Monte Carlo simulation

Christian, Jonathan, Marcus, Puk & Sofia

May 5, 2025

Contents

1 Introduction

2 Statistical theory

2.1 Probability space

The **sample space**, S, is the set of all possible outcomes.
If a sample space contains a finite number of possibilities or an unending sequence with as many elements as there are whole numbers, it is called a **discrete sample space**.
Example: When rolling a standard six-sided die form the discrete sample space, the possible outcomes are $S = 1, 2, 3, 4, 5, 6$

If a sample space contains an infinite number of possibilities equal to the number of points on a line segment, it is called a **continuous sample space**.
Example: Measuring the heights of people in a population. This is a continuous sample space, because height can take any real value within a given range.
An **event** is a subset, $A \subseteq S$, of the sample space. The event is the amount that contains all possible events. An example of a discrete event could be rolling a die and getting an uneven number, this would be the event $A = 1, 3, 5$.
For the continuous event, it could be that a person is between 160 cm and 170 cm tall.

The probability of an event A, $P(A)$, is the sum of the weights of all sample points in A. The probability of the whole sample space is 1, $P(S) = 1$ The probability of any event being between 0 and 1,$0 < P(A) < 1$ The probability of the empty set being 0, $P(\emptyset) = 0$

Probability of mutually exclusive events
If A and B are mutually exclusive, $A \cap B = \emptyset$, then
$P(A \cup B) = P(A) + P(B)$

Where A and B never occur at the same time, so their union is equal to the two events added together.
Probability of union
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Here, the union of the two events is A added to B, but minus their common event, since it otherwise would be added twice.
Two events A and B are independent, if
$P(A|B) = P(A)$
The equivalent definition to this is:
Two events A and B are independent if and only if
$P(A \cap B) = P(A)P(B)$ This says that the probability of both event A and B happening, is equal to the product of the two events. If we are interested in certain parameters of a population distribution, we can look at a sample. From this, we can make a **point estimate**.
Examples of this are,
$\bar{x}$ is a point estimate of μ

s is a point estimate of σ

This is often supplemented with a **confidence interval**
This is an interval around the point estimate, where we are confident that the population parameter is located.

For μ, we have different ways of estimating it. We can use the sample mean $\bar{X}$, or the average X_T of the sample upper and lower quartiles. But in this case, we have to look out for **bias**. If the distribution of a population is skewed, then X_T is biased. The result of this is, that in the long run, this estimator will systematically over or under estimate the value of μ. This is written as,
$E(X_T) \neq \mu$.
It is generally preferred that the estimator is **unbiased**. In this case, $\bar{X}$ is an unbiased estimate of the population mean μ.

The standard error of $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$. Here, the standard error decreases, when the sample size increases. If an estimator has this property, it is called **consistent**. If we compare, the estimator X_T is also consistent, but has a greater variance than $\bar{X}$.
It is generally preferred that the estimator has the smallest possible variance, and in that case it is called efficient. So $\bar{X}$ is an efficient estimator.
When estimating a parameter, the symbol $\hat{}$ is used above it. For μ, $\hat{\mu} = \bar{X}$.
We can calculate $\bar{X}$ using the following formula,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

For the variance σ, we can estimate it by using the formula for S^2,

$$S^2 = \frac{1}{n-1}\sum_{i=1}(X_i - \bar{X})$$

2.2 Probability distribution

Data can come in various distributions depending on different parameters such as degrees of freedom. The distribution is the shape of the data and it will have an effect on statistical models. Therefore it is important to have an understanding of distributions.

2.2.1 Normal distribution

In the world of statistics, the most common distribution is the normal distribution. It is constructed as a bell shape. The normal distribution is a continuous distribution, with this density function:

$$n(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)}{2\sigma^2}}$$

The distribution is dependent on the mean(μ) and the standard deviation(σ), where changes to the mean will result in a change in the positioning of the normal distribution. Whereas a change in the standard deviation will change the spread of the curve. The normal distribution also always contains an area under the curve that is equal to one. This is to ensure that the normal distribution correctly models probability.

There is a special case of the normal distribution, called the standard normal distribution, where the mean is zero and the standard deviation is one. All variations of a normal distribution can be standardized by a transformation of the distribution, using the Z-score formula.

$$Z = \frac{X - \mu}{\sigma}$$

Z in the Z-score represents the amount of standard deviations a given X value, deviates from the mean.

