# Some questions are marked with \*\*. These are more challenging and may require a few minutes of research beyond what was covered in the course. Feel free to skip them—they're intended for interested readers who want to explore further.

# Probability

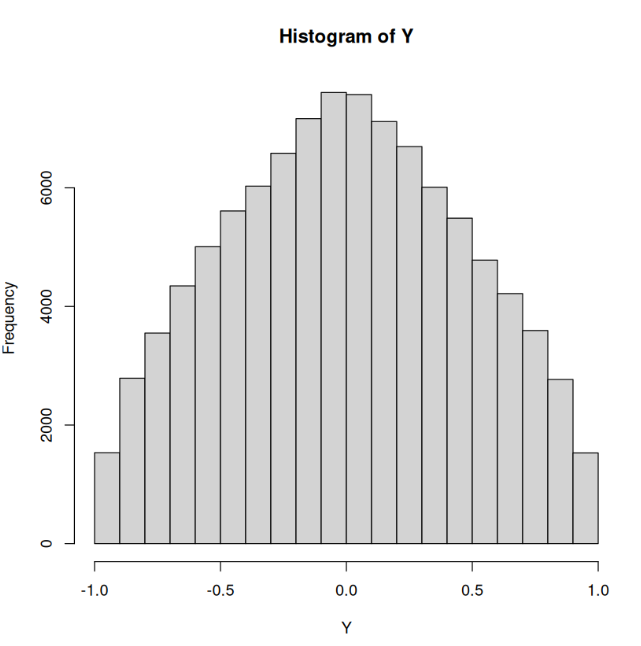
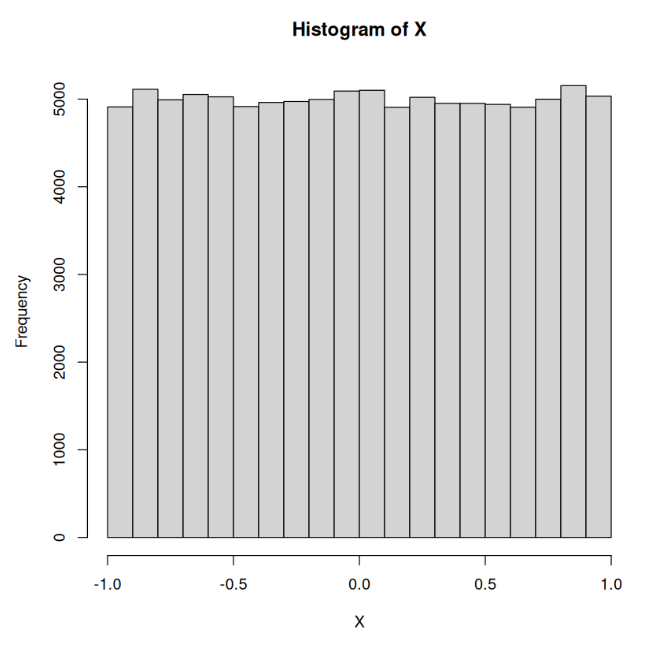
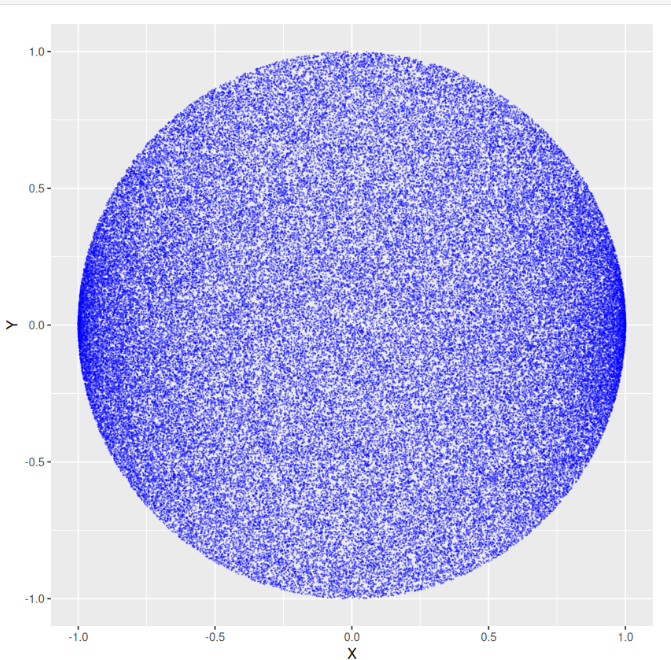
**1) \*\* 20 points**

