R-bloggers

R news and tutorials contributed by (750) R bloggers

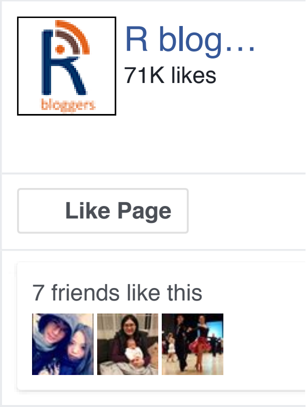
- Home
- About
- RSS
- add your blog!
- Learn R
- R jobs���
- Contact us

Welcome!

Follow @rbloggers 57.7K

Here you will find daily **news and tutorials about R**, contributed by over 750 bloggers.
There are many ways to **follow us -**
By e-mail:

On Facebook:



R blog…
71K likes

Like Page

7 friends like this

If you are an R blogger yourself you are invited to add your own R content feed to this site (**Non-English** R bloggers should add themselves- here)

Jobs for R-users

- Data Scientist @ Garching bei München, Bayern, Germany
- Software Developer
- Senior Quantitative Analyst, Data Scientist
- R data wrangler
- Senior Data Scientist

Search & Hit Enter

Popular Searches

- googlevis
- heatmap
- twitter
- latex
- sales forecasting
- sql
- web Scraping
- eof
- hadoop
- Jeff Hemsley
- random forest
- 3 d clusters
- anova
- blotter
- boxplot
- coplot
- decision tree
- discriminant
- financial
- ggplot background grid colour
- how to import image file to r
- maps
- purrr
- rattle
- Trading
- bar chart
- barplot
- Binary
- climate
- contingency table data frame

Recent Posts

- R Interface to Spark
- Data Science for Business – Time Series Forecasting Part 2: Forecasting with timekit
- Run massive parallel R jobs cheaply with updated doAzureParallel package
- Introduction to Set Theory and Sets with R
- Campaign Response Testing no longer published on Udemy
- Neural networks Exercises (Part-1)
- Introducing the MonteCarlo Package
- Words growing or shrinking in Hacker News titles: a tidy analysis
- Test-driving Microsoft Cognitive Toolkit in R using reticulate
- Deep Learning with R
- Add P-values and Significance Levels to ggplots
- UK R Courses

- Unconf projects 4: cityquant, notary, packagemetrics, pegax
- Tic Tac Toe Part 3: The Minimax Algorithm
- What is the tidyverse?

Other sites

- SAS blogs
- Jobs for R-users

Chi-Squared Test

August 14, 2016
By Selva Prabhakaran

Like 2 Share Share

(This article was first published on **DataScience+**, and kindly contributed to R-bloggers)

147 SHARES Share Tweet

Before we build stats/machine learning models, it is a good practice to understand which predictors are significant and have an impact on the response variable.

In this post we deal with a particular case when both your response and predictor are categorical variables.

By the end of this you'd have gained an understanding of what predictive modelling is and what the significance and purpose of chi-square statistic is. We will go through a hypothetical case study to understand the math behind it. We will actually implement a chi-squared test in R and learn to interpret the results. Finally you'll be solving a mini challenge before we discuss the answers.

Background knowledge – Predictive modelling

For the sake of completeness, let's begin by understanding how predictive modelling works so you can better appreciate the significance of chi-squared tests and how it fits in the process.

Predictive modelling is a technique where we use statistical modelling or machine learning algorithms to predict a response variable/s based on one or more predictors.

The predictors are typically features that influence the response in some way. The models work best if the features are meaningful and have a significant relationship with the response.

