# If you are a Data Science Enthusiast, then this article about the ignorance of Statistics is for you.

I want to start this article by highlighting the issues related to data science that I think as a statistician are important for its success as an emerging new field and I'm going to motivate this piece initially at least with a story about cats which comes from an article that appeared in The New York Times in August of 1989.

You might consider it to be an early example of data journalism, So without further ado let's start with what the piece — On Landing Like a Cat: It Is a Fact, suggested.

Every year scores of cats falls from open windows in New York City, from June the 4th through November the 4th of 1984 for instance 132 such victims
were admitted to the Animal Medical Center in Manhattan.

The article goes on to recite some statistics from the data set like 21 of 22 falling 7 or more stories actually survived, 2 of them fell together, 40% fell at night. The height of the buildings ranged from 2 to 32 stories with an average of 5.5.

Most of the cats landed on concrete most survived. And the weird thing was that there seemed to be a positive relationship between the length of the fall and the probability of survival.

*May be the Cat was able to turn or relax his body or even kind of a flying squirrel thing where they're like flaps of skin underneath its arms.* □□

Well, this is exceeding strange because you might ask what's going on here. Some of them surmised, maybe it's the 9 lives hypothesis or maybe they interviewed some experts who'd say that the cat had some time to fall.

The giveaway is actually in the first paragraph of the article that I am highlighting out to you as it said on three two such victims were admitted to the Animal Medical Center all right dead cats are not transported to Animal Hospitals.

The whole story had no basis and the reason was it was based on this sort of found data. It's sort of a tale of caution in the interpretation of found data and the reason is because this is what statisticians might call a convenient sample.

It's expedient as the data is not collected via any sort of sampling scheme but its available. It is being used to answer questions that weren't really intended by the collectors of the data.

The main problem with this data is *missingness*. Missing data arises everywhere with experiments, especially with observational studies in some cases.

In this piece the missingness was highly correlated with the outcome of interest which was whether or not the cat survived.

This data is what we'd call a non representative sample and by representative we mean that it's related. The data set in hand is the same as the underlying target population of interest these are statistical issues right the inference from the Cats data was unreliable because of ignorance of statistics essentially.

---

## Thanks for making it to the end 

If you liked this article, I've got a practical reads for you one about the Skills in Python for every Data Scientist and one about How to learn Data Science with Python.

*I've also got this Data-Centric newsletter that you might be into. I send a tiny email once or twice every quarter with some useful resource I've found. Don't worry, I hate spam as much as you. Feel free to subscribe. ↗*