chandlerrc_hw2

September 18, 2023

1 Homework 2

1.1 References

- Lectures 4-8 (inclusive).

1.2 Instructions

- Type your name and email in the "Student details" section below.
- Develop the code and generate the figures you need to solve the problems using this notebook.
- For the answers that require a mathematical proof or derivation you should type them using latex. If you have never written latex before and you find it exceedingly difficult, we will likely accept handwritten solutions.
- The total homework points are 100. Please note that the problems are not weighed equally.

[ ]:
```
import matplotlib.pyplot as plt
%matplotlib inline
import matplotlib_inline
matplotlib_inline.backend_inline.set_matplotlib_formats('svg')
import seaborn as sns
sns.set_context("paper")
sns.set_style("ticks")

import numpy as np
import scipy
import scipy.stats as st
import urllib.request
import os

def download(
    url : str,
    local_filename : str = None
):
    """Download a file from a url.

    Arguments
    url            -- The url we want to download.
    local_filename -- The filemame to write on. If not
                      specified
```

```
    """
    if local_filename is None:
        local_filename = os.path.basename(url)
    urllib.request.urlretrieve(url, local_filename)
```

1.3 Student details

- **First Name:** Robert
- **Last Name:** Chandler
- **Email:** chandl71@gmail.com
- **Used generative AI to complete this assignment (Yes/No):** No
- **Which generative AI tool did you use (if applicable)?:** N/A

1.4 Problem 1 - Joint probability mass function of two discrete random variables

Consider two random variables X and Y. X takes values $\{0, 1, \ldots, 4\}$ and Y takes values $\{0, 1, \ldots, 8\}$. Their joint probability mass function, can be described using a matrix:

[ ]:
```
P = np.array(
    [
        [0.03607908, 0.03760034, 0.00503184, 0.0205082 , 0.01051408,
         0.03776221, 0.00131325, 0.03760817, 0.01770659],
        [0.03750162, 0.04317351, 0.03869997, 0.03069872, 0.02176718,
         0.04778769, 0.01021053, 0.00324185, 0.02475319],
        [0.03770951, 0.01053285, 0.01227089, 0.0339596 , 0.02296711,
         0.02187814, 0.01925662, 0.0196836 , 0.01996279],
        [0.02845139, 0.01209429, 0.02450163, 0.00874645, 0.03612603,
         0.02352593, 0.00300314, 0.00103487, 0.04071951],
        [0.00940187, 0.04633153, 0.01094094, 0.00172007, 0.00092633,
         0.02032679, 0.02536328, 0.03552956, 0.01107725]
    ]
)
```

The rows of the matrix correspond to the values of X and the columns to the values of Y. So, if you wanted to find the probability of $p(X = 2, Y = 3)$ you would do:

[ ]:
```
print(f"p(X=2, Y=3) = {P[2, 3]:.3f}")
```

```
p(X=2, Y=3) = 0.034
```

A. Verify that all the elements of P sum to one, i.e., that $\sum_{x,y} p(X = x, Y = y) = 1$.

[ ]:
```
np.allclose(P.sum(), 1)
```

[ ]:
```
True
```

B. Find the marginal probability density of X:

$$p(x) = \sum_y p(x, y).$$

You can represent this as a 1-dimensional vector of length 5.

[ ]:
```
P_x = P.sum(axis=1)
P_x
```

[ ]:
```
array([0.20412376, 0.25783426, 0.19822111, 0.17820324, 0.16161762])
```

C. Find the marginal probability density of Y. This is a 1-dimensional vector of length 9.

[ ]:
```
P_y = P.sum(axis=0)
P_y
```

[ ]:
```
array([0.14914347, 0.14973252, 0.09144527, 0.09563304, 0.09230073,
       0.15128076, 0.05914682, 0.09709805, 0.11421933])
```

D. Find the expectation and variance of X and Y.

[ ]:
```
vals_X = np.arange(5)
vals_Y = np.arange(9)

X = st.rv_discrete(values=(vals_X, P_x))
Y = st.rv_discrete(values=(vals_Y, P_y))

print(f"E[X] = {X.expect()}")
print(f"E[Y] = {Y.expect()}")

print(f"V[X] = {X.var()}")
print(f"V[Y] = {Y.var()}")
```

```
E[X] = 1.83535668
E[Y] = 3.69345081
V[X] = 1.8718956371793776
V[Y] = 7.1905898441103435
```

E. Find the expectation of $E[X + Y]$.

[ ]:
```
print(f"E[X + Y] = E[X] + E[Y] = {X.expect() + Y.expect()}")
```

```
E[X + Y] = E[X] + E[Y] = 5.52880749
```

F. Find the covariance of X and Y. Are the two variable correlated? If yes, are they positively or negatively correlated?

[ ]:
```
C_XY = 0

for i in range(vals_X.size):
    for j in range(vals_Y.size):
        x = vals_X[i]
        y = vals_Y[j]
        C_XY += (x - X.expect()) * (y - Y.expect()) * P[i, j]
```

```
C_XY
```

[ ]: 0.31867736582709294

The variables are positively correlated, but not strongly so

G. Find the variance of $X + Y$.

[ ]:
```
print(f"V[X + Y] = V[X] + V[Y] + 2C[X,Y] = {X.var() + Y.var() + 2 * C_XY}")
```

V[X + Y] = V[X] + V[Y] + 2C[X,Y] = 9.699840212943908

J. Find the probability that $X + Y$ is less than or equal to 5. That is, find $p(X + Y \leq 5)$. Hint: Use two for loops to go over all the combinations of X and Y values, check if $X + Y \leq 5$, and sum up the probabilities.