2.2.2 The central limit theorem

A very effective theorem in statistics is the central limit theorem. This theorem states that if a random sample $\overline{X}$, with the size n, is taken from a population with a mean and a finite variance, then as n goes towards infinity, the distribution will resemble a normal distribution. If used with the Z-score formula, the distribution will resemble a standard normal distribution. The formula for the Z-score, when in conjunction with the central limit theorem, looks like this:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Where $\overline{X}$ is a random sample of size n and μ is the mean of the true population. The standard error i represented by $\sigma/\sqrt{n}$, where σ is the standard deviation and n is the sample size. Usually the standard deviation is unknown, for these situations it's possible to use the estimator S^2. This estimates the variance of the population from the variance of the sample, by this formula:

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})}{n-1}$$

The square root of the variance is the standard deviation, therefore the square root of the estimator S^2 would be the estimated standard deviation. The problem with using the estimator S^2, is that with small samples the variance is small and therefore it contains a lot of bias. In this situation the t-distribution would be used instead of the normal distribution, because the t-distribution takes the bias into account the bias of the standard deviation. It does this by having thicker tails, meaning that the probability of more extreme values are higher.

2.2.3 The t-distribution

The t-distribution is shaped as the standard normal distribution, in a bell shape and symmetrical around the mean of zero, the difference is that the t-distribution is more variable. This comes from the fact that the t-distribution is dependent on the degrees of freedom. When the degrees of freedom surpasses 30, the rule of thumb is that the distribution will resemble a normal distribution. So before 30 degrees of freedom, the distribution contains more variance. The t-distribution will come to resemble the standard normal distribution, when in surpasses 30 degrees of freedom, this makes sense, since the two distributions have the same formula:

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

The only difference is the estimated standard deviation S.

2.3 Statistical methods

2.3.1 Confidence intervals

The confidence interval is a good tool to use, when trying to estimate a parameter of a population. Its used to create an interval, where the parameter has a probability to lie inside of. This probability is called the confidence level and it's a chosen value, usually the chosen confidence level is either 95% or 99%. The confidence interval will become bigger with a larger confidence level. A good confidence interval is small with a large confidence level, this will usually occur when the sample size is large. The chosen confidence level relates to an α-value, where as an example the chosen confidence level is 95%, then the α-value would be 5% or normally written as 0.05. The α-value will sometimes be needed to find the critical value, that is used to calculate the margin of error, as an example it's used when trying to find the critical value of the confidence interval, when working with a t-distribution.
To set up a confidence interval, the margin of error needs to be computed and then that will be both added and subtracted from the point estimate. This will give the values of the outer bounds of the interval. The margin of error is calculated from this formula:

$$Margin\_of\_error = critical\_value \pm standard\_error$$

The standard error will change depending on which parameter that the confidence interval is estimating, but the general formula for the standard error is:

$$\frac{\sigma}{\sqrt{n}}$$

An example of computing a confidence interval of the mean while working with a standard normal distribution, then the formula for the confidence interval would be this:

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

Where $1 - \alpha$ is the confidence level. As it's the mean that is being estimated, then instead of Z-score, then μ must be isolated and that is done by multiplying $\frac{\sigma}{\sqrt{n}}$ and subtracting $\bar{X}$ on all sides, then multiplying all side by -1 to remove the minus sign. So the formula for a confidence interval of the mean will look like this:

$$P(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

This formula will give the upper and lower bounds of the confidence interval.

The interpretation of a confidence interval
To interpret a confidence interval, it would be incorrect to interpret the confidence level of some value x, as the probability of the true parameter being inside of the interval. The reason behind this is that the computed interval is static, so either the value x is inside the interval or it's not. So the correct way of interpreting the confidence interval is by taking multiple samples and computing the confidence interval for all samples, then the value x would reside inside 95% of the confidence intervals. **Kilde for fortolkningen af kofidense intervaller:**
$http://www.drhuang.com/science/mathematics/book/probability_and_statistics_for_engineering_and_the_scienc$

2.3.2 Hypothesis testing

A hypothesis test is used to test an assumption about a population. This is done from a sample of the population, as the information about the population is usually hard to come by. A hypothesis test is set up, by having a null hypothesis and an alternate hypothesis.

