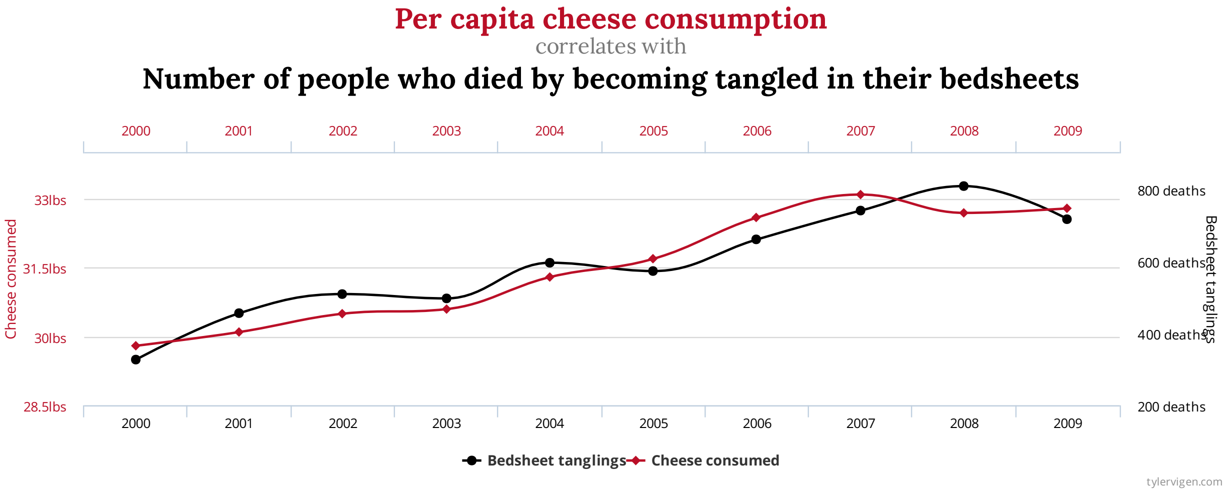
Spurious Correlations and Collider Bias

James Hunter

March 26, 2020

Introduction

We hear a lot "*correlation* does not imply *causation*" or an association between two variables does not mean one causes the other. In fact, correlations can be purely random in nature. From time to time, we hear of weird and wonderful cases such as the 0.94 correlation between the per capital consumption of cheese in the United States with the number of people who died as a result of becoming entangled in their bedsheets.[Vigen] Besides thinking of them as ridiculous, we call these correlations "spurious".



Hollywood Stars and Mind vs. Body

Frequently, spurious correlations come about because of a bias in sampling that can be introduced into a data analysis, unintentionally or with malice aforethought. Let's invent our version of a case that has appeared in various forms before. [von Jouanne-Diedrich, Rossman] How do Hollywood stars become stars. They start out as simple actors, waiting on tables, acting as parking valets, going to classes to improve their craft. Let's imagine that casting directors only have two characteristics of the actors to work with, their intellectual capacity and their physical attractiveness – their *minds* and their *bodies*. Let's assume in the population of actors that these characteristics are normally and randomly distributed. So, our **population** of actors is has no special standout characteristics that we can measure in this study. We will also assume that our population of actors has a total of 1,000 individuals.

Let us see how this population looks on the two dimensions we have information on using R.