[ ]:
```
p_less_than_5 = 0
for i in range(vals_X.size):
    for j in range(vals_Y.size):
        x = vals_X[i]
        y = vals_Y[j]
        p_less_than_5 += P[x, y] if x + y <= 5 else 0

p_less_than_5
```

[ ]: 0.5345903099999999

1.5 Problem 2 - Zero correlation does not imply independence

The purpose of this problem is to show that zero correlation does not imply independence. Consider the random variable X and Y following a standard normal distribution. Define the random variable as $Z = X^2 + 0.01 \cdot Y$. You will show that the correlation between X and Z is zero even though they are not independent.

A. Take 100 samples of X and Z using numpy or scipy. Hint: First sample X and Y and use the samples to get Z.

[ ]:
```
X = st.norm()
Y = st.norm()

sample_X = X.rvs(size=100)
sample_Y = Y.rvs(size=100)

sample_Z = sample_X**2 + 0.01 * sample_Y
sample_Z
```

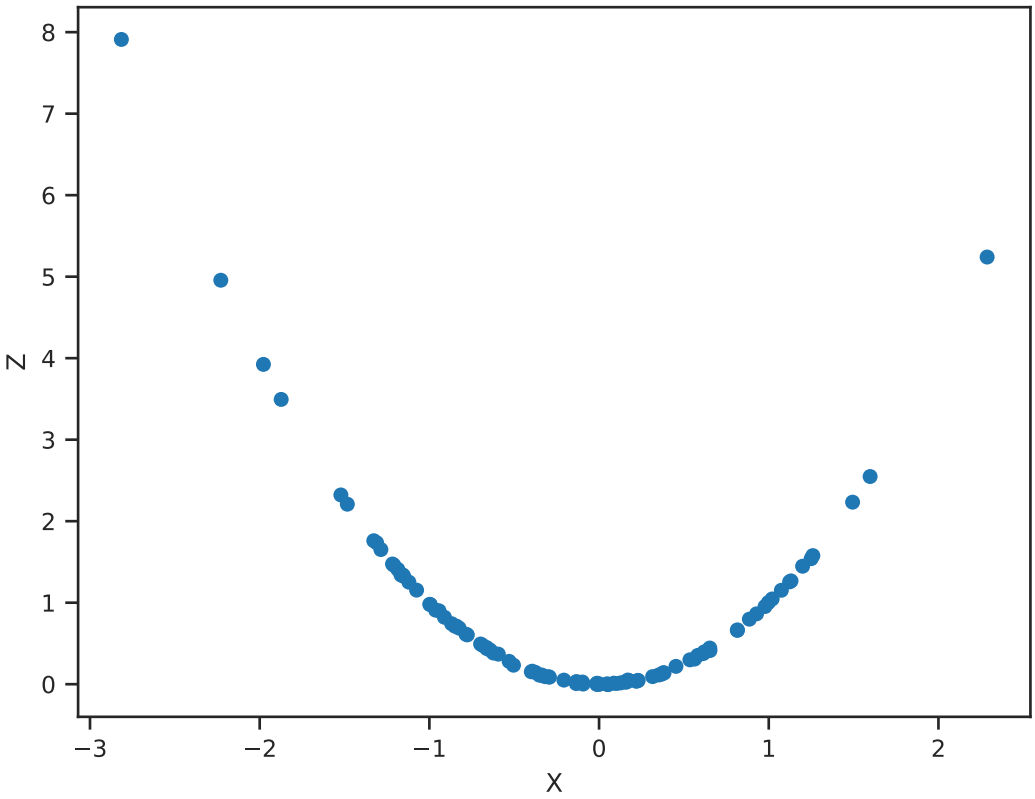
[ ]:
```
array([ 1.52418335e-01,  5.09340401e-02,  8.63360962e-02,  1.17689350e-02,
        9.53101944e-01,  7.52293915e-03,  4.49822287e-01,  2.32186077e+00,
        3.82624447e-01,  3.08349168e-01,  4.12871589e-01,  3.49396006e+00,
        7.91043880e+00,  1.59533813e-01,  7.11927057e-01,  8.21441636e-01,
```

```
1.47458841e+00,  1.45924058e+00,  1.26733599e+00,  9.22404586e-02,
1.44121746e-01,  1.60727866e-02,  1.65201288e+00,  2.99030736e-01,
5.24075928e+00,  6.65993885e-01,  1.37696814e-01,  1.12546453e-01,
4.43570835e-01,  1.41699868e-01,  1.32358332e+00,  6.88042148e-01,
1.01944311e-01,  9.31497954e-02,  2.54804452e+00, -1.21466864e-03,
4.70634553e-02, -4.41976918e-03,  4.26043938e-01,  2.80292262e-01,
1.40861904e+00,  9.80380731e-01,  4.77876695e-01,  2.65142349e-02,
3.53125269e-01,  1.13073171e-01,  9.08750974e-01,  4.95616358e+00,
1.25715129e-02,  1.00182301e+00,  3.58047475e-02,  6.60599027e-01,
3.83741448e-03,  2.33128052e-01,  1.15409144e+00,  3.26029402e-02,
1.25208000e+00,  4.27900097e-02,  4.37152438e-01,  1.73555715e+00,
1.04575575e+00,  9.00475049e-01,  4.93231269e-01,  4.19845865e-01,
2.23328893e+00, -3.24320159e-03,  1.76023087e+00,  5.01416275e-02,
2.20846120e+00,  6.12948390e-01,  2.18808257e-01,  9.75519616e-01,
6.05130551e-01,  7.12183582e-01,  1.25550760e+00,  1.54019477e+00,
4.39221706e-03,  3.92371357e+00,  1.10910871e-01,  7.97910545e-01,
8.62883236e-01,  2.54769437e-02,  1.44740148e+00,  3.76291975e-01,
3.12617263e-03,  7.42125765e-01,  8.60009823e-04,  3.68743452e-01,
9.12811905e-02,  1.57762001e+00,  1.23398554e-01,  1.91409186e-03,
1.33955773e+00,  1.34047862e+00,  2.73158751e-02,  4.47600117e-02,
1.15230939e+00,  3.96599141e-01,  8.86390768e-03, -4.57307897e-03])
```

B. Do the scatter plot between X and Z.

[ ]:
```
fig = plt.scatter(sample_X, sample_Z)
fig.axes.set_xlabel("X")
fig.axes.set_ylabel("Z")
```

[ ]:
```
Text(0, 0.5, 'Z')
```



C. Use the scatter plot to argue that X and Z are not independent.

Answer:
There is a clear dependence between X and Z where the points sampled from these distributions follow a parabolic shape when plotted in a cartesian plane. If X and Z were independent, we would expect not to see any such pattern, but rather an amorphous blob of points.