We want to create a suitable distribution for a circle centered at the origin with a radius of 1 so that we can select a point on this surface with equal probability. Before solving this problem in the correct way, we ask you to reject the two methods that we will describe below using python code. Then, solve the problem by correctly choosing random variables. (Note that the output for this question should include 3 explanations and 3 pieces of code, where each output should be an image or graph that demonstrates whether the chosen method for the random variables is efficient.)

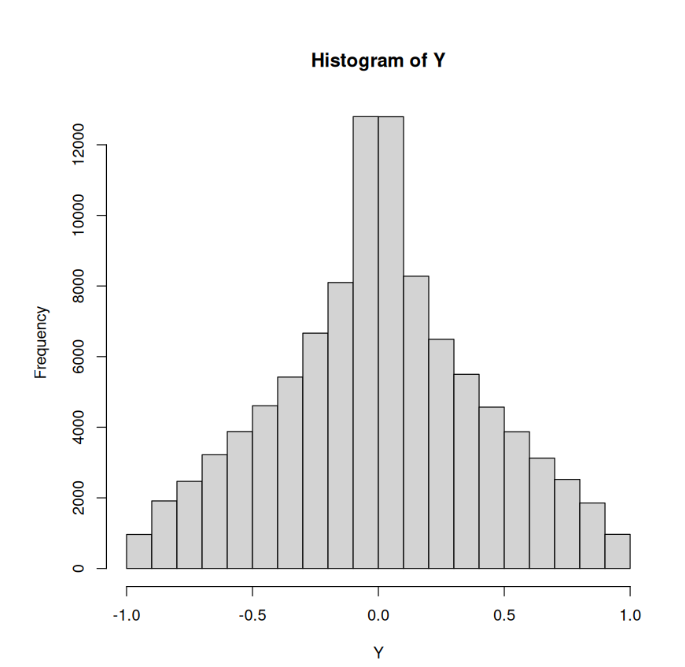
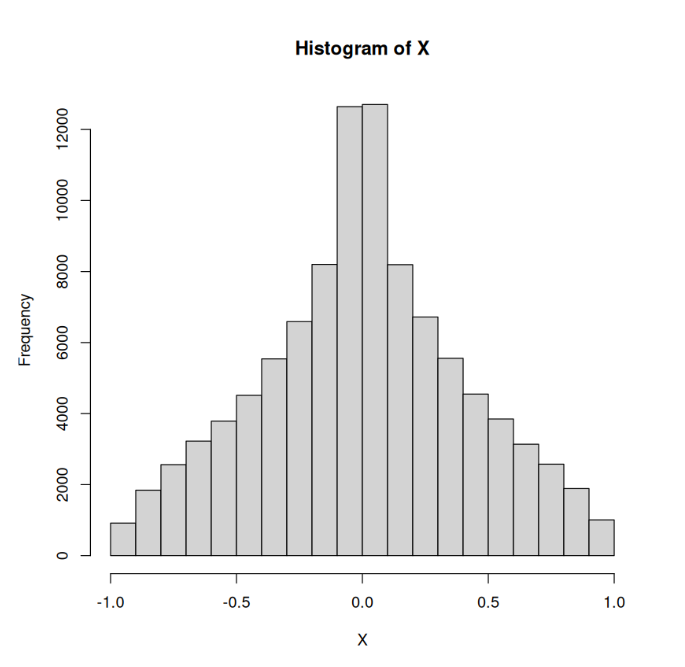
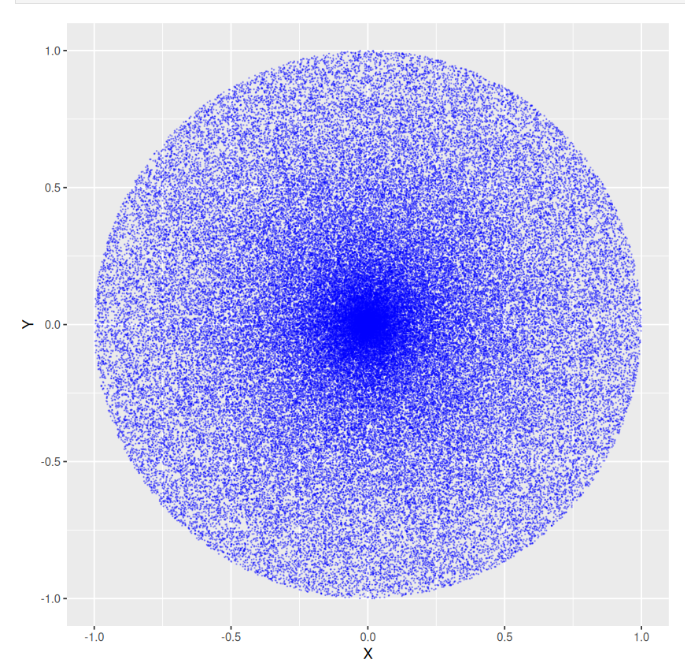
1. Both uniformly distributed in the ranges:
2. Both uniformly distributed in the ranges:

