MATH 414: STOCHASTIC SIMULATION

Rafiki's Notes

Rafael Barroso
Ingenierie Mathematique
École Polytechnique Fédérale de Lausanne
September 20, 2025

Contents

1 Pseudo Random Number Generators

In this section, we explore how random (you'll see that they're not so random) numbers are generated in our computers and how the programs designed for creating these numbers work.

Definition 1.1. A random number generatorn (RNG) is a procedure that produces an *infinite* stream of independent, identically distributed (i.i.d.) random variables $U_1, U_2, \ldots \sim \mu$ accoording to some *probability distribution* μ.

Recall that a probability dustribution is basically a fancy way of saying "how likely different things are to happen." Imagine you have a bag of Skittles, and you want to know the chances of grabbing each color. A probability distribution is like a chart or rulebook that says: 'Red has a 30% chance, green has 20%, purple is rare like a unicorn at 5%, etc.'

Remark 1.2. Note that an RNG is a *uniform* random number generator if μ is the *uniform distribution* in $(0, 1)$. Where the uniform distribution is the "everyone gets a fair share" version of probability. Every outcome has the same chance of happening.

Since all of the current RNG's are based on algorithms, they produce a *purely deterministic* (i.e. the outcome is already locked in once you know the rules) stream of variables $U_1, U_2, U_3, \ldots$ which 'look like' a stream of i.i.d. random variables. For this reason, algorithmic generators are called pseudo-random number generators. Here's a lil' sexy pseudo code for how such implementation of an RNG could go about:

Let $\mathcal{S}$ be a finite state space and $\mathcal{U}$ be the output space. Let $f : \mathcal{S} \to \mathcal{S}$ (i.e. a function that 'evolves' the system, goes from one state to annother) and $g : \mathcal{S} \to \mathcal{U}$ be given functions.

Algorithm: Pseudo-RNG

1. take $X_0 \in \mathcal{S}$ // seed
2. for $k = 1, 2, \ldots$ do
3. $X_k = f(X_{k-1})$ // recursion on state variable $X_k \in \mathcal{S}$
4. $U_k = g(X_k)$ // output $U_k \in \mathcal{U}$
5. end

Remark 1.3. Important notions to notice:

- X_0 is called the seed.
- A Pseudo-RNG starting from a given seed will always produce the same sequence $U_1, U_2, \ldots$.
- Since the state space $\mathcal{S}$ is finite, the generator eventually revisits an already visited state. All Pseudo-RNGs are *periodic* (i.e. eventually return to the initial state).
- Good generators have period $p = |\mathcal{S}|$.

Given the information we now know, we can *compress* some of the 'good' qualities a uniform pseudo RNG might possess:

- Have a large period
- Pass a battery of statistical tests for uniformity and independence.
- Be fast and efficient
- Be reproducible
- Have the possibility to generate multiple streams.
- Avoid producing the numbers 0 and 1.

We would now like to evaluate if these generators produce pure bullshit or if they're actually giving us sequences that resemble what we're looking for; an i.i.d. sequence of random variables accoording to some probability distribution.

Remark 1.4. The Cummulative Distribution Function (CDF) denotes the probability that a random variable X is less than or equal to some value x. Formally speaking,

$$\text{CDF}(X) := \mathbb{P}(X \leq x)$$

In other words, instead of just looking at one outcome, it keeps a running total of probabilities up to that point.

- It always starts at 0 (nothing less than the smallest value has happened yet).
- Ends at 1 (because by the time you've included all possible values, the probability is 100%).

So, while a probability distribution $\mathbb{P}(X = x)$ shows "the chance of this exact outcome," the CDF shows "the chance of being at or below this outcome."

The framework we're working with is the following. Let $U \in I \subset \mathbb{R}$ be a random variable with $\text{CDF}(X) = F(x) = \mathbb{P}(U \leq x)$. We define $U := \{U_1, \ldots, U_n\}$ to be a sample produced by a RNG with empirical distribution function $\hat{F}(x)$.

