- Twitter
- alt=
- shiny
- boxplot
- hadoop
- time series
- animation
- ggplot2
- latex
- ggplot
- trading
- excel
- finance
- PCA
- quantmod
- googlevis
- RSTUDIO
- how to import image file to R
- market research
- rattle
- eclipse
- knitr
- SQL
- rcmdr
- coplot
- tutorial
- map

Recent Posts

- Post your resume to R-users.com (+6 new R jobs for 2016-02-22)
- more e's [and R's]
- Convert Car2go Austin into an EV Fleet
- Large scale eigenvalue decomposition and SVD with rARPACK
- inegiR version 1.2
- Working with the Clinton State Dept Email Dumps in R (Part 1: Graphs)
- Read from hdfs with R. Brief overview of SparkR.
- Toying with models: The Luria–Delbrück fluctuation test
- Microsoft R Open 3.2.3 update
- How I build up a ggplot2 figure
- bayesboot: An R package for doing the Bayesian bootstrap
- Shinyapps.io Update Notification
- Dynamic Stochastic General Equilibrium models made (relatively) easy with R
- Password protect Shiny Apps
- Big Business Backs Hillary: Small Bernie

Other sites

Comparison of two proportions: parametric (Z-test) and non-parametric (chi-squared) methods

July 29, 2009
By Todos Logos

Consider for example the following problem.
The owner of a betting company wants to verify whether a customer is cheating or not. To do this want to compare the number of successes of one player with the number of successes of one of his employees, of which he is certain that he is not cheating. In a month's time, the player performs 74 bets and wins 30; the player in the same period of time making 103 bets, wins 65. Your client is a cheat or not?

A problem of this kind can be solved in two different ways: using a parametric and a non-parametric method.

*** Solution with the parametric method: Z-test.**

You can use a Z-test if you can do the following two assumptions: the probability of common success is approximate 0.5, and the number of games is very high (under these assumption, a binomial distribution is approximate a gaussian distribution). Suppose that this is the case. In R there is no function to calculate the value of Z, so we remember the mathematical formula, and we write our function:

$$Z=\frac{\frac{x_1}{n_1}-\frac{x_2}{n_s}}{\sqrt{\widehat{p}(1-\widehat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}$$

```
z.prop = function(x1,x2,n1,n2){
  numerator = (x1/n1) - (x2/n2)
  p.common = (x1+x2) / (n1+n2)
  denominator = sqrt(p.common * (1-p.common) * (1/n1 + 1/n2))
  z.prop.ris = numerator / denominator
  return(z.prop.ris)
}
```

`z.prop` function calculates the value of Z, receiving input the number of successes (x1 and x2), and the total number of games (n1 and n2). We apply the function just written with the data of our problem:

```
z.prop(30, 65, 74, 103)
[1] -2.969695
```

We obtained a value of z greater than the value of z-tabulated (1.96), which leads us to conclude that the player that the director was looking at is actually a cheat, since its probability of success is higher than a non-cheat user.

*** Solution with the non-parametric method: Chi-squared test.**

Suppose now that it can not make any assumption on the data of the problem, so that it can not approximate the binomial with a Gauss. We solve the problem with the test of chi-square applied to a 2×2 **contingency table**. In R there is the function prop.test.

```
prop.test(x = c(30, 65), n = c(74, 103), correct = FALSE)

    2-sample test for equality of proportions without continuity correction

data: c(30, 65) out of c(74, 103)
X-squared = 8.8191, df = 1, p-value = 0.002981
alternative hypothesis: two.sided
95 percent confidence interval:
  -0.37125315 -0.08007196
sample estimates:
prop 1     prop 2
0.4054054  0.6310680
```

`Prop.test` function calculates the value of chi-square, given the values of success (in the vector `x`) and total attempts (in the vector `n`). The

vectors x and n can also be previously declared, and then be retrieved as usual: `prop.test (x, n, correct = FALSE)`.

In the case of small samples (low value of n), you must specify `correct = TRUE`, so as to change the computation of chi-square based on the **continuity of Yates**:

```
prop.test(x = c(30, 65), n = c(74, 103), correct=TRUE)

    2-sample test for equality of proportions with continuity correction

data: c(30, 65) out of c(74, 103)
X-squared = 7.9349, df = 1, p-value = 0.004849
alternative hypothesis: two.sided
95 percent confidence interval:
  -0.38286428 -0.06846083
sample estimates:
prop 1     prop 2
0.4054054  0.6310680
```

In both cases, we obtained p-value less than 0.05, which leads us to reject the hypothesis of equal probability. In conclusion, the customer is a cheat. For confirmation we compare the value chi-square-value calculated with the chi-square-tabulation, which we calculate in this way:

```
qchisq(0.950, 1)
[1] 3.841459
```

`qchisq` function calculates the value of chi-square as a function of alpha and degrees of freedom. Since chi-square-calculated is greater than chi-square-tabulation, we conclude by rejecting the hypothesis H0 (as stated by the p-value, and the parametric test).

Sponsors

Jobs for R users

- Root word and Noun Phrase Extraction
- Data analyst
- Statistician R developer for agriculture