These are the expected outputs:

a)



b)



1. theoretically prove that method a and b are not correct ways to uniformly sample from a circle centered at origin.

**2) 20 points**Load the data from [two\_currencies.csv](https://drive.google.com/file/d/1lNvmxD7wFUqTwfoO_xRXQgKkK3gWH-j4/view?usp=sharing) file.  
For 20 days, you have invested $10,000 each morning at 7 a.m. in cryptocurrency A and sold it each evening at 6 p.m. Each day, you experienced some profit or loss. For example, on the first day, you lost $2, meaning you sold cryptocurrency A for $9,998.

Similarly, you have invested $10,000 each morning at 7 a.m. in cryptocurrency B and sold it each evening at 6 p.m. On the first day, you made a profit of $482, selling cryptocurrency B for $10,482.

Assume that the profit from trading cryptocurrencies A and B each day is independent of other days, meaning there is no specific trend over time. In other words, the daily profits for A and B are independently and identically distributed (i.i.d.) samples from random variables ​ and ​, which represent the profit distribution per dollar invested in A and B, respectively. While the daily profits are independent over time, there may be a correlation between A and B.

You have a total budget of $20,000 to invest in both cryptocurrencies.

1. Your goal is to determine how to allocate this budget between A and B to maximize your expected profit. Expected profit of a portfolio is defined as with the constraint that
2. Your goal is to determine how to allocate this budget between A and B to minimize your investment risk. Risk is defined as the variance of your portfolio, calculated as , where ​ is the amount invested in A and ​ is the amount invested in B, with the constraint that   
   Hint: Use grid search to explore different combinations of ​ and ​ to find the optimal investment split.You can also solve it as an optimization problem instead of grid search.
3. \*\* find a portfolio that has maximum expected profit, while having maximum 0.0148 variance. Use grid search to explore different combinations of ​ and ​ to find the optimal investment split. You can also solve it as an optimization problem instead of grid search.

# Statistics

In this part you should answer the questions based on [this dataset.](https://archive.ics.uci.edu/dataset/320/student+performance)

You can load the data from [student\_performance.csv](https://drive.google.com/file/d/15AjF3I9XADapTBjlMvlrb5fL7ZhlWgcg/view?usp=sharing) file.

**3) 5 points**

Is there a significant difference in average G1 grade between males and females ? answer this question with two approaches:

1. Check the correlation between gender of students and their G1 grades. You can convert female to 1 and male to 0.
2. Run a statistical test to compare average G1 grade between males and females, and analyze the result with 0.05 significance level. What statistical test would be appropriate here?

