

Log inSign up



Cross Validated is a question and answer site for people interested in statistics, machine learning, data analysis, data mining, and data visualization. It only takes a minute to sign up.

Sign up to join this community

Anybody can ask a question

Anybody can answer

The best answers are voted up and rise to the top

**Cross Validated**
SPONSORED BY  **aws** machine learning


[Ask Question](#)

P-value for point biserial correlation in R

Asked 4 years, 2 months ago Active 7 months ago Viewed 9k times

Does anybody know of an R package that produces a p-value for point biserial correlations?

3 I've tried all of the major packages that I know (with some help from Google) and haven't found any. If a package doesn't come to mind, is there some way that I can intuitively calculate the p-value?

 r correlation p-value

 share cite improve this question follow

edited Apr 11 '18 at 13:48



cbrnr

237 1 12

asked Jul 28 '16 at 18:20



user2917781

323 3 13

[add a comment](#)

3 Answers

Active Oldest Votes

4 The [point-biserial correlation](#) is equivalent to calculating the Pearson correlation between a continuous and a dichotomous variable (the latter needs to be encoded with 0 and 1). Therefore, you can just use the standard `cor.test` function in R, which will output the correlation, a 95% confidence interval, and an

```
set.seed(1)
x <- sample.int(100, 50, replace=TRUE)
y <- sample(c(0, 1), 50, replace=TRUE)
cor.test(x, y)
```

This yields a correlation of $r = 0.202$, which is not significant ($t = 1.429$, $df = 48$, $p = 0.1595$):

```
Pearson's product-moment correlation

data:  x and y
t = 1.429, df = 48, p-value = 0.1595
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.08088534  0.45478598
sample estimates:
      cor 
0.2020105
```

As @sal-mangiafico and @igor-p point out, the function `biserial.cor` from the `ltm` package produces slightly different results. This is because `cor.test` uses the population standard deviation, whereas `biserial.cor` uses the sample standard deviation. Furthermore, the result of `biserial.cor` has the opposite sign than the result of `cor.test`. This can be adjusted by specifying the argument `level=2` in `biserial.cor`.

share cite improve this answer follow

edited Feb 18 at 9:00

answered Jul 13 '17 at 8:06



cbrnr

237 1 12

add a comment

In response to @user9413061,

2 I think I discovered the source of the problem.

In the [standard definition of biserial correlation](#), the population standard deviation is used.

`ltm::biserial.cor` uses the sample standard deviation.

In the following, a function is defined to calculate the population standard deviation. The function `biserial.cor.new` is defined, which is the same as `ltm::biserial.cor` with `sd.pop` used instead of `sd`.

I think `biserial.cor.new` will return the same result as `cor.test`.

```
sd.pop = function(x){sd(x)*sqrt((length(x)-1)/length(x))}

biserial.cor.new =
function (x, y, use = c("all.obs", "complete.obs"), level = 1)
{
  if (!is.numeric(x))
    stop("'x' must be a numeric variable.\n")
  y <- as.factor(y)
  if (length(levels(y)) > 2)
```

[Log in](#)[Sign up](#)

```
stop("'x' and 'y' do not have the same length")
use <- match.arg(use)
if (use == "complete.obs") {
  cc.ind <- complete.cases(x, y)
  x <- x[cc.ind]
  y <- y[cc.ind]
}
ind <- y == levs[level]
diff.mu <- mean(x[ind]) - mean(x[!ind])
prob <- mean(ind)
diff.mu * sqrt(prob * (1 - prob))/sd.pop(x)
}
```

And an example:

```
x = c(3,4,5,6,7,5,6,7,8,9)
y = c(0,0,0,0,0,1,1,1,1,1)

library(ltm)

### DIFFERENT RESULTS WITH ltm::biserial.cor

biserial.cor(x,y, level=2)

### [1] 0.5477226

cor.test(x,y)

### Pearson's product-moment correlation
### sample estimates:
###      cor
### 0.5773503

### SAME RESULTS WITH new function

biserial.cor.new(x,y, level=2)

### [1] 0.5773503

cor.test(x,y)

### Pearson's product-moment correlation
### sample estimates:
###      cor
### 0.5773503
```

[share](#) [cite](#) [improve this answer](#) [follow](#)

edited Apr 9 '18 at 23:41

answered Apr 9 '18 at 18:26



Sal Mangiafico

6,082 2 8 22

[add a comment](#)

For my understanding you don't have to code the dichotome variable with 0 and 1. Therefore using other values results in exactly the same output. Try for example:



```
y2 <- rep(c(-786,345), 50)
cor.test(x, y)
cor.test(x, y2)
```

Both gives you an r of 0.01732137. The only thing that can happen by coding the dichotome variable differently is that you get -0.01732137, which will be the case if the first number is bigger than the second, e.g.

```
y3 <- rep(c(0,1), 50)
cor.test(x, y3)
```

results in -0.01732137.