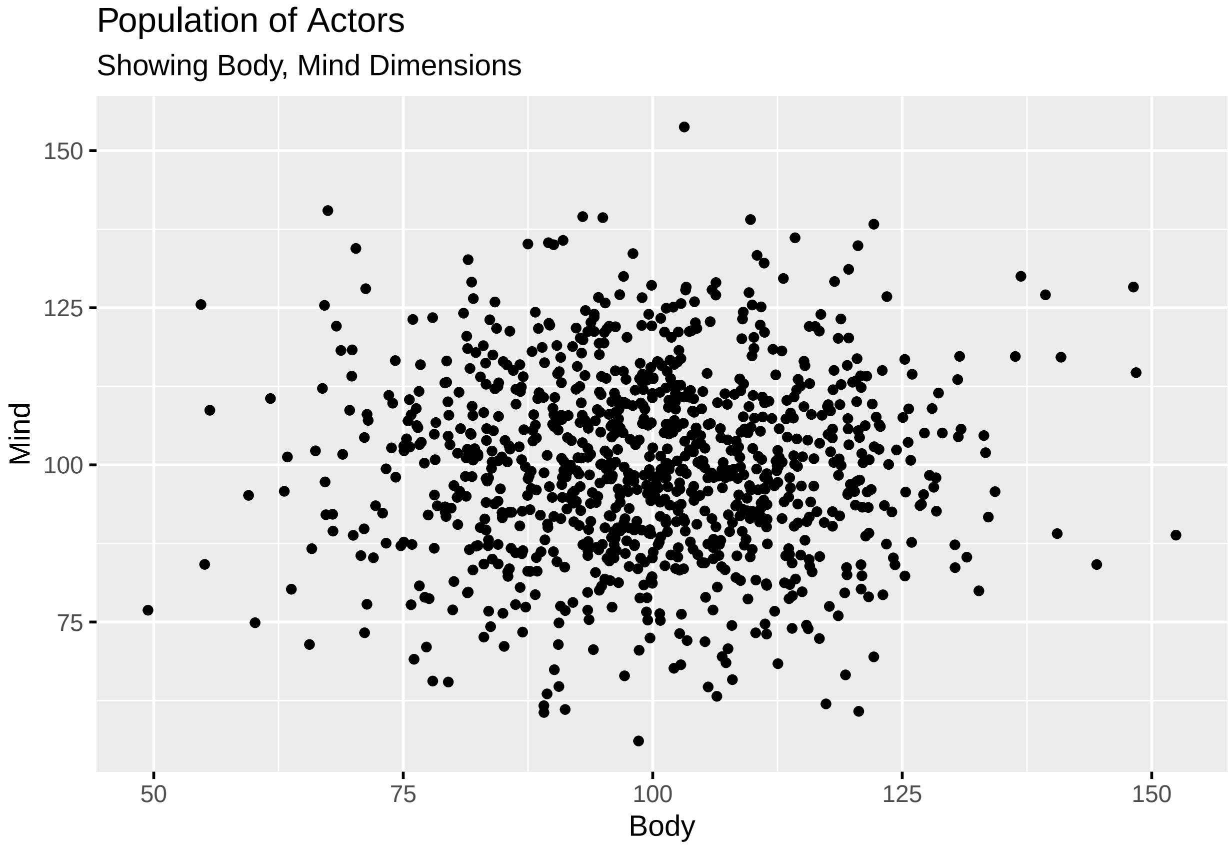
```
# load packages
pacman::p_load(tidyverse, ggpubr, glue, here)

# create the random population

set.seed(42)
```

```r
body <- rnorm(1000, 100, 15) # random normal numbers with a mean of 100, sd of 15
mind <- rnorm(1000, 100, 15)
pop <- tibble(body, mind)

pop %>%
  ggscatter(x = "body",
            y = "mind",
            shape = 20,
            fill = "grey",
            palette = "uchicago",
            title = "Population of Actors",
            subtitle = "Showing Body, Mind Dimensions",
            xlab = "Body",
            ylab = "Mind",
            ggtheme = theme_gray())
```



```r
cor(x = pop$body, y = pop$mind, method = "pearson")
```

```
## [1] 0.009981927
```

Appears pretty random. The population correlation is 0.01, showing almost pure independence between the variables.