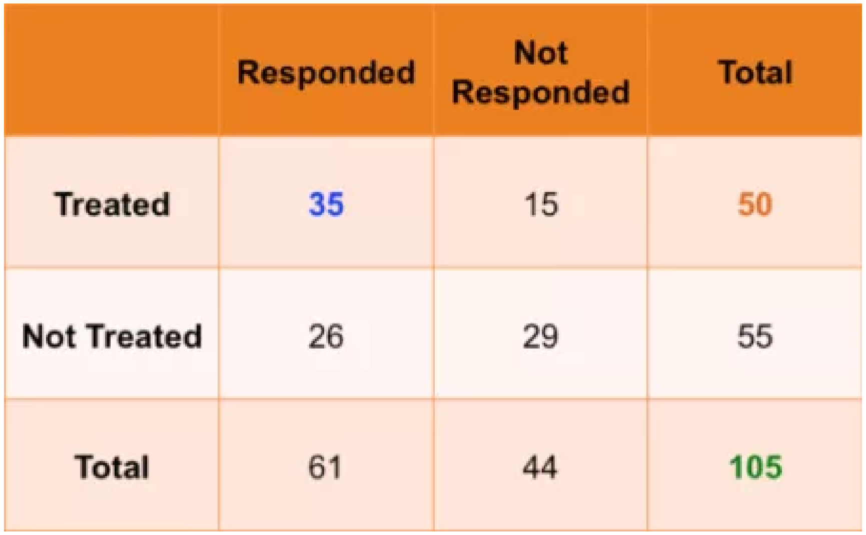
But, you normally wouldn't know beforehand if the response is dependent on a given feature or not. We can use the chi-squared test to determine if they are dependent or not, provided, both response and predictors are categorical variables.

Hypothetical Example: Effectiveness of a Drug Treatment

Let's consider a hypothetical case where we test the effectiveness of a drug for a certain medical condition.

Suppose we have 105 patients under study and 50 of them were treated with the drug. The remaining 55 patients were kept as control samples. The health condition of all patients was checked after a week.

The following table shows if their condition improved or not. Just by looking at it, can you tell if the drug had a positive effect on the patients.



	Responded	Not Responded	Total
Treated	35	15	50
Not Treated	26	29	55
Total	61	44	105

As you can see, 35 out of the 50 patients showed improvement. Suppose if the drug had no effect, the 50 would have been split the the same proportion as the patients who were not given the treatment. But in this case, about 70% of patients showed improvement, which is significantly higher than the control case.

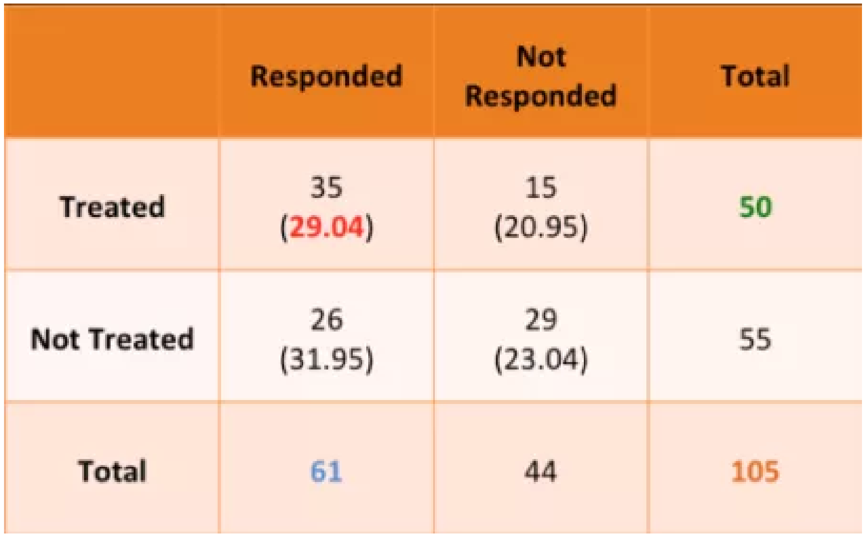
Since both categorical variables have only 2 levels, it was sort of intuitive to day that the drug treatment and health condition are dependent. But, as the number of categories increase, we need to quantifiable statistic to definitively say if they are dependent or not.

One such a metric is the **chi-squared statistic**.

