Data and Sampling Distributions

Random Sampling and Sample Bias

- In the era of big data, we still need sampling to work efficiently with a variety of data.

- Random sampling is a process in which each available member of the population being sampled has an equal chance of being chosen at each draw. The resulting sample is called simple random sample.

- Data science mostly care about the quality of the data. But statistics also cares about the representativeness of the data, in other words, how good our data in hand represent the population.

- Self-selection bias occurs when people who choose to participate in a study are not representative of the population you're trying to study.

Though it can be mitigated to a certain degree if compared to a similarly biased sample.

Bias

- Bias comes in different forms, and may be observable or invisible. When a result does suggest bias (by reference to a benchmark or actual values), it is often an indicator that a statistical or machine learning model has been misspecified, or an important variable left out.

- There are variety of methods to achieve representativeness, but most of them originates from random sampling.

- Strata is the plural form of stratum. A stratum is defined as a homogenus subgroup of a population with common characteristics. (e.g. over-weighting Hispanic and black people when sampling from a typical US state)

This is called stratified sampling.

Selection Bias

- Data snooping, is the act of extensive data hunting to find something interesting.

There is a saying among statisticians: "If you torture the data long enough, sooner or later it will confess."

- We can guard against selection bias by using a holdout set.

Regression to the Mean

- Regression to the mean refers to a phenomenon where extreme observations tend to be followed by more central ones.

To give you an example: A student who got a very high score in an exam tends to get a "good" or "average" score from the next exam.

- Another example is when the children of extremely tall people tend not to be as tall as their parents.

Sampling Distribution of a Statistic

- A sample is typically collected with the purpose of either measuring something or modeling something.

- Because the estimate is based on a sample, it is subject to error.

The result might be different if we were to draw a different sample.

- This uncertainty is referred to as sampling variability.

- By collecting many brand-new samples from the population, calculating the statistic for each, and then calculating the standard deviation of those results, one could directly estimate the standard error.

- The distribution of a sample statistic such as the mean is likely to be more regular and bell-shaped than the distribution of the population. See the coding part for a visualization.

Central Limit Theorem

- What we just described is called central limit theorem.

- It means that the means drawn from multiple samples will resemble the familiar bell-shaped normal curve, if the sample size is large enough and the departure of the data from normality is not too great.

- Formal hypothesis tests and confidence intervals play a small role in data science, and since CLT underlies the machinery of those topics, we will not go deeper.

- In other words, the CLT is not so central in the practice of data science.

Standard Error

- Standard error = $SE = \frac{s}{\sqrt{n}}$ s means standard deviation of the sample distribution, n means the fixed sample size.

- As the fixed sample site increases, the SE decreases.

- In practice, this method is not feasible since it is very wasteful.

The Bootstrap

- Let's say we have a population of 1000 people, and we picked 100 people for sampling.

After that, we calculate the mean of the sample of 100 people, but we have a question in mind.

- What would the mean look like if we took another sample of 100 people? And what could we do it we don't have the ability to take another sample from the population.

We use the bootstrap method, it works like this: After taking the 100 people sample, we take 100 random value from the sample with replacement, which means we may have the same values multiple times in our new sample, gathered using the bootstrap method.

- After doing these steps a lot of times, we actually conclude with a result resembling the CLT result, where we can find the standard error by finding the SD of the distribution of the samples.

Confidence Intervals

- To this point, we only have focused on point estimates. Which creates an undue faith to the estimation we calculated.

To counteract this tendency, we use confidence intervals which represents our estimation as a range rather than a single value.

- We can calculate the confidence intervals by using the classical t-based formulas. Where we calculate the error margin using Student's t-table.

- But we can also use the bootstrapping method to calculate the confidence interval, which is far more practical when we have the required technological resources.

- Suppose your bootstrap means (sorted) look like this:

[94, 95, 96, ..., 108, 109]

And you want to calculat the %60 CI,

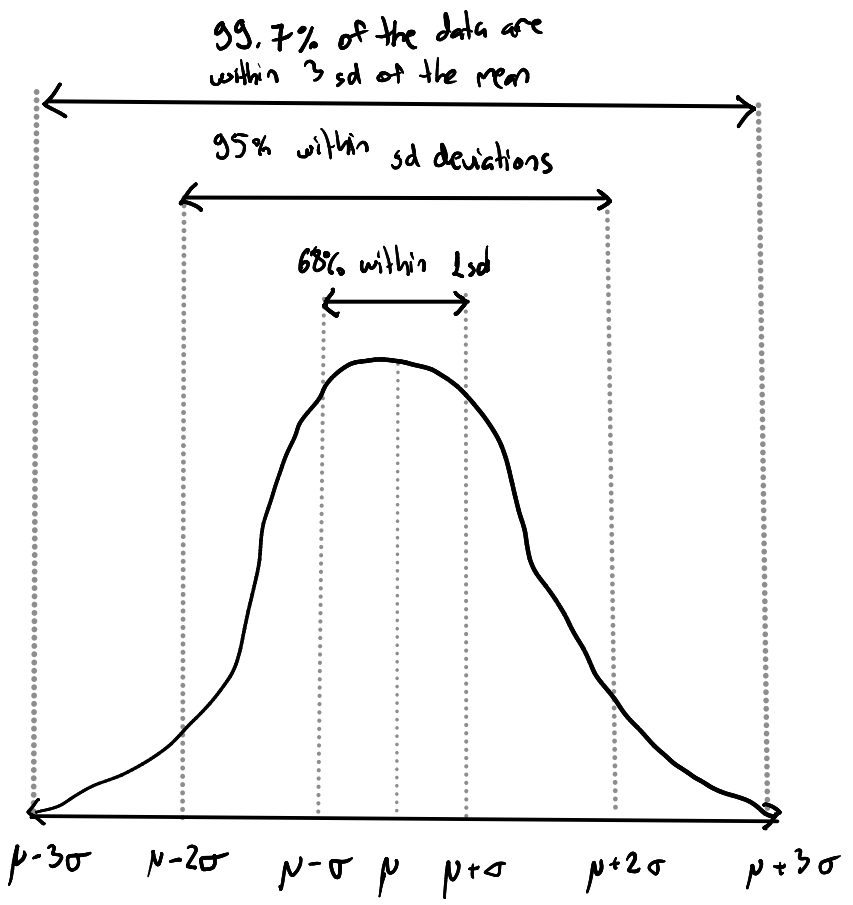
You can directly use the 20th and 80th percentiles to find the CI.

20th percentile ≈ 97 80th percentile ≈ 105

60% boostrap CI → [97, 105], which translates to: "If you repeated the entire sampling and interval-building process many times, about 95% of those intervals would contain the true population value."

Normal Distribution

- The fact that distributions of sample statistics are often normally shaped has made it a powerful tool in the development of mathematical formulas that approximate those distributions.



Standard Normal and QQ-Plots

- To compare the data to standard normal distribution, we calculate the z-score with its formula:

$$z = \frac{x - \mu}{\sigma}$$

x = value you're checking
μ = mean of the dataset
σ = the standard deviation.

More on the relevant coding side.