Simulation of correlated categorical and continuous data

Asked 2 years, 7 months ago Modified 2 years, 7 months ago Viewed 502 times

Part of R Language Collective

1

I want to simulate correlated categorical and continuous data. How to achieve that in R?

```
#For example, how to simulate the data in a way that these two variable are correlated?
x <- sample( LETTERS[1:4], 1000, replace=TRUE, prob=c(0.1, 0.2, 0.65, 0.05) )
#Categorical variable
y <- runif(1000,1,5) #Continuous variable
```

Any ideas will be greatly appreciated!

r simulation correlation categorical-data continuous Edit tags

Share Edit Follow Close Flag

asked Feb 21, 2021 at 19:36

cliu

933 6 13

2 Answers

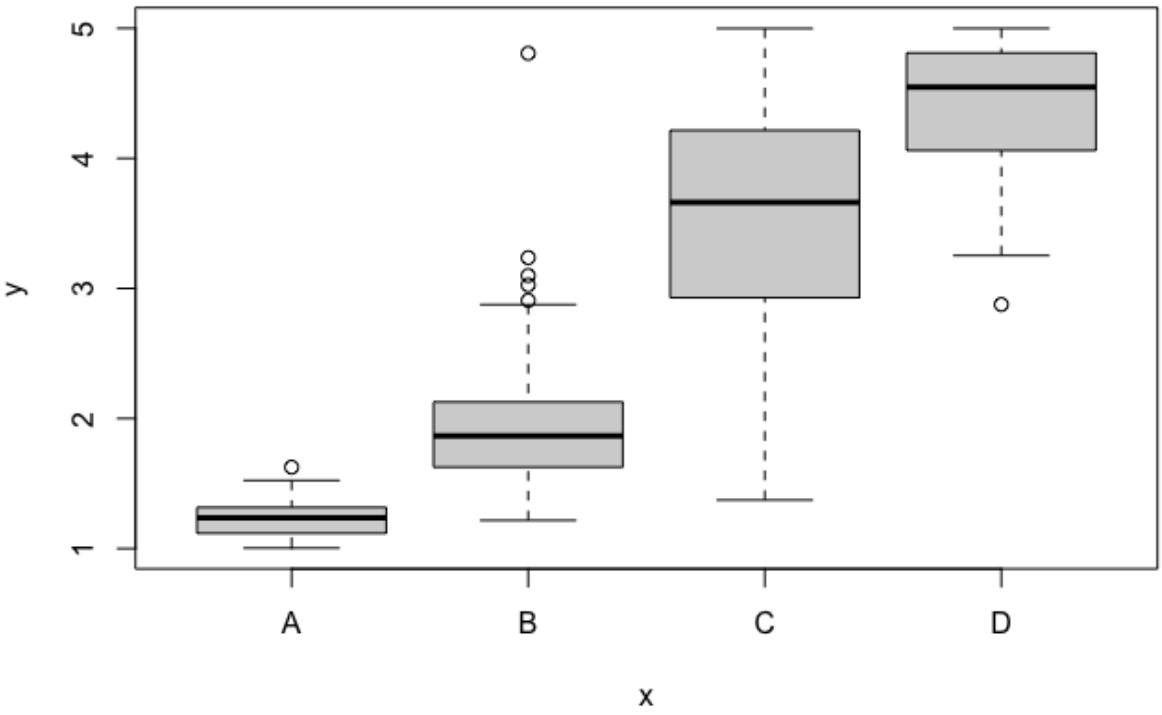
Sorted by:
Reset to default

Date modified (newest first)

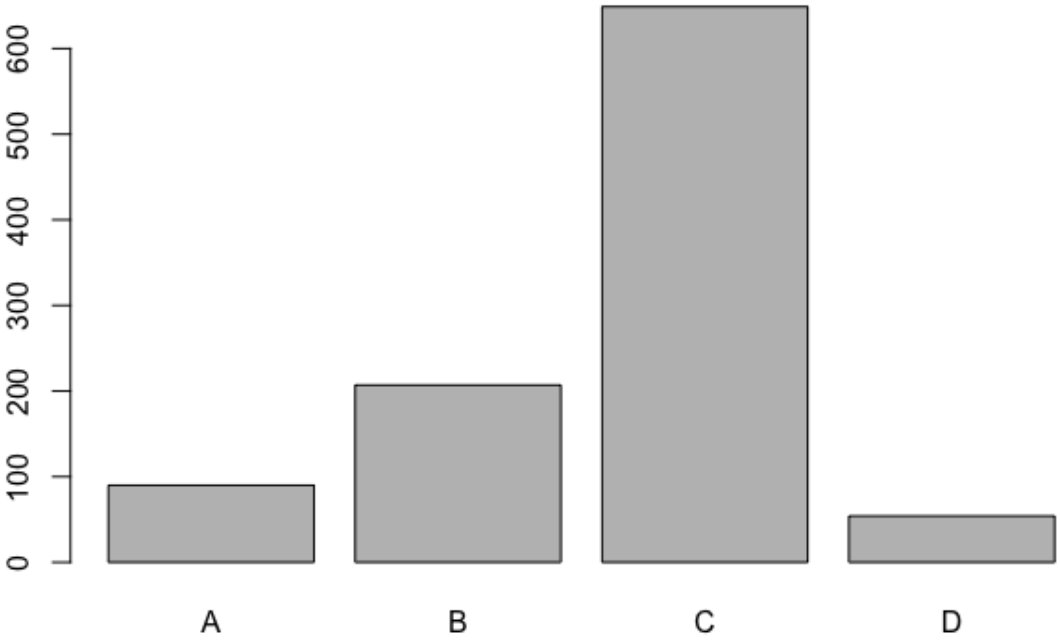
0

Here's a method using copulas. Use larger values of `alpha` for higher correlation.

