 1973 American Statistician **paper** that correlation coefficients are not enough to communicate correlation. Visualizations are also necessary. We recreate his example using `ggplot`.

Bar Plots to Illustrate Medians
We introduce the concept of the interpolated median, an adjusted measure of median to more precisely measure the center of a distribution. We create bar plots of the interpolated median and use error bars to explore evidence for statistically significant differences.

5 Introduction to Regression

Bivariate Regression
In its most simple form, a regression estimates a linear relationship between an explanatory variable and an outcome variable. In this tutorial, we explore the relationship mathematically, graphically, and learn how to estimate it.

Estimating and Interpreting Coefficients
Given an estimate for a regression line, we delve further into the meaning and interpretation of the regression coefficients and their statistical significance.

Non-Linearities in Regression
In its most simple form, a regression looks at a linear relationship between two variables. Here, we explore how to allow for non-linear relationships using logarithms for the explanatory and/or outcome variables. We explore graphically what such structures imply and we delve into the new meaning of the coefficients with such a structure.

6 Multiple Regression

Introduction to Multiple Regression

Multiple regression analysis extends simple bivariate regression analysis with the inclusion of more than one explanatory variable. The procedure is used to determine how one or more explanatory variables influence an outcome variable, while holding all other explanatory variables fixed.

Variance Decomposition

We use the regression model to break down how variability in the outcome variable is explained by variability in the explanatory variables, and how much is unexplained. We use these measures to construct estimates for how well the model fits or explains the data.

Standardized Regression

While we have an understanding of how to interpret individual regression coefficients, in this tutorial we examine a method for estimating a regression to allow us to compare regression coefficients. The method may also allow for more convenient interpretations of individual regression coefficients.

General Linear Restrictions

Sometimes we are interested in testing more complicated combinations of regression coefficients, rather than only looking at a single regression coefficient. Here we learn to generally compare regression models with no restrictions, relative to regression models with one more more restrictions imposed.

Multicolinearity

Multicolinearity is a problem when two or more explanatory variables are related to one another so that there is limited ability to differentiate the effect of one variable versus another. We learn how to examine evidence for multicolinearity and learn a strategy for gaining insights from the data when it is present.

Interaction Effects

Two different explanatory variables may each affect an outcome variable. Sometimes the two variables acting jointly have a different affect than the sum of each individually. These are interaction effects. Many interesting insights can be found using interaction effects. We learn how to incorporate this into the regression model and how to interpret the results.

Linear Combinations

With interaction effects, a variable in a regression can appear after more than one coefficient, and we need to draw statistical inferences on the magnitude of a combination of regression coefficients in order to understand the effect of a single variable. We walk

through the estimation and interpretation of an example.

Dummy Variables

You're a dummy variable! Dummy is not an insult to a variable, but a description of when a variable is a 0 or 1, i.e. a categorical variable for an observation having a trait or not having a trait. Dummy variables can enter as an explanatory variable rather easily. When we introduce interaction effects with explanatory variables, we can make some interesting insights, but how we interpret the coefficients takes great care. We walk through an example.

7 Heteroskedasticity

Introduction to Heteroskedasticity

Heteroskedasticity is the property when the variance of the error term changes predictably with one or more of the explanatory variables. In all the previous tutorials, we implicitly assumed homoskedasticity (when the variance of the error term is constant). We learn how to detect heteroskedasticity and account for it in our estimation procedure.

Inference with Heteroskedasticity

Here we learn how to conduct hypothesis tests and confidence intervals on coefficients when there is heteroskedasticity.

8 Binary Dependent Variables

Binary/Dummy Variables

Binary/dummy variables are variables that take on a value of 0 or 1 indicating whether or not an observation has some characteristic. We explore about how we can use binary variables in the same way as numeric variables, or example taking the mean and estimating hypothesis tests and confidence intervals, and what that means.

Linear Probability Model

Binary/dummy variables can also be the outcome/dependent variable. In such a model, the goal is to predict whether or not a some outcome will occur and possibly identify explanatory variables that affect the probability of the outcome. The linear probability model is one of multiple approaches to estimate regression models with dummy dependent variables. It simply replaces the usually continuous y-variable with the 0-1

dummy variable. This causes heteroskedasticity, though, so we account for this problem accordingly. We discuss some benefits and problems to using this approach for dummy dependent variables.

Logistic Regression

Logistic regression is another approach for estimating models with a binary/dummy dependent variable. It addresses some of the problems associated with the linear probability model, but has its own limitations. We walk through an example for how to estimate the model and make inferences about the marginal effects of the explanatory variables.

9 Panel Regression

Pooled Panel

A panel regression is when you estimate a regression on a panel data set. Panel data sets have multiple individuals/sampling units and each individual has multiple measurements across time. Therefore, you have data across individuals and across time. In this tutorial, we learn how to estimate and interpret the most simple framework for a panel data set.

Fixed Effects Regression

Fixed effects regression allows one to make inferences in regression while taking to account any variable, measurable or not, identifiable or not, that affects the individual but not over time. An effect on the individual but is fixed over time, is the "fixed effect."

Difference-in-Difference Regression

Like fixed-effects regression, this technique again allows for inferences on causation, taking into account any variable, measurable or not, identifiable or not, that does not change with time. The process is essentially comparing a treatment and control group (difference), before and after (another difference) some treatment.

10 Introduction to Forecasting

Introduction to Forecasting

Time series data is data on one individual or sampling unit, but with many observations across time. Most economic and financial data is time series data. In this introductory tutorial, we explore graphically and statistically, some common features of time series

data. Finally, we introduce some common statistical models that are used to forecast future outcomes for time series variables.