$$H_0 = Null\ hypothesis$$

$$H_a = Alternate\ hypothesis$$

When working with hypothesis testing, the hypothesis H_0 is usually represented as the status quo, where as the hypothesis H_a is represented as the opposition. It is also important to note that there is only two outcomes of a hypothesis test, either H_0 is rejected in favor of H_a or H_0 is failed to be rejected. Therefore in no situation can H_0 be stated to be an absolute truth, as there might be other samples where H_0 will be rejected. Therefore in a hypothesis test H_0 needs to be the thing that can be rejected and if H_0 gets rejected, then H_a will become the new status quo until proven otherwise.
In a hypothesis test H_0 will be the assumption that a parameter for two populations is the same, where as H_a can be either one of three assumptions, depending

on the intention of the hypothesis test.

$$H_0 : \theta = \theta_0$$

1. $H_a : \theta \neq \theta_0$
2. $H_a : \theta < \theta_0$
3. $H_a : \theta > \theta_0$

When the direction of the rejection is not important and also is unknown, then (1) will be the case. This scenario sets up a two-tailed-test, where the hypothesis test is used to reject H_0 if H_a is either significantly larger or smaller than H_0, this means that the critical area is on both sides of the difference of θ and θ_0. Either (2) or (3) will set up a one-tailed-test, where depending on what is important, either the hypothesis test is used to determine if H_a is significantly bigger or smaller than H_0. This means that the critical area only spans one side of the difference between θ and θ_0.

Error in hypothesis testing

When making a hypothesis test there is four different possible outcomes. The results are separated by correct decisions and errors. There exist two types of hypothesis errors, called type 1 error and type 2 error. The type 1 error occurs when H_0 is mistakenly rejected and H_0 is true. Type 2 error is the opposite, where H_a is rejected and H_a is true. The types of outcomes occurring from a hypothesis test can be seen in Table 1

	H_0 is true	H_0 is false
Does not reject H_0	Correct decision	Type 2 error
Reject H_0	Type 1 error	Correct decision

Table 1: Outcomes of a hypothesis test

3 Polynomial regression

3.1 Assumptions

3.1.1 Homoscedasticity

One of the assumptions of a polynomial regression is that homoscedasticity is fulfilled. Homoscedasticity is the assumption of constant error variance, where observations in a dataset would exhibit errors that have roughly the same spread across all levels of the independent variable.
If this assumption is not upheld, then this will cause the standard error to be biased and therefore not trustworthy. This problem causes further testing involving this standard error to become wrong, an example is the hypothesis test. The reason for the assumption needs to be upheld, comes from how the regression is created. The regression is created via the ordinary least square method, that requires the assumption of homoscedasticity to be upheld.
A way to display homoscedasticity is through the variance-covariance matrix. The matrix shows whether the data contains homoscedasticity or heteroscedasticity through the diagonal values. If the matrix contains all the same values through the diagonal, then the assumption of homoscedasticity is upheld, else the data contains heteroscedasticity. This is a showcase of the variance-covariance matrix with homoscedasticity:

$$\mathrm{Var}(\varepsilon) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 & 0 \\ 0 & \sigma^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma^2 & 0 \\ 0 & 0 & \cdots & 0 & \sigma^2 \end{bmatrix} = \sigma^2 I$$

The general way of writing the variance-covariance matrix, is by this formula:

$$\mathrm{Var}(\epsilon) = \mathbb{E}(\epsilon\epsilon') = \begin{bmatrix} \sigma_{1,1}^2 & \sigma_{1,2} & \sigma_{1,3} & \cdots & \sigma_{1,l-1} & \sigma_{1,l} \\ \sigma_{2,1} & \sigma_{2,2}^2 & 0 & 0 & 0 & \\ \sigma_{3,1} & 0 & \ddots & \vdots & \vdots & \\ \vdots & & \ddots & \ddots & 0 & \\ \sigma_{l-1,1} & 0 & & 0 & \sigma_{l-1,l-1}^2 & \\ \sigma_{l,1} & 0 & & & & \sigma_{l,l}^2 \end{bmatrix}$$

Every position in the matrix is calculated, then the diagonal will tell if the data contains homoscedasticity or heteroscedasticity. The positions that are not on the diagonal should be zero else the data contains another problem, that is autocorrelation, meaning that the observations in the data set are correlated.