**4) 5 points**

Is there a significant difference between G1 and G2 grades ? What statistical test would be appropriate here?

**5) 5 points**

use bootstrapping to estimate a 0.9 confidence interval for average of G3 grades. You should decide whether parametric or non-parametric bootstrap fits better to your problem.

**6) 5 points**

use CLT to estimate 0.9 confidence interval for average of G3 grades.

**7) 3 points**

compare your results from question 5 and 6, and justify the results.

**8) \*\* 5 points:** read about QQ plot and draw QQ plot of the bootstrapped average G3 grades and assess if it has a normal distribution. in addition use QQ plot on G3 grades (not the averages) to assess whether it is normal.

**9) 5 points**  
Select an appropriate likelihood function for modeling the number of absences and estimate its parameter(s) using Maximum Likelihood Estimation (MLE).

**10) 10 points:**  
Fit an exponential distribution to the number of absences and compare the likelihood of this fitted exponential distribution with the likelihood of the distribution you estimated in question 9. Based on these likelihoods, which distribution provides a better fit for the data? Hint: To compare the likelihoods of a continuous probability density function (PDF) with a discrete dataset, consider each discrete data point as belonging to an interval. For example, you might treat the value 2 as falling within the interval [1.5, 2.5). Explain why this approach is needed when comparing likelihoods of a discrete dataset, derived from a continuous PDF with those from a discrete PDF.

**11) 2 points**

plot your likelihood function on top of the empirical PDF of the absences and compare them.

**12) 2 points**

plot CDF of your likelihood function, on top of the empirical CDF of the absences and compare them.

**13) 2 points**

Based on the two previous answers, assess whether the distribution you considered for the number of absences properly explains its distribution.

**14) 5 points**

Is there dependence between father job (Fjob) and family size (famsize) in the data? What hypothesis test can you run to assess that assumption ? report the result of your hypothesis test with 0.05 significance level.

**15) 5 points**

Among the quality of family relationships(famrel), free time after school(freetime), going out with friends(goout), workday alcohol consumption(Dalc), weekend alcohol consumption(Walc) and current health status(health), which one has the strongest relationship with G3 grades and in which direction?

Hint: Covariance and correlation are key concepts to explore relationships between variables. How can you visualize this relationship?

**16) 5 points**

We are willing to know if weekend alcohol consumption(Walc) has greater than -0.2 correlation with G3 grades (greater in magnitude). use bootstrapping to estimate p-value for the correlation between Walc and G3 grades and report the result of your hypothesis test with a 0.05 significance level. If your sample size was 100 students, do you expect to get a bigger or smaller p-value ?

**17) \*\* 5 points**  
points: can we use CLT to estimate confidence intervals for correlation ?

**18) 5 points**

Simple Linear Regression: Build a simple linear regression model to predict G3 using the variable you identified in the previous step.

Interpretation: Report and interpret the model's intercept and slope. What do they mean in the context of this data?

**19) 15 points**We want to estimate the proportion of students who want to pursue higher education (column 'higher'). Assume a Beta(1,1) distribution for the probability that a random student wants to pursue higher education. Using simulation, estimate the PDF of the posterior distribution. In the simulation, you cannot use the fact that the Beta distribution is a conjugate prior to the Bernoulli distribution. First compute the posterior distribution using the simulation and then use the fact that the Beta distribution is a conjugate prior to the Bernoulli distribution and compute the theoretical posterior distribution, and finally compare these results.

steps:

1. Compare the Maximum A Posteriori (MAP) estimate with the Maximum Likelihood Estimate (MLE). Is there a meaningful difference? Why?
2. Compute the prior and likelihood of the data for each prior in np.linspace(0.001,0.999,1000).
3. Calculate the Bayes numerator, then normalize it to compute the posterior distribution.
4. To prevent floating-point issues that can arise from multiplying many likelihoods less than 1, use the sum of the log likelihoods instead. Afterward, apply np.exp to obtain the correct posterior probabilities. Why is this necessary?
5. Compare the MAP and MLE. Discuss the **bias and variance** of the MLE versus the MAP estimate with this Beta(1,1) prior. Which estimator do you expect to have lower variance, and why might the MAP be considered biased?
6. Compute the theoretical distribution and compare the result with your simulation result.