Remark 1.5. The empirical distribution function is basically the "data-based version" of the cumulative distribution function (CDF). Instead of coming from a theoretical model, it is built directly from a sample of data.
For a sample $X_1, X_2, \ldots, X_n$, the empirical distribution function is:

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} = \frac{\text{number of } \{U_i \leq x\}}{n}$$

where n = numaber of data points, $\mathbb{1}_{\{X_i \leq x\}}$ is an indicator function (it equals 1 if $X_i \leq x$, and 0 otherwise). So, $\hat{F}(x)$ is just the proportion of sample points that are less than or equal to x.

1.1 Testing, testing and more testing

One of the things statisticians love to do is to test the hypothesis H_0 that U has been drawn independently from the distribution F. This would tell us if each U_i does not depend form any other $U_j(\ \forall i \neq j)$ and so, giving us confidence that our data is 'random' in some sense. Annother thing we would also like to know, is to decide if our empirical data U generated by a RNG, actually 'fits' with the theoretical data. In other words, we would like to see if U matches F. We begin by exploring some non-parametric (i.e. we do not rely on parameters like mean or variance, just pure raw data son) goodness-of-fit tests for a given sample sequence U.

1. **Q-Q Plot (Quantile v. Quantile plot):** Imagine you have two classes of students. You suspect they both come from a population with the same height distribution. To check, you do the following:

   (a) Line up all the students in **Class A** (your generated data) from shortest to tallest.

   (b) Do the same for **Class B** (the perfect, theoretical distribution).

   (c) Compare them one-on-one: the shortest from A vs. the shortest from B, the second shortest vs. the second shortest, and so on.

   If you plot their heights against each other (Class A's heights on the y-axis, Class B's on the x-axis), and they really do have the same distribution, the points should form a perfectly straight diagonal line. The shortest students are about the same height, the median students are the same height, and the tallest students are the same height. If, however, Class A was secretly full of basketball players, their heights would be consistently taller than Class B's for the same rank, and the line would curve away (the Q-Q plot does exactly this, but with data points instead of people!).

   A Q-Q plot compares the quantiles of our data's empirical distribution $\hat{F}_n$ versus the quantiles of the theoretical distribution F.

Definition 1.6. A *quantile* is just a fancy word for a point below which a certain fraction of the data falls. For instance, the 0.25 quantile (or 25th percentile) is the value that is greater than 25% of the data.
Mathematically, the lecture defines the quantile from our data as the estimator $\hat{q}_p = U^{(j)}$ and the exact theoretical quantile as $q_p = \text{argmin}_x\{F(x) \geq p\}$, where $p = \frac{j}{n+1}$. We plot $\hat{q}_p$ (on the y-axis) against q_p (on the x-axis).

The Verdict After plotting all the points, we simply look at the result:

- If the points form a nice, straight line along $y = x$, we pop the champagne! Our data is a good fit for the theoretical distribution. We can trust our random number generator.
- If the points deviate from the line in a systematic pattern (like an S-curve or a banana shape), our generator is fucking up. The shape of the deviation can even tell us *how* our data is different (e.g., if it has "heavier tails").

2. **Kolmogorov-Smirnov Test:** This non-parametric statistical tool is used to determine if a given data sample comes from a specific continuous distribution. In essence, it quantifies the maximum vertical distance between the distribution of our data and the ideal distribution we are comparing it against.

The Null Hypothesis (H_0)

The core assumption we are testing is the null hypothesis, H_0: The data sample $(U_1, \ldots, U_n)$ was drawn independently from the specified theoretical continuous distribution with CDF $F(x)$.

The Test Statistic (D_n)

The test is built around the K-S statistic, D_n, which measures the largest absolute difference between the **empirical CDF ($\hat{F}_n(x)$)** of the sample and the **theoretical CDF ($F(x)$)**. The test statistic is then the supremum (for a finite sample, the maximum) of the absolute difference across all possible values of x:

$$D_n = \sup_x |\hat{F}_n(x) - F(x)|$$

The Decision Rule

Under the null hypothesis, the distribution of the scaled statistic, $\sqrt{n}D_n$, converges to a known distribution called the Kolmogorov distribution, regardless of what $F(x)$ is.