D. Use the samples you took to estimate the variance of Z.

[ ]:
```
# np.var() is biased by default but since n is 100, this should only affect our answer by ~1%
print(f"A (biased) estimate of the sample variance of Z is V[Z] = {sample_Z.var()}")
```

```
A (biased) estimate of the sample variance of Z is V[Z] = 1.4587942132124894
```

E. Use the samples you took to estimate the covariance between X and Z.

[ ]:
```
cov_XZ = np.cov(sample_X, sample_Z)[0, 1]
# np.cov() is unbiased by default... see comment in cell above
print(f"An (unbiased) estimate of the covariance between X and Z is: C[X, Z] = {cov_XZ}")
```

```
An (unbiased) estimate of the covariance between X and Z is: C[X, Z] =
-0.33956302372120456
```

F. Use the results above to find the correlation between X and Z.

[ ]:
```
# np.corrcoef(sample_X, sample_Z)[0,1]
cov_XZ / np.sqrt(sample_X.var() * sample_Z.var())
```

[ ]:
```
-0.31458926500506607
```

G. The correlation coefficient you get may not be very close to zero. This is due to the fact that we estimate it with Monte Carlo averaging. To get a better estimate, we can increase the number of samples. Try increasing the number of samples to 1000 and see if the correlation coefficient gets closer to zero.

[ ]:
```
sample_X = X.rvs(size=1000)
sample_Y = Y.rvs(size=1000)

sample_Z = sample_X**2 + 0.01 * sample_Y

np.corrcoef(sample_X, sample_Z)[0, 1]
```

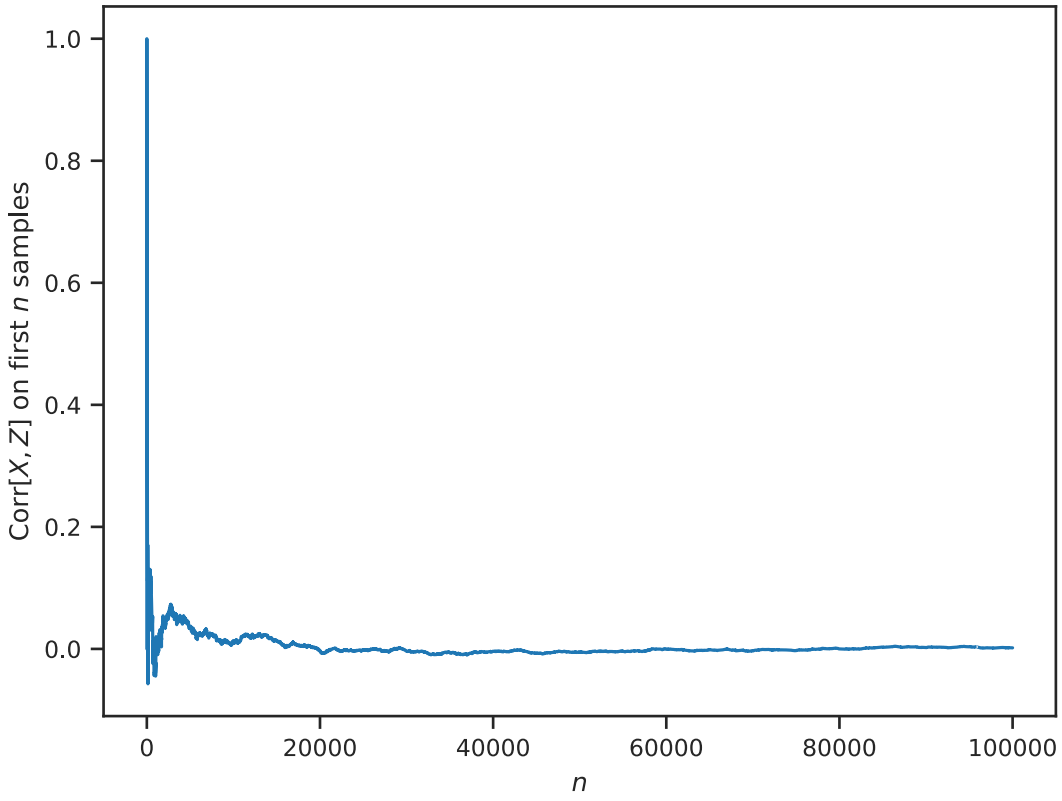
[ ]:
```
0.021251624898300996
```

H. Let's do a more serious estimation of Monte Carlo convergence. Take 100,000 samples of X and Z. Write code that estimates the correlation between X and Z using the first n samples for $n = 1, 2, \ldots, 100,000$. Plot the estimates as a function of n. What do you observe? How many samples do you need to get a good estimate of the correlation?

