✓ 100 XP

Good, bad, and missing data

3 minutes

Machine learning’s predictive power comes from the fact that it's shaped by data. A side effect is that models trained only on small amounts of data rarely perform well in the real-world. This is because small amounts of data aren't usually good representations of the real-world. For example, if we selected four people at random from the world, they wouldn't be very representative of the average person on Earth. By contrast, if we selected 1 billion people, our data would probably be quite representative.

But it’s not all about quantity. It can be equally important that data:

- Aren't just large, but representative
- Don't contain errors
- Aren't missing key information

Here we’ll cover these topics before moving on to a practical exercise working with our Titanic dataset.

What do you mean ‘representative’?

Statisticians rely on two key concepts: *populations* and *samples*, which help us to think about if the data we have is what we want.

A population is what we are interested in or, if you like, in every conceivable datapoint. For our Titanic scenario, we’re interested in knowing which factors led to people surviving—including stowaways who weren't listed on the official records. As an alternative example, if we're investigating the relationship between personality traits and likelihood to sink a ship, our population would be every ship captain who has ever lived.

A sample refers to data that we have available (a subset of the population). For our Titanic dataset, this sample will just be the people who are listed on the official manifest. For our alternative example, our sample might be every ship captain who we can convince to do a personality test down at our local marina.

it's important to think about whether your sample is representative of (‘like’) the population. For our Titanic example, our sample is so large that it’s probably a very good match. By contrast, for our alternative example, only talking to ship captains at our local marina probably isn’t a good cross-section of the kinds of sailors that exist across the globe. Using data from our local marina might build a model that works well for local captains but isn't very helpful when discussing captains from other countries.

What is a data error?

Data errors simply mean data that is wrong. If bad enough, these errors can skew a model to consistently make predictions that are wrong. In short, if you put bad data in, you’ll get bad predictions out.

Errors in data are a fact of life and largely come from two places:

Measurement errors mean that when the data was collected, they were measured badly. These errors are often subtle and difficult or impossible to eliminate.

Data entry errors mean data was collected correctly but entered into a spreadsheet, or similar, incorrectly. Data entry errors are sometimes easier to spot than measurement errors. For example, we might measure someone as 1.8 meters tall, but miss the decimal place and write 18 m, which is easy to identify because a tree-sized person is unrealistic.

What is complete data?

A dataset that is complete has no missing data. Data can be missing in two ways. Consider that we record the heights and weights of Dylan, Reece, and Tom:

Name	Height (m)	Weight (kg)
Dylan	1.8	75
Reece		82

Our data are incomplete because we have a sample missing: Tom is missing entirely, while Reece's height is missing.

Ideally, we always work with complete data, but this practice isn’t always realistic. When we have incomplete data, we can:

- Choose a model that can work with incomplete data, or
- Remove samples (rows) that have incomplete data, and work with what remains, or
- Artificially add values that are missing with reasonable substitutes.

In most circumstances, picking a model that can handle missing data is best, though not always possible. Removing data is the easiest and usually a valid solution, though care needs to be taken that removing data doesn't cause a sample to misrepresent the population. Adding data in artificially is usually a method of last resort.

In the next unit, we'll work with our Titanic dataset, identifying and patching up incomplete data.

Next unit: Exercise - Visualize missing data

Continue >