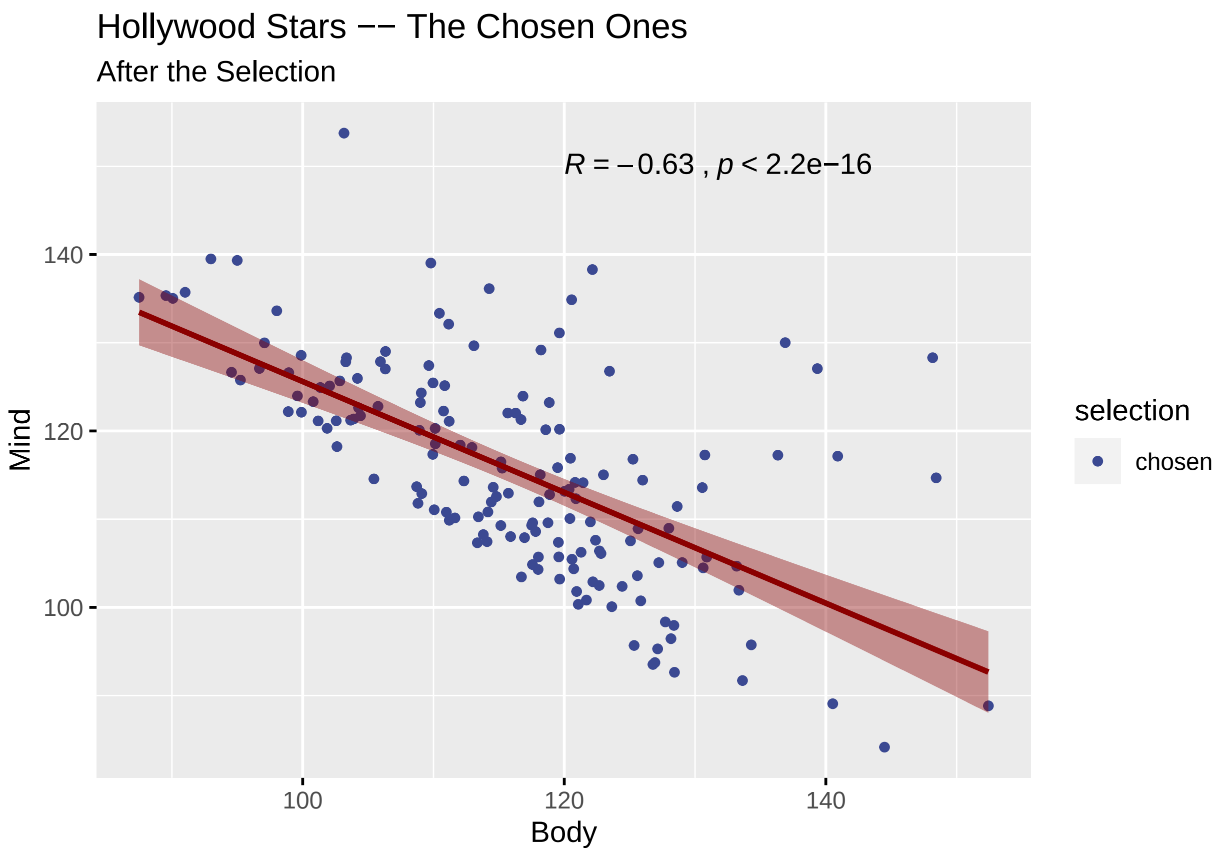
However, casting directors need some decision criterion to decide who gets the parts in films and who will remain a waitperson and "aspiring" actor. In our simulated world, they choose to add together the scores for the `bodyand` `mind` variables and use the sum to choose the twenty percent of the actors who will get the parts. The directors have determined that actors with a combined score of 220 or more will be chosen.

Let's choose the Hollywood stars with R by adding the scores together and then marking our 1,000 actors as being *chosen* or *rejected*. We will then take a look at how these variables are correlated for our Hollywood stars.

```r
pop <- pop %>%
  mutate(body_mind = body + mind) %>%
  mutate(selection = factor(ifelse(body_mind >= 220, "chosen", "rejected")))

# Graph of the stars and calculation of the correlation
star_graph <- pop %>%
  filter(selection == "chosen") %>%
  ggscatter(x = "body",
            y = "mind",
            shape = 20,
            color = "selection",
            palette = "aaas",
            add = "reg.line",
            add.params = list(color = "darkred"),
            conf.int = TRUE,
            cor.coef = TRUE,
            cor.coeff.args = list(method = "pearson", label.x = 120, label.y = 150),
            show.legend.text = FALSE,
            title = "Hollywood Stars -- The Chosen Ones",
            subtitle = "After the Selection",
            xlab = "Body",
            ylab = "Mind",
            ggtheme = theme_gray())

star_graph
```



```
## Conduct a Pearson correlation test on the stars

stars <- pop %>%
  filter(selection == "chosen")

c_test <- cor.test(stars$body, stars$mind, alternative = "two.sided")
c_test
```