[ ]:
```
%%time

n = 100_000
sample_X = X.rvs(size=n)
sample_Y = Y.rvs(size=n)

sample_Z = sample_X ** 2 + 0.01 * sample_Y

correlation_XZ_n = np.zeros(n)

for i in range(n):
    # Two scalar values cannot really have a correlation since they each
    # have zero variance, so we would end up dividing by zero
    if i == 0:
        correlation_XZ_n[i] = 0
        continue

    correlation_XZ_n[i] = np.corrcoef(sample_X[:i + 1], sample_Z[:i + 1])[0,1]
```

```
CPU times: user 28.6 s, sys: 8.63 ms, total: 28.6 s
Wall time: 28.6 s
```

```
[ ]: plt.plot(np.arange(n) + 1, correlation_XZ_n)
     plt.xlabel("$n$")
     plt.ylabel(r"$\operatorname{Corr}[X,Z]$ on first $n$ samples");
```



The correlation value is quite unstable at the very beginning and continues to be for the first 5000 iterations or so, but as n increases, we see the correlation converge closer and closer to its true value of zero. The number of samples required to get a good estimate of the correlation depends on how "good" is defined, but around the 20,000 mark, the jitter of the value seems to have mostly diminished and hovers fairly stably around the true value.

1.6 Problem 3 - Creating a stochastic model for the magnetic properties of steel

The magnetic properties of steel are captured in the so-called $B-H$ curve, which connects the magnetic field H to the magnetic flux density B. The $B-H$ curve is a nonlinear function typically measured in the lab. It appears in Maxwell's equations and is, therefore, crucial in the design of electrical machines.

The shape of the $B-H$ curve depends on the manufacturing process of the steel. As a result, the $B-H$ differs across different suppliers but also across time for the same supplier. The goal of this problem is to guide you through the process of creating a stochastic model for the $B-H$ curve using real data. Such a model is the first step when we do uncertainty quantification for the design of electrical machines. Once constructed, the stochastic model can generate random samples of

the $B-H$ curve. We can then propagate the uncertainty in the $B-H$ curve through Maxwell's equations to quantify the uncertainty in the performance of the electrical machine.

Let's use some actual manufacturer data to visualize the differences in the $B-H$ curve across different suppliers. The data are here. Explaining how to upload data on Google Colab will take a while. We will do it in the next homework set. You should know that the data file `B_data.csv` needs to be in the same working directory as this Jupyter Notebook. I have written some code that allows you to put the data file in the right place without too much trouble. Run the following:

[ ]:
```
url = "https://github.com/PredictiveScienceLab/data-analytics-se/raw/master/
 ↪lecturebook/data/B_data.csv"
download(url)
```

If everything worked well, then the following will work:

[ ]:
```
B_data = np.loadtxt('B_data.csv')
B_data
```

[ ]:
```
array([[0.        , 0.00490631, 0.01913362, …, 1.79321352, 1.79337681,
        1.79354006],
       [0.        , 0.00360282, 0.01426636, …, 1.8367998 , 1.83697627,
        1.83715271],
       [0.        , 0.00365133, 0.01433438, …, 1.77555287, 1.77570402,
        1.77585514],
       …,
       [0.        , 0.00289346, 0.01154411, …, 1.7668308 , 1.76697657,
        1.76712232],
       [0.        , 0.00809884, 0.03108513, …, 1.7774044 , 1.77756225,
        1.77772007],
       [0.        , 0.00349638, 0.0139246 , …, 1.76460358, 1.76474439,
        1.76488516]])
```

The shape of this dataset is:

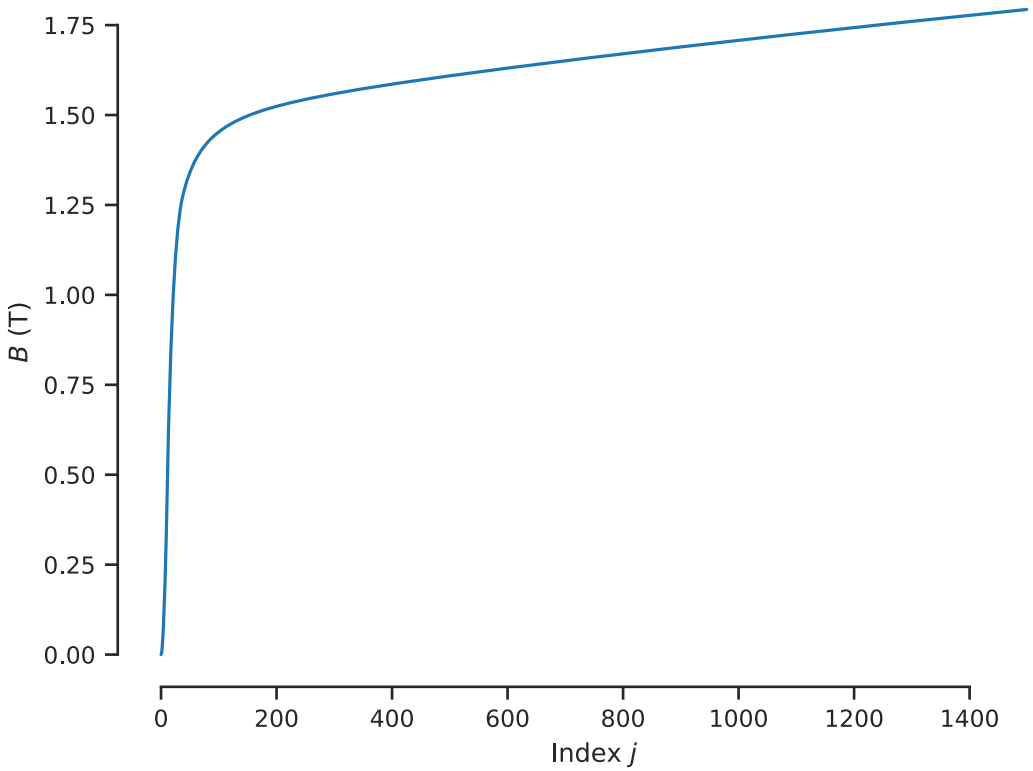
[ ]:
```
B_data.shape
```

[ ]:
```
(200, 1500)
```

The rows (200) correspond to different samples of the $B-H$ curves (suppliers and times). The columns (1500) correspond to different values of H. That is, the i,j element is the value of B at the specific value of H, say H_j. The values of H are equidistant and identical; we will ignore them in this analysis. Let's visualize some of the samples.