**20) \*\* 5 points**

Use Beta(500,500) as prior and compare MAP with MLE. Is there a meaningful difference ? Why? Explain how this strong prior influences the **bias and variance** of the MAP estimate compared to the MLE.

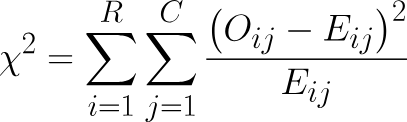
**21) 20 points**this question, you will:

1. Empirically demonstrate that the Chi-squared test statistic for contingency tables follows a Chi-squared distribution with  
    Degree of freedom =(number of rows−1)(number of columns−1)  
    under the null hypothesis of independence.
2. Use your empirically obtained distribution to perform a hypothesis test in a case where the null hypothesis is false.

Part 1 – Building the Empirical Chi-Squared Distribution

Background:

The Chi-squared test statistic for a contingency table is given by:



Since we haven’t proved this mathematically, you’ll validate it empirically by simulating data under the null hypothesis (independence) to see if the values match the theoretical distribution.

where:

- is the observed frequency in cell ,

- is the expected frequency under the null hypothesis of independence,

- R is the number of rows, and C is the number of columns.

Under the null hypothesis, this statistic asymptotically follows a Chi-squared distribution with degrees of freedom df = (R-1)(C-1), provided the sample size is sufficiently large and expected frequencies are at least 5 in most cells.

**Steps:**

1. Choose table dimensions R and C (e.g., 3\*4; ensure R >=2 2 and C >=2 2 for meaningful degrees of freedom).

2. Define a null probability model assuming independence between rows and columns. For example:

* Specify marginal row probabilities (summing to 1) and column probabilities (summing to 1)
* The joint probability for cell (i,j) is

3. Fix a sample size n (e.g., n = 100; choose n large enough so that most to satisfy the test's assumptions).

4. Simulate at least 10,000 samples under the null model. For each sample:

* Generate a multinomial random sample of size n using the probabilities to obtain the observed counts .
* Compute the expected counts
* Calculate the statistic.

5. Plot a histogram of the simulated values (use appropriate binning for clarity).

6. Overlay the theoretical Chi-squared probability density function (PDF) with df = (R-1)(C-1) degrees of freedom on the histogram.

7. Discuss your results, including:

* Whether the empirical distribution approximates the theoretical one.
* How increasing the number of simulations (e.g., try 1,000 vs. 10,000 vs. 100,000) affects the convergence.
* Any deviations observed and potential reasons (e.g., small n violating assumptions).

Part 2 – Using the Empirical Distribution for Hypothesis Testing

1. Define an alternative probability model where rows and columns are \*\*not independent\*\* (ensuring the null hypothesis is false). For example:

* Start with the same marginal probabilities and from Part 1.
* Introduce dependence
* ensure probabilities sum to 1 and are non-negative; e.g., add positive values to some cells and subtract from others).
* Alternatively, use a specific dependent model like increasing probabilities in the diagonal cells.

2. Draw one sample of size n (same as in Part 1) from this alternative model to obtain observed counts

3. Compute the statistic for this sample, using the same formula as in Part 1 (with E\_i,j calculated under the assumption of independence).

4. Using your empirical distribution from Part 1 (not the theoretical one), compute the p-value: proportion of the null-simulated values that are greater than or equal to your observed statistic.

5. Determine whether you can reject the null hypothesis at significance level

6. Report your conclusion (e.g., "Reject" or "Fail to reject") and briefly discuss the result, including:

* The power of the test in this scenario (e.g., why it detected dependence).
* A comparison to the theoretical p-value (computed using the Chi-squared CDF) and any differences observed.
* Potential improvements, such as increasing n to better detect dependence.

Implementation Note: Reuse code from Part 1 where possible. Explain how you constructed the alternative model and why it represents dependence.