Chi-Squared Statistic

For sake of understanding, lets see how Chi-squared statistic is computed for the 2 by 2 case.
To begin with, we will assume that the 2 variables are not related to each other. In that is the case, can you tell what would the expected value of each cell? For example, the first cell will take the value: 50 times 75 by 105, which equals 35.7.

All the expected values can be computed this way (shown in brackets).



	Responded	Not Responded	Total
Treated	35 (29.04)	15 (20.95)	50
Not Treated	26 (31.95)	29 (23.04)	55
Total	61	44	105

Once that is done, the Chi-Sq statistic is computed as follows.

$$\chi^2= \sum_{i=1}^{n} \frac{(O_i – E_i)^2}{E_i}$$

Numeric Computation

$\ Chi-Sq = ((35-29.04)^2 / 29.04) + ((15-20.95)^2 / 20.95) + $
$\ ((26-31.95)^2 / 31.95) + ((29-23.04)^2/23.04) = 5.56 $

This value will be larger if the difference between the actual and expected values widens.

Also, the more the categories in the variables the larger the chi-squared statistic should be.

Chi-Squared Test

In order to establish that 2 categorical variables are dependent, the chi-squared statistic should be above a certain cutoff. This cutoff increases as the number of classes within the variable increases.

Alternatively, you can just perform a chi-squared test and check the p-values.

Like all statistical tests, chi-squared test assumes a null hypothesis and an alternate hypothesis. The general practice is, if the p-value that comes out in the result is less than a pre-determined significance level, which is 0.05 usually, then we reject the null hypothesis.

H0: The The two variables are independent
H1: The two variables are related.

The null hypothesis of the chi-squared test is that the two variables are independent and the alternate hypothesis is that they are related.

R Code

Let's work it out in R by doing a chi-squared test on the treatment (X) and improvement (Y) columns in treatment.csv
First, read in the treatment.csv data.

```
df <- read.csv("https://goo.gl/j6lRXD")
table(df$treatment, df$improvement)
              improved not-improved
  not-treated       26           29
  treated           35           15
```

Let's do the chi-squared test using the `chisq.test()` function. It takes the two vectors as the input. We also set `correct=FALSE` to turn off Yates' continuity correction.

```
# Chi-sq test
chisq.test(df$treatment, df$improvement, correct=FALSE)
        Pearson's Chi-squared test

data:  df$treatment and df$improvement
X-squared = 5.5569, df = 1, p-value = 0.01841
```

We have a chi-squared value of 5.55. Since we get a p-Value less than the significance level of 0.05, we reject the null hypothesis and conclude that the two variables are in fact dependent. Sweet!

Mini-Challenge

For this challenge, find out if the 'cyl' and 'carb' variables in 'mtcars' dataset are dependent or not. Go ahead, pause and try it out. I will be showing the answers in a few seconds. Good Luck!

So here is my answer:
Let's have a look the table of `mtcars$carb vs mtcars$cyl`.

```
table(mtcars$carb, mtcars$cyl)
    4  6  8
```

```
  1 5 2 0
  2 6 0 4
  3 0 0 3
  4 0 4 6
  6 0 1 0
  8 0 0 1
```

Since there are more levels, it's much harder to make out if they are related. Let's use the chi-squared test instead.

```
# Chi-sq test
chisq.test(mtcars$carb, mtcars$cyl)
        Pearson's Chi-squared test

data:  mtcars$carb and mtcars$cyl
X-squared = 24.389, df = 10, p-value = 0.006632
```

We have a high chi-squared value and a p-value of less that 0.05 significance level. So we reject the null hypothesis and conclude that `carb` and `cyl` have a significant relationship.

Congratulations of you got it right!

If you liked this post, you might find my latest video course 'Introduction to R Programming' to be quite resourceful. Happy Learning!

Related Post

1. Missing Value Treatment
2. R for Publication by Page Piccinini
3. Assessing significance of slopes in regression models with interaction
4. First steps with Non-Linear Regression in R
5. Standard deviation vs Standard error

programming (RStudio, Sweave, LaTeX, SQL, Eclipse, git, hadoop, Web Scraping) statistics (regression, PCA, time series, trading) and more...