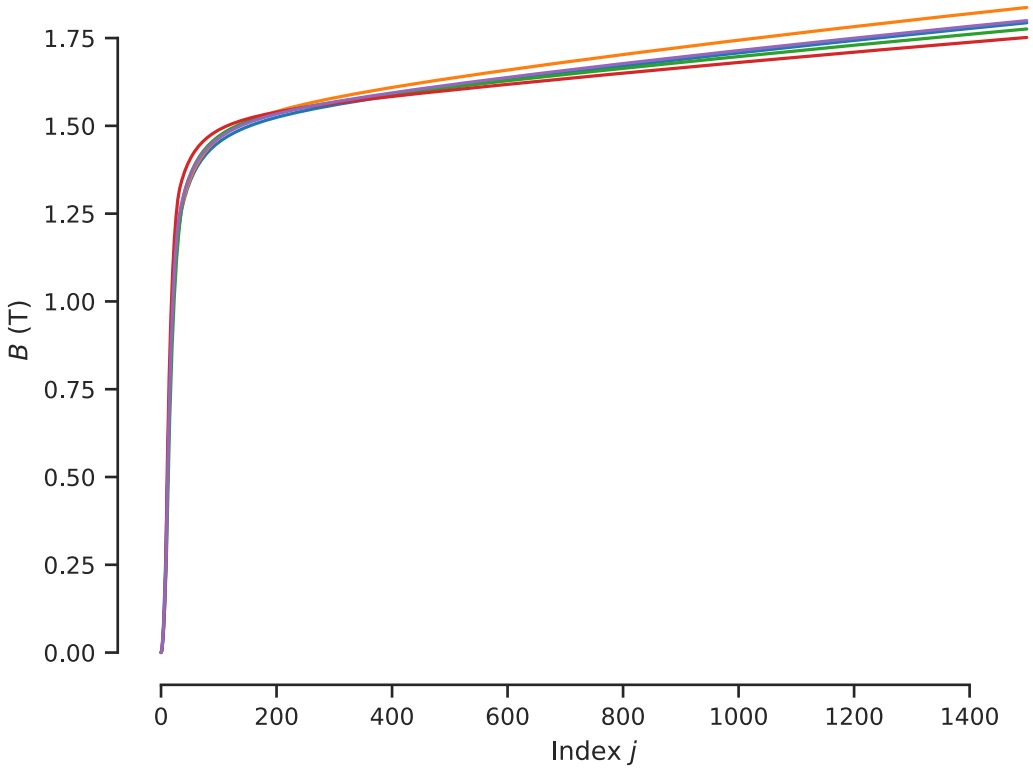
Here is one sample:

[ ]:
```
fig, ax = plt.subplots()
ax.plot(B_data[0, :])
ax.set_xlabel(r"Index $j$")
ax.set_ylabel(r"$B$ (T)")
sns.despine(trim=True);
```



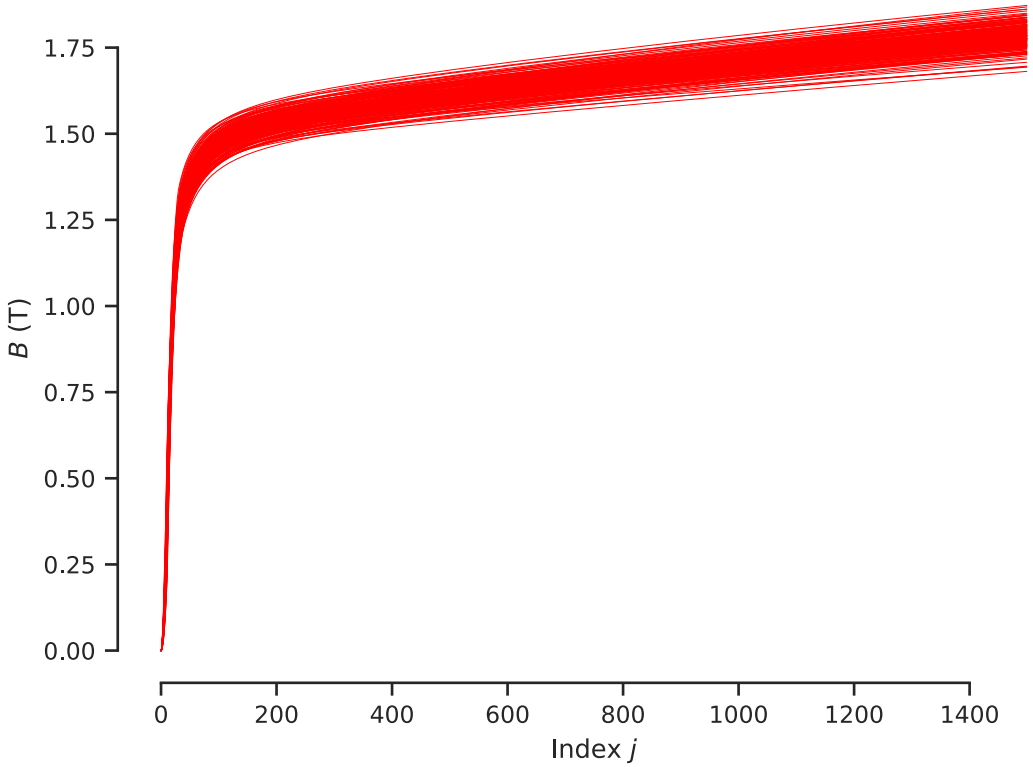
Here are five samples:

[ ]:
```
fig, ax = plt.subplots()
ax.plot(B_data[:5, :].T)
ax.set_xlabel(r"Index $j$")
ax.set_ylabel(r"$B$ (T)")
sns.despine(trim=True);
```