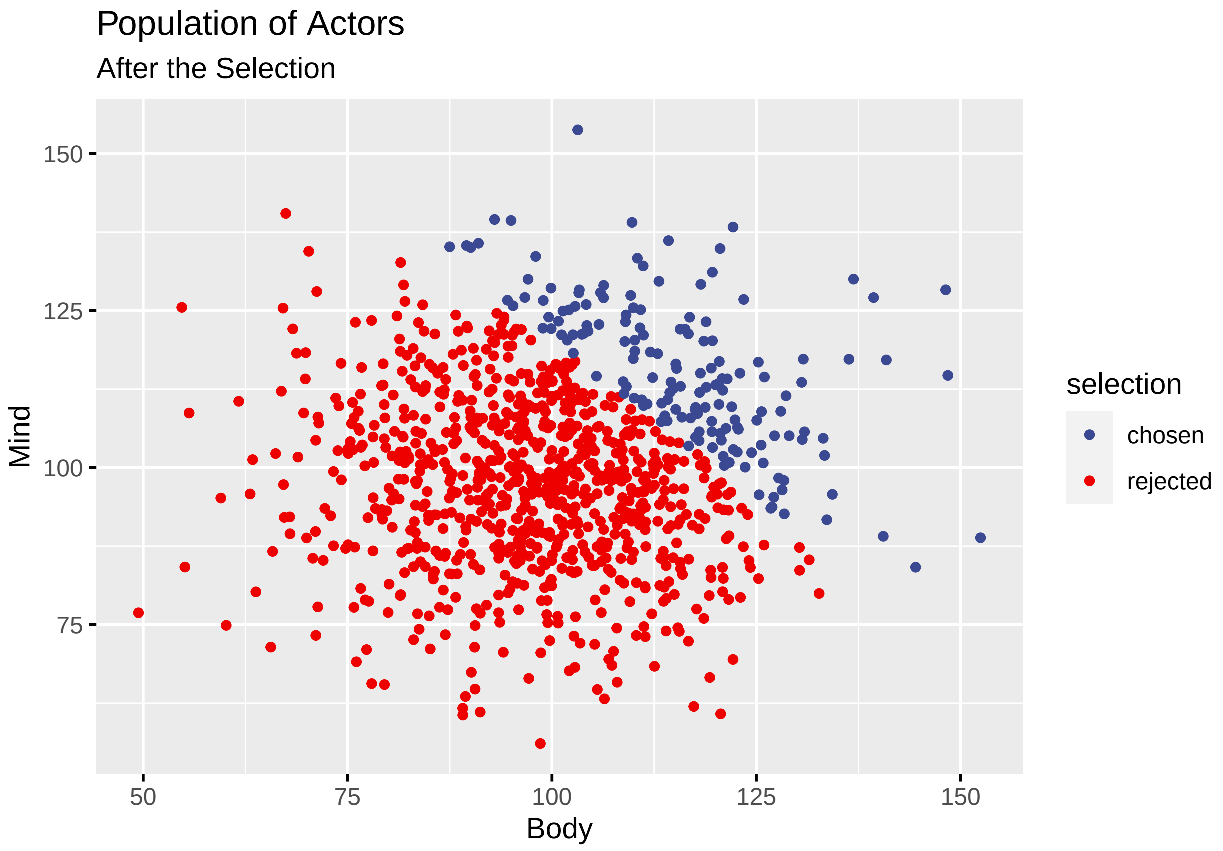
```
##
##  Pearson's product-moment correlation
##
## data:  stars$body and stars$mind
## t = -10.268, df = 159, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7161700 -0.5283726
## sample estimates:
##        cor
## -0.6314422
```

But, These Results Don't Reflect the Population Result

Our *stars* have a very different relationship between `body` and `mind` than the overall population they are drawn from. Among the *stars*, they are either smart or attractive, but not both. As their `body` scores go up, their `mind` scores decline. Hence, they have a strong negative correlation (r = -0.6314) that appears to be highly significant ($2.7095138 \times 10^{-19}$). However, we know that this sample comes from a population with no correlation between `body` and `mind`. How does this happen?

At work is a phenomenon called either "selection bias" or "collider bias". The casting directors are selecting what they define as an elite group based on a variable that was not directly measured: the `body_mind` score. Even though, or especially because, it is the sum of the two variables we have measures for, it introduces a bias into the sample of actors chosen to be stars. You can see this in a graph of all the actors that shows which were chosen and which were not.

```r
# Graph of the full population
pop %>%
  ggscatter(x = "body",
            y = "mind",
            shape = 20,
            color = "selection",
            palette = "aaas",
            title = "Population of Actors",
            subtitle = "After the Selection",
            xlab = "Body",
            ylab = "Mind",
            ggtheme = theme_gray())
```



The blue *chosen* group in the top right portion of the graph clearly shows the tendency of having relatively higher scores on one variable accompanied by relatively lower scores on the other variable. This selection bias comes about because the `body_mind` variable is a collider. Colliders are variables that are influenced by at least two other variables. This has the result of creating a negative correlation in a sample where no correlation actually exists in the population.

Collider variables are the opposite of "confounder" variables, which frequently play havoc with datasets by hiding relationships among variables that exist because of their relationship to one or more variables that you have not measured.

There it is. A short lesson in statistical bias and in causation and correlation. Because of what we have seen here, correlations are much more carefully evaluated today than in the past in terms of whether they are meaningful as a measure.

However, when you look at movie stars in the real world, do you see a parallel tendency to be either bright or attractive? Does this simple simulation reflect a real-world impression?

References

Rossman, Gabriel, 2012, https://www.theatlantic.com/business/archive/2012/05/when-correlation-is-not-causation-but-something-much-more-screwy/256918/

Vigen, Tyler, https://www.tylervigen.com/spurious-correlations

von Jouanne-Diedrich, Holger K., 2020, https://blog.ephorie.de/collider-bias-are-hot-babes-dim-and-eggheads-ugly