We reject the null hypothesis H_0 at a significance level α if our calculated statistic is greater than a critical value K_α obtained from tables of the Kolmogorov distribution.

$$\text{If } \sqrt{n}D_n > K_\alpha, \text{ reject } H_0.$$

If we reject H_0, we have statistical evidence that our data sample does not follow the theoretical distribution F. Otherwise, if $\sqrt{n}D_n \leq K_\alpha$, we fail to reject H_0, meaning our data is consistent with the proposed distribution.

3. χ^2 **Test:** This one is particularly useful for binned or categorical data. Essentially, it compares the histogram of the sample with the exact one.

> **Remark 1.7.** Recall that a histogram is a type of bar chart that graphically shows the distribution of quantitative data by displaying the frequency of data points within predefined bins or intervals.

The Null Hypothesis (H_0)

The hypothesis being tested is, H_0: The observed data is drawn from the specified theoretical distribution. In other words, the observed frequencies are consistent with the expected frequencies.

The Test Statistic ($\hat{Q}_m$)

To calculate the statistic, we first partition the data's range into $m+1$ discrete bins or classes, I_j. We then define:

- N_j: The **Observed Frequency**, which is the count of data points from our sample that fall into bin j.
- p_j: The **Theoretical Probability** that any single random value from the distribution would fall into bin j. For a uniform distribution, p_j is the same for all bins of equal size.
- np_j: The **Expected Frequency**, which is the theoretical count of data points we expect to see in bin j for a sample of size n.

The χ^2 test statistic, denoted $\hat{Q}_m$, is the sum of the squared differences between observed and expected frequencies, normalized by the expected frequency for each bin:

$$\hat{Q}m = \sum j = 1^{m+1} \frac{(N_j - np_j)^2}{np_j}$$

A large value of $\hat{Q}_m$ implies a significant discrepancy between the observed data and the theoretical model.

The Decision Rule

Under the null hypothesis, the test statistic $\hat{Q}_m$ asymptotically follows a **Chi-Squared distribution with m degrees of freedom** (where m is the number of classes minus one).

We reject the null hypothesis H_0 at a significance level α if our calculated statistic is greater than a critical value $q_{1-\alpha}$ from the $\chi^2(m)$ distribution table.

$$\text{If } \hat{Q}m > q_{1-\alpha}, \text{ reject } H_0.$$

Rejecting H_0 means the observed frequencies are significantly different from the expected ones, suggesting the data does not follow the theoretical distribution. If $\hat{Q}m \leq q_{1-\alpha}$, we fail to reject H_0, indicating that the data is a good fit.

We now explore some more tests in order to help us deretmine if our data $U := \{U_1, \ldots, U_n\}$ (generated by a unifrom RNG) is mutually independent i.e. two or more random variables are independent from one annother. For these tests, we consider our null hypothesis H_0; that $\{U_i\}_i$ are mutually independent and uniformly distributed in $(0, 1)$.

1. **Serial Test:** Group our data into k blocks of lenght d (such that $kd = n$) i.g. $V_j := \{U_{(j-1)d+1}, \ldots, U_{jd}\}$ for $j = 1, \ldots k$. Once we've divvy'd up our sampled data, we test (using any of our previous defined tests) to see if $V := \{V_1, \ldots, V_k\}$ has a joint uniform distribution.

2. **Gap Test:** Let $T_1, T_2, \ldots$ be the times when the sequence visits a given interval (α, β). The gap length between consecutive visits is $Z_i = T_i - T_{i-1} - 1$. Under H_0, Z_i are iid with a geometric distribution with parameter $\rho = \beta - \alpha$, i.e., $\mathbb{P}(Z = j) = \rho(1 - \rho)^j$ for $j = 0, 1, 2, \ldots$. One can then use a χ^2 test to check if the $\{Z_i\}_i$ have the correct geometric distribution.