Here are all the samples:

[ ]:
```
fig, ax = plt.subplots()
ax.plot(B_data[:, :].T, 'r', lw=0.1)
ax.set_xlabel(r"Index $j$")
ax.set_ylabel(r"$B$ (T)")
sns.despine(trim=True);
```

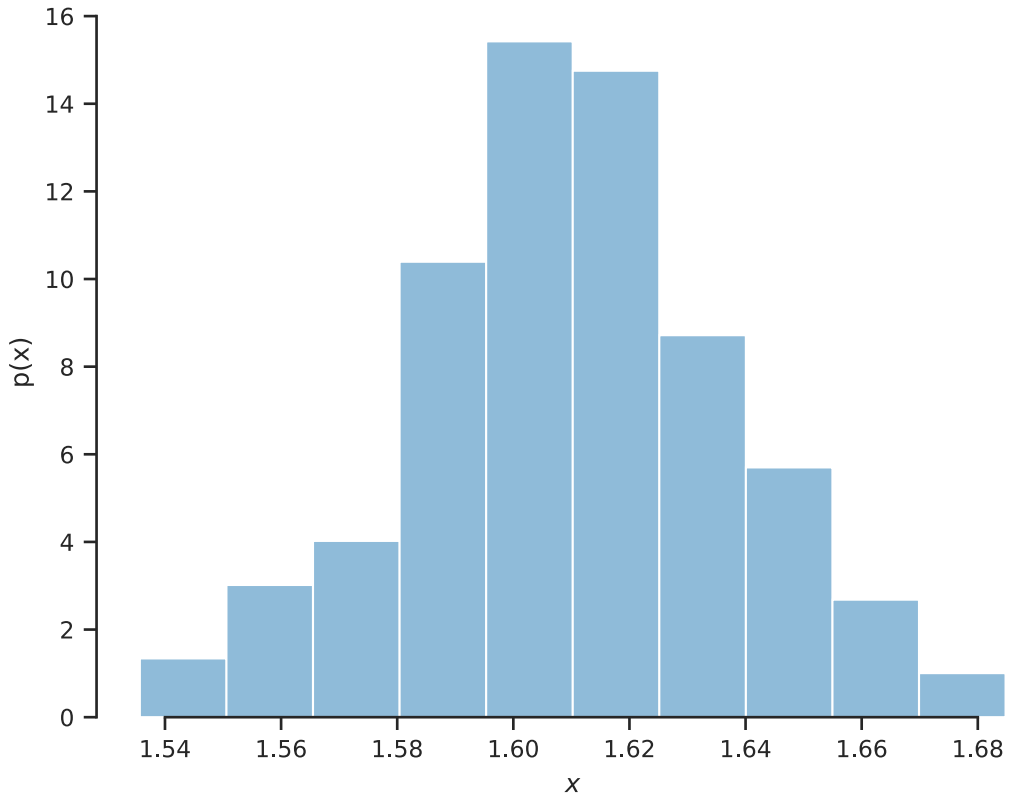


A. We are going to start by studying the data at only one index. Say index $j = 500$. Let's define a random variable

$$X = B(H_{500}),$$

for this reason. Extract and do a histogram of the data for X:

[ ]:
```
X_data = B_data[:, 500]
fig, ax = plt.subplots()
ax.hist(X_data, alpha=0.5, density=True)
ax.set_xlabel(r"$x$")
ax.set_ylabel(r"p(x)")
sns.despine(trim=True);
```



This looks like a Gaussian $N(\mu_{500}, \sigma^2_{500})$. Let's try to find a mean and variance for that Gaussian. A good choice for the mean is the empirical average of the data:

$$\mu_j = \frac{1}{N}\sum_{i=1}^{N} B_{ij}.$$

By the law of large numbers, this is a good approximation of the true mean as $N \to \infty$. Later we will learn that this is also the *maximum likelihood* estimate of the mean.

So, the mean is:

[ ]:
```
mu_500 = X_data.mean()
print(f"mu_500 = {mu_500:.2f}")
```

```
mu_500 = 1.61
```

Similarly, for the variance a good choice is the empirical variance defined by:

$$\sigma_j^2 = \frac{1}{N}\sum_{i=1}^{N}(B_{ij} - \mu_j)^2.$$

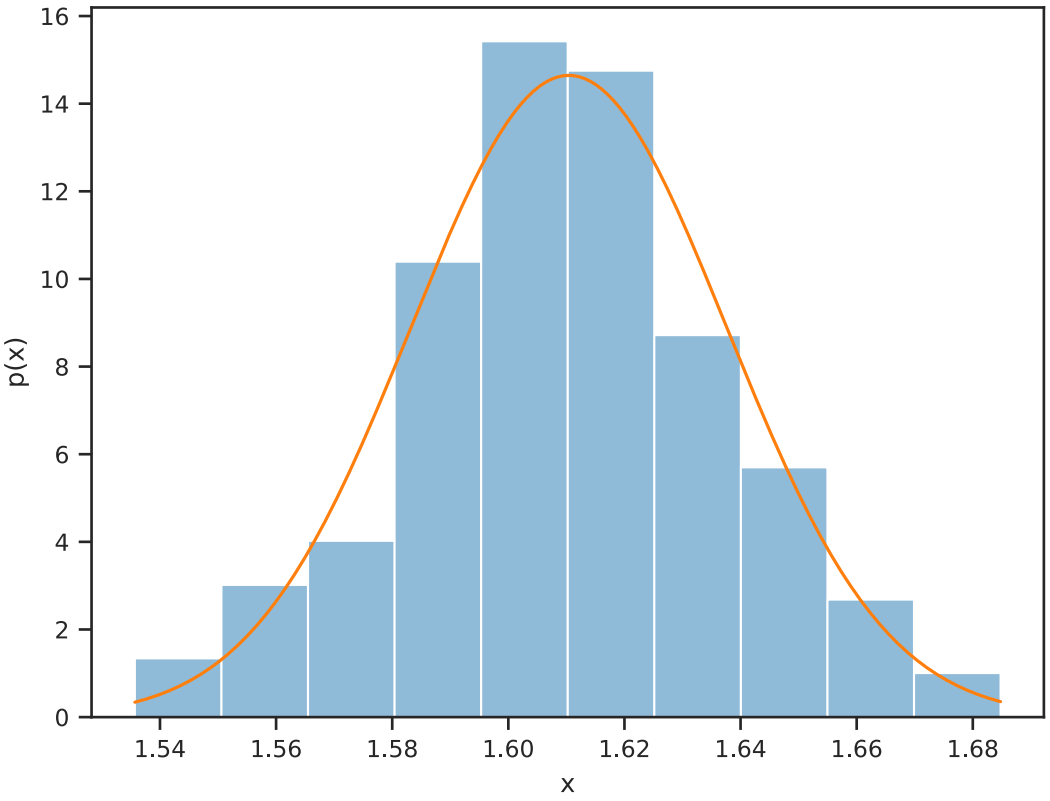
This also converges to the true variance as $N \to \infty$. Here it is:

[ ]:
```
sigma2_500 = np.var(X_data)
print(f"sigma_500 = {sigma2_500:.2e}")
```

```
sigma_500 = 7.42e-04
```

Repeat the plot of the histogram of X along with the PDF of the normal variable we have just identified using the functionality of `scipy.stats`.

[ ]:
```
fig, ax = plt.subplots()
ax.set_xlabel("x")
ax.set_ylabel("p(x)")

ax.hist(X_data, alpha=0.5, density=True)
pdf_sample_points = np.linspace(X_data.min(), X_data.max(), 1000)
X_norm = st.norm(loc=mu_500, scale=np.sqrt(sigma2_500))
ax.plot(pdf_sample_points, X_norm.pdf(pdf_sample_points));
```



B. Using your normal approximation to the PDF of X, find the probability that $X = B(H_{500})$ is greater than 1.66 T.

[ ]:
```
print(f"p(X > 1.66) = {1 - X_norm.cdf(1.66)}")
```

```
p(X > 1.66) = 0.034355748631309635
```

C. Let us now consider another random variable

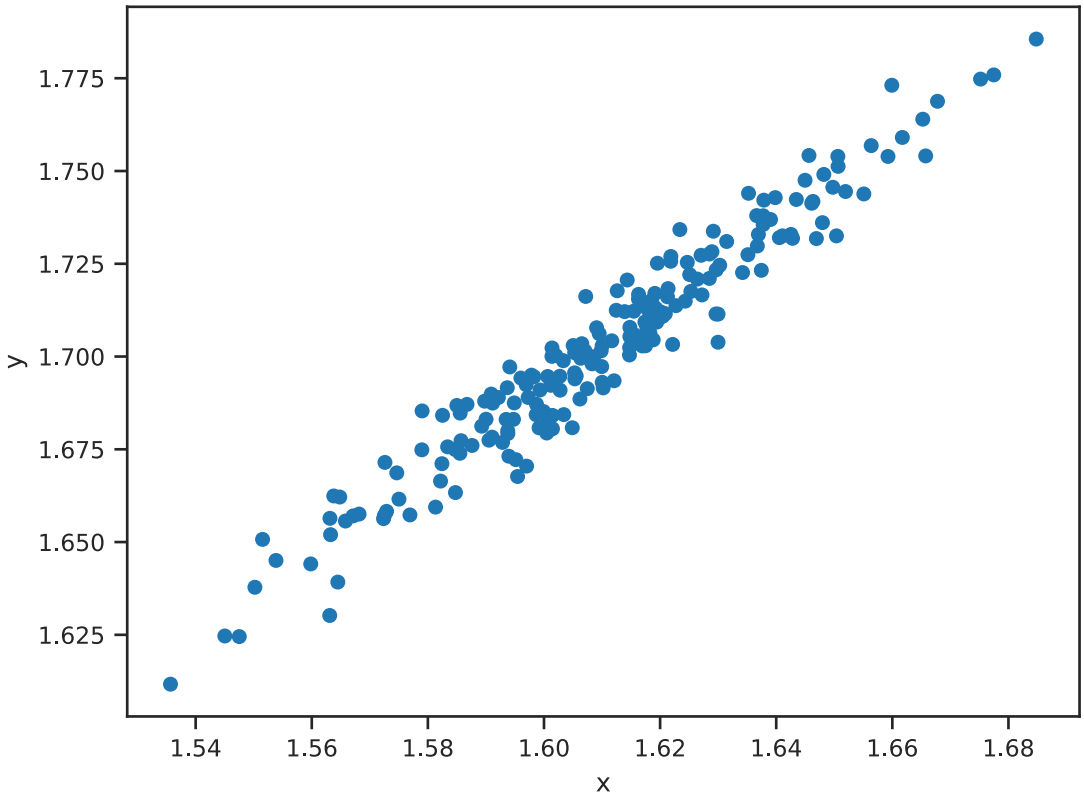
$$Y = B(H_{1000}).$$

Isolate the data for this as well:

[ ]:
```
Y_data = B_data[:, 1000]
```

Do the `scatter` plot of X and Y:

[ ]:
```
fig, ax = plt.subplots()
ax.set_xlabel("x")
ax.set_ylabel("y")

ax.scatter(X_data, Y_data);
```



D. From the scatter plot, it looks like the random vector

$$\mathbf{X} = (X, Y),$$

follows a multivariate normal distribution. What would be the mean and covariance of the distribution? First, organize the samples of X and Y in a matrix with the number of rows being the

number of samples and two columns (one corresponding to X and one to Y).