Source: https://openpublishing.library.umass.edu/pare/article/id/1590/

3.1.2 No multicollinearity

Perfect multicollinearity is a term used for describing a perfect linear relationship between two or more independent variables. This relationship occurs when an independent variable can be perfectly predicted from other independent variables. In mathematical terms, this could be written as a linear regression:

$$X_1 = c + \beta_1 \cdot X_2 + ... + \beta_n \cdot X_n$$

Where $X_1...X_n$ is all the independent variables that have a perfect linear relationship. The coefficients are represented by $\beta_1...\beta_n$ and they are the amount that X_1 changes when their relative independent variable changes. Lastly c is the intercept and represents the value of X_1, when all other independent variables are zero.
The regression model can feel the effects of multicollinearity even without there being perfect multicollinearity. A strong linear relationship is enough to have an effect on the model. The problem caused by multicollinearity, is that as it increases the variance of the value that the coefficients can receive also increases. Where as perfect multicollinearity will make the model unable to estimate a value of one coefficient, due to the perfect linearity between the independent variables.
The VIF (variance inflation factor) is used to measure how much the variance of the coefficients are inflated due to multicollinearity. The formula for the VIF, is as followed:

$$\frac{1}{1 - R^2}$$

Where R^2 is the coefficient of dertermination. This coefficient of dermination is calculated through a linear regression that is set up with an independent variable as the dependent variable and all the other independent variables staying as independent variables. This is done for all independent variables, so each independent variable has a VIF-value. A rule of thumb is that a VIF-value greater that ten is significant and therefore needs to be dealt with.

Source: https://ekja.org/upload/pdf/kja-19087.pdf

4 Pseudo random number generator

5 Pseudo Random Number Generator

To generate the dataset, we use a Pseudo-Random Number Generator (PRNG). PRNGs are algorithms that produce sequences of numbers that appear random but are actually deterministically generated from an initial seed value. While these numbers are not truly random, they are sufficiently unpredictable for many practical applications. Random numbers are widely used in fields such as statistics, game theory, cryp- tography, and simulations. These applications require numbers that behave as if they were random, yet can be reproduced when needed. This is where PRNGs come in—they allow for repeatable randomness, making them ideal for controlled experiments, testing, and security. This chapter will explore the key concepts behind PRNGs. Before going into the mechanics of these generators, it is important to first understand what 'random' means and the characteristics that define truly random numbers.

5.1 Properties of PRNGs

The quality of a PRNG is determined by several key factors that influence its use for different applications. Some of the properties of a good PRNG is properties: Independency, a large period and reproducibility The numbers produced by the PRNG should be statistically independent, en- suring that each generated value exhibits no correlation with previous numbers or other sequences. This implies that knowledge of previously generated num- bers or sequences provides no advantage in predicting the next output. A PRNG operates within a specific interval before its sequence begins to repeat. A high-quality PRNG has a long interval, delaying repetition and enhancing its unpredictability. Conversely, a PRNG with a shorter period becomes more pre- dictable and less suitable for practical use. A key feature of a PRNG is its ability to reproduce the same sequence of num- bers when given a specific seed. This property is particularly useful in testing and simulation scenarios, where it is essential to generate identical sequences multiple times for consistency and reproducibility In addition, a PRNG must be fast and efficient to prevent it from introducing performance bottlenecks within an application. The speed of number generation directly impacts computational efficiency, especially in applications requiring a large volume of random numbers. An inefficient PRNG can significantly slow down processes, undermining the overall performance of the system. Therefore, balancing randomness and efficiency is essential for practical applications

5.2 Linear Congruential Generator

Linear Congruential Sequence (LCS) is a commonly used approach to generate pseudo-random numbers. LCS generates a sequence of numbers using a linear recurrence relation LCS is expressed as:

$$X_{n+1} = (aX_n + c) \bmod m.$$

where X_0 is the seed, a is the multiplier, c is the increment, and m is the modulus.

Example: Given $a = 5$, $c = 1$, $m = 16$, and $X_0 = 7$:

$$X_1 = (5 \cdot 7 + 1) \bmod 16 = 4$$

$$X_2 = (5 \cdot 4 + 1) \bmod 16 = 5$$

$$X_3 = (5 \cdot 5 + 1) \bmod 16 = 10$$

$$X_4 = (5 \cdot 10 + 1) \bmod 16 = 3$$

This sequence has a period of 16. In an LCG, the period can be as large as m, but choosing parameters carefully is crucial to achieving long periods.

6 Monte Carlo

6.1 Assumptions

7 Problem statement

8 Classical regression

9 Monte Carlo regression

10 Comparison between regressions

11 Discussion

12 Conclusion

13 Litteratur