If you got this far, why not **subscribe for updates** from the site? Choose your flavor: e-mail, twitter, RSS, or facebook...

Like 2 Share Share

Comments are closed.

Search & Hit Enter

Recent popular posts

- Deep Learning with R
- Add P-values and Significance Levels to ggplots
- Introducing the MonteCarlo Package
- How to create dot-density maps in R

Most visited articles of the week

1. How to write the first for loop in R
2. Installing R packages
3. Using apply, sapply, lapply in R
4. How to Make a Histogram with Basic R
5. Tutorials for learning R
6. How to perform a Logistic Regression in R
7. Freedman's paradox
8. In-depth introduction to machine learning in 15 hours of expert videos
9. Shiny app to explore ggplot2

Sponsors

EARL
CONFERENCE
Abstract submissions open!
earlconf.com/abstracts

DataCamp
Learn R
by doing.

BUILDING TIDY TOOLS
HADLEY WICKHAM, RStudio
May 17 & 18, 2018 | San Francisco
LEARN MORE AND REGISTER

Beginner's Guide to
Spatial, Temporal and Spatial-Temporal Ecological Data Analysis with R-INLA
Zuur, Ieno, Saveliev

Register for free updates
in all areas of statistics
STATISTICS VIEWS
Bringing Statistics Together
WILEY

Acquire Data
Science Skills 2018
Join in
May 1-4 | Boston | ODSC EAST

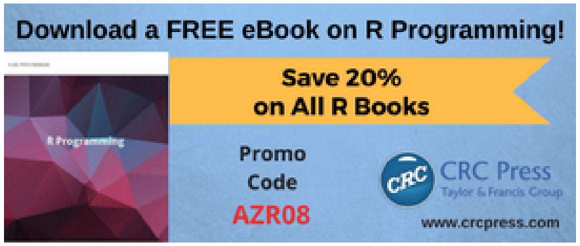
The Future of Data
Science gathers here.
#ODSC
Boston.
London.
San Francisco.

R Courses for Professionals
Download R templates for machine learning
The fastest way to learn
data science!
VIEW SNEAK PEAK
DATA SOCIETY

R-Akademie
on the road.
April 2018 | Hamburg

R for Machine Learning:
A Hands-On Introduction
with Max Kuhn
Predictive Analytics World Business
June 4, 2018
Las Vegas
LEARN MORE

TABLEAU'S VIS
ON BIG DATA
GET THE WHITEPAPE
+ a b | e a u

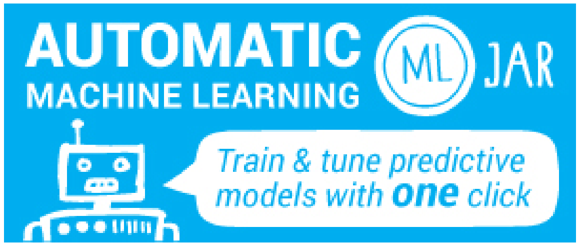

Download a FREE eBook on R Programming!
Save 20%
on All R Books
Promo
Code
AZR08
R Programming
CRC Press
Taylor & Francis Group
www.crcpress.com

jumping rivers
TRAINING: R, SCALA, STAN

STATWORX
Data Science Service
Data Science | Consulting | Development | Training

Try the FASTEST ML for R
Click for a Free Trial
YOTTAMINE ANALYTICS

SIGMA X
SIGMA X


AUTOMATIC MACHINE LEARNING
ML JAR
Train & tune predictive models with one click

Contact us if you wish to help support R-bloggers, and place **your banner here**.

Jobs for R users

- Data Scientist @ Garching bei München, Bayern, Germany
- Software Developer
- Senior Quantitative Analyst, Data Scientist
- R data wrangler
- Senior Data Scientist
- Manager, Statistical Consulting & Data Science
- Financial Controller

Search & Hit Enter

Full list of contributing R-bloggers