[ ]:
```
XY_data = np.hstack([X_data[:, None], Y_data[:, None]])
```

In case you are wondering, the code above takes two 1D numpy arrays of the same size and puts them in a two-column numpy array. The first column is the first array, the second column is the second array. The result is a 2D numpy array. We take sampling averages over the first axis of the array.

The mean vector is:

[ ]:
```
mu_XY = np.mean(XY_data, axis=0)
print(f"mu_XY = {mu_XY}")
```

```
mu_XY = [1.61041566 1.70263681]
```

The covariance matrix is trickier. We have already discussed how to find the diagonals of the covariance matrix (it is simply the variance). For the off-diagonal terms, this is the formula that is being used:

$$C_{jk} = \frac{1}{N}\sum_{i=1}^{N}(B_{ij} - \mu_j)(B_{ik} - \mu_k).$$

This formula converges as $N \to \infty$. Here is the implementation:

[ ]:
```
# Careful with np.cov because it requires you to transpose the matrix we␣
↪defined in class
C_XY = np.cov(XY_data.T)
print(f"C_XY =")
print(C_XY)
```

```
C_XY =
[[0.00074572 0.00082435]
 [0.00082435 0.00096729]]
```

Use the covariance matrix `C_XY` to find the correlation coefficient between X and Y.

[ ]:
```
corr_XY = C_XY[0,1] / np.sqrt(X_data.var() * Y_data.var())
print(corr_XY)
```

```
0.9754923538193253
```

Are the two variables X and Y positively or negatively correlated? **Answer:** They are positively correlated

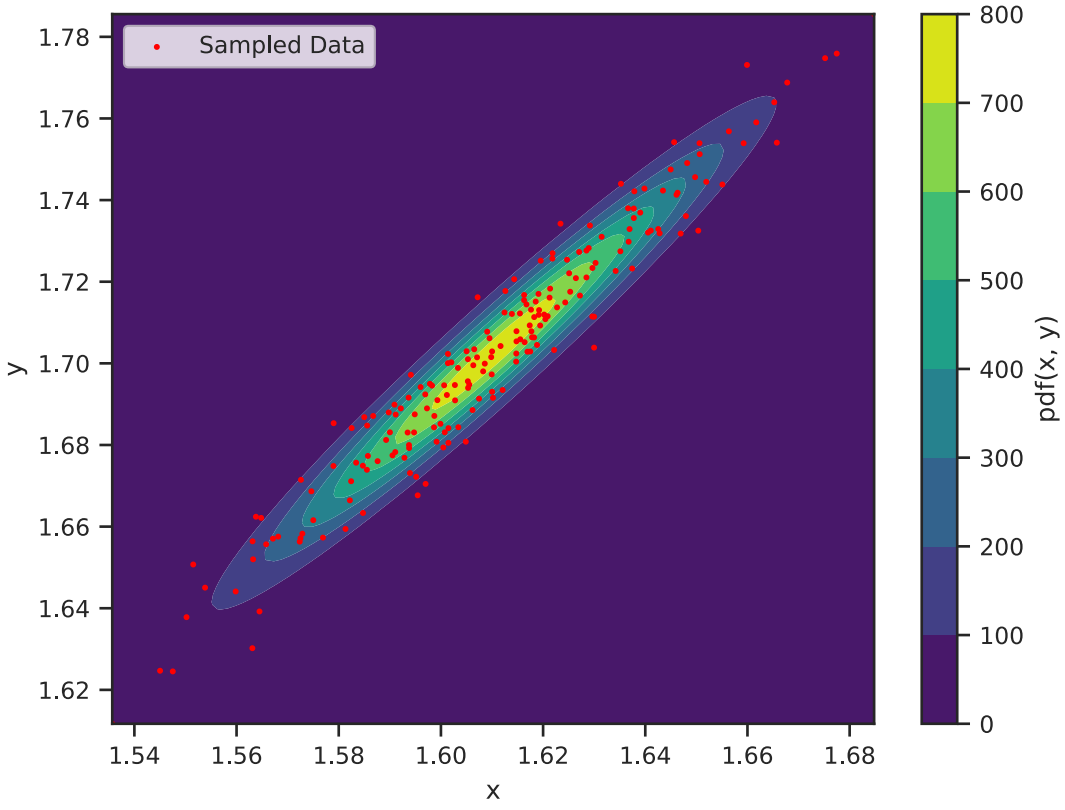
E. Use `np.linalg.eigh` to check that the matrix `C_XY` is indeed positive definite.

[ ]:
```
if np.all(np.linalg.eigh(C_XY)[0] > 0):
    print("All eigenvalues of C_XY are positive, so C_XY is positive definite")
```

```
All eigenvalues of C_XY are positive, so C_XY is positive definite
```

F. Use the functionality of `scipy.stats.multivariate_normal` to plot the joint probability function of the samples of X and Y in the same plot as the scatter plot of X and Y.

[ ]:
```
fig, ax = plt.subplots()
x, y = np.meshgrid(np.linspace(X_data.min(), X_data.max(), 100), np.
 ↪linspace(Y_data.min(), Y_data.max(), 100))

pos = np.dstack((x, y))

norm_XY = st.multivariate_normal(mean=[X_data.mean(), Y_data.mean()], cov=C_XY)

contour = ax.contourf(x, y, norm_XY.pdf(pos), cmap='viridis');

cbar = fig.colorbar(contour)
cbar.set_label('pdf(x, y)')

ax.scatter(X_data, Y_data, s=2, color='red', label='Sampled Data');
ax.legend()
ax.set_xlabel('x')
ax.set_ylabel('y');
```



G. Now, consider each $B-H$ curve a random vector. That is, the random vector $\mathbf{B}$ corresponds to the magnetic flux density values at a fixed number of H-values. It is:

$$\mathbf{B} = (B(H_1), \dots, B(H_{1500})).$$

It is like $\mathbf{X} = (X, Y)$ only now we have 1,500 dimensions instead of 2.