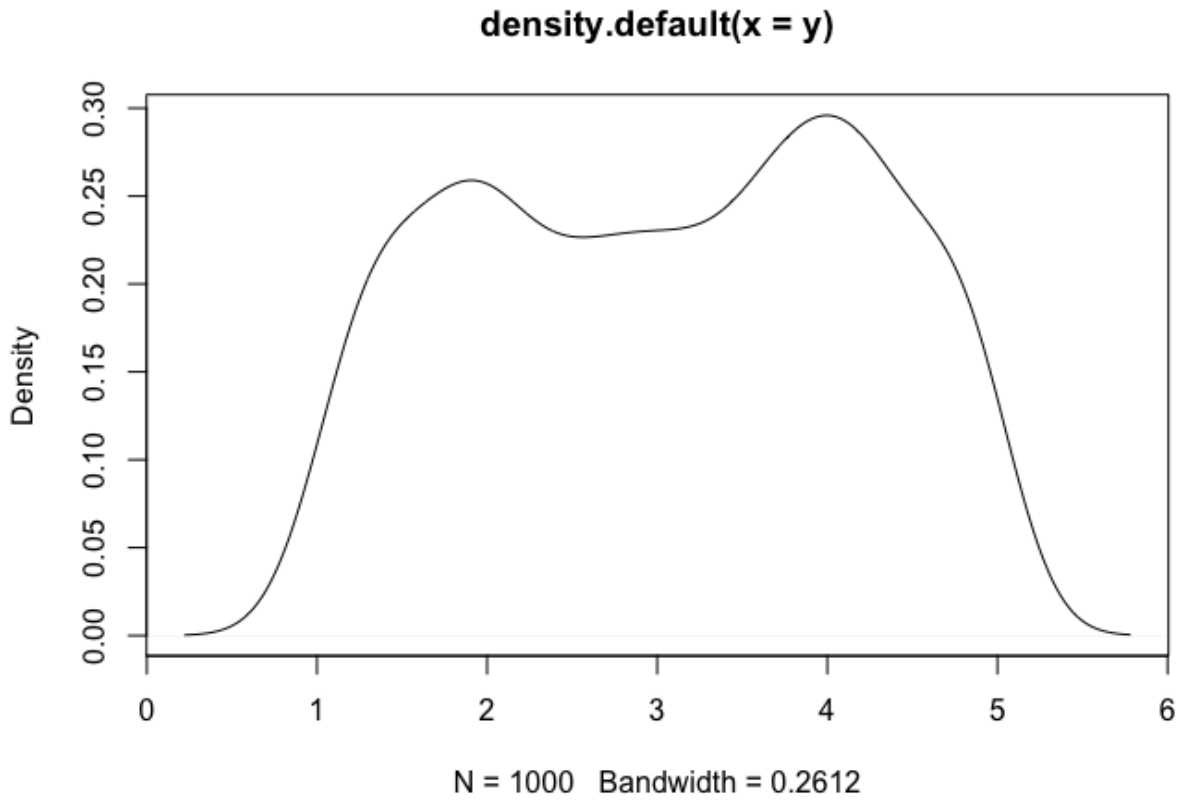
```
library(copula)
n <- 1000
alpha <- 5
u <- rCopula(n, claytonCopula(alpha))
u1 <- u[,1]
u2 <- u[,2]
x <- ifelse(u1 < 0.1, "A",
     ifelse(u1 < 0.3, "B",
     ifelse(u1 < 0.95, "C", "D")))
y <- qunif(u2, 1, 5)
plot(factor(x), y)
```




```
plot(factor(x))
```




```
plot(density(y))
```



Created on 2021-02-21 by the reprex package (v0.3.0)

Share Edit Follow Flag

answered Feb 21, 2021 at 20:40

user2554330
37.5k 4 43 90

Thanks for the answer@user2554330. Any ways to specify the correlation coefficient in the code though? – cliu Feb 21, 2021 at 21:13

What kind of correlation coefficient do you mean? The usual one needs numeric data. – user2554330 Feb 21, 2021 at 21:22

Yes, let's say we change `x` to this `x <- ifelse(u1 < 0.1, 1, ifelse(u1 < 0.3, 2, ifelse(u1 < 0.95, 3, 4)))` . If I want to fix r=0.2, how to specify that? – cliu Feb 21, 2021 at 21:25

1 Change the value of `alpha` . The relation between `alpha` and the correlation will depend on the distributions in some ugly way; I'd do it by simulation (i.e. compute correlation for large samples with a few values of `alpha` , and interpolate). For your distributions it looks like `alpha <- 0.3` will come close to `r=0.2` . – user2554330 Feb 21, 2021 at 21:32

5

Does this give you something like what you're looking for? You can change the sd value to modify the amount of correlation.

```
k <- 1:4
n <- 1000
x <- sample( LETTERS[k], n, replace=TRUE, prob=c(0.1, 0.2, 0.65, 0.05) )
y <- as.vector(sapply(k,function(x) rnorm(round(n/length(k)),mean=x,sd=2)))
```

Share Edit Follow Flag

answered Feb 21, 2021 at 19:52

rdodhia
350 2 9