Furthermore, I read on different pages that "the point-biserial correlation is equivalent to calculating the Pearson correlation between a continuous and a dichotomous variable", but in fact I get different results if I conduct a Pearson and a point-biserial correlation on same data. An example:

```
x <- 1:100
y <- rep(c(0,1), 50)

cor.test(x, y)
```

gives me 0.01732137, but `biserial.cor(x, y)` results in -0.01723455.

I understand that it is okay to get positive and negative values, but the absolute value should be the same, which is not the case. The results are also different if I use other data, e.g. `x <- rnorm(100, 100, 15)` instead of `x <- 1:100`.

For this reason I am unsure whether it is acceptable to use `cor.test()` and report that you have conducted a point-biserial correlation.

[share](#) [cite](#) [improve this answer](#) [follow](#)

edited Apr 9 '18 at 8:40



Ferdi

4,516 5 36 56

answered Apr 9 '18 at 8:10



user203567

Please state, which package you take `biserial.cor` from. I guess, it's the `ltm` package, but who knows. So the correlation is 0.017 in both cases, the rest probably some rounding error? If one function was superior, how do you know, which? – Bernhard Apr 9 '18 at 8:54

It appears that `ltm::biserial.cor` and `stats::cor.test` do not produce the same result. See `ltm::biserial.cor` for the formula used there. – Sal Mangiafico Apr 9 '18 at 15:43


Yes, I used the ltm package. – user203567 Apr 10 '18 at 9:41


[add a comment](#)


Your Answer

B *I*          

Sign up or **log in**

 Sign up using Google

 Sign up using Facebook

 Sign up using Email and Password

Post as a guest

Name

Email


Required, but never shown


Post Your Answer

By clicking "Post Your Answer", you agree to our [terms of service](#), [privacy policy](#) and [cookie policy](#)

Not the answer you're looking for? Browse other questions tagged [r](#) [correlation](#) [p-value](#) or **ask your own question**.

Featured on Meta

 Goodbye, Prettify. Hello highlight.js! Swapping out our Syntax Highlighter

 Hot Meta Posts: Allow for removal by moderators, and thoughts about future...

Linked

0

Which is the optimal method to check association between continuous and binary data?

Related

[Log in](#)[Sign up](#)

autoregressive parameters in R using GLS with ARMA(p=1) correlation

1

Calculating average variance extracted (AVE) in R for checking discriminant validity (Fornell-Larcker criterion)

1

What does correlation mean in error propagation?

5

How to determine the cut-point of continuous predictor in survival analysis, optimal or median cut-point?

2

C-index for survival analysis with confidence interval (R)

3

Understanding very high p value with Spearman's rank correlation

3

Correlation of random variables

0

Erdem correlation for time series

Hot Network Questions



What is the mean absolute difference between values in a normal distribution?



Algebraic independence of shifts of the Riemann zeta function



Is there a technical (or slang) term for triggering a game loss due to attempting to draw from an empty library?



What book(s) would you recommend for structuring and pricing Exotic Products?



Employer planning on making a change that I'm prepared to quit over. How should I tell manager?



How can I identify the reason that makes a MILP model hard for solvers such as CPLEX?



Why is Schenker so influential in US academia? Is it the same elsewhere?



Can this be formulated as one inequality



How do I get rid of a mental barrier with respect to continuing research work that I have already done?



Output the International Phonetic Alphabet



Does having the (accurate) shape of an Aboleth give one all their ancestral-genetic memories?



The randomness of modular squaring



Why "Giraffe" as a name for the animal?



Good name for object between squared and rounded



Can employer legally stop paying time & 1/2 to exempt employee after stating in the offer that they would do so?



Log in

Sign up



"What does it sound like" vs "How does it sound like"



How can some USB 2.0 audio interfaces support phantom power through USB alone?



What would happen if I don't replace worn drivetrain components?



Did the switchblade in Se7en violate the Chekhov's gun principle?



Why does the Amiga 500 have half the HSYNC pulses separately compared to those in CSYNC?



Grammar of Negative Verb + 限り



What's the word for asking someone to deliver their promise?



Why does the Quantum Realm behave different for Janet van Dyne than for Scott Lang?



Question feed

CROSS VALIDATED

[Tour](#)

[Help](#)

[Chat](#)

[Contact](#)

[Feedback](#)

[Mobile](#)

[Disable Responsiveness](#)

COMPANY

[Stack Overflow](#)

[For Teams](#)

[Advertise With Us](#)

[Hire a Developer](#)

[Developer Jobs](#)

[About](#)

[Press](#)

[Legal](#)

[Privacy Policy](#)

STACK EXCHANGE NETWORK

[Technology](#) >

[Life / Arts](#) >

[Culture / Recreation](#) >

[Science](#) >

[Other](#) >

[Blog](#) [Facebook](#) [Twitter](#) [LinkedIn](#) [Instagram](#)

site design / logo © 2020 Stack Exchange Inc;
user contributions licensed under [cc by-sa](#).
rev 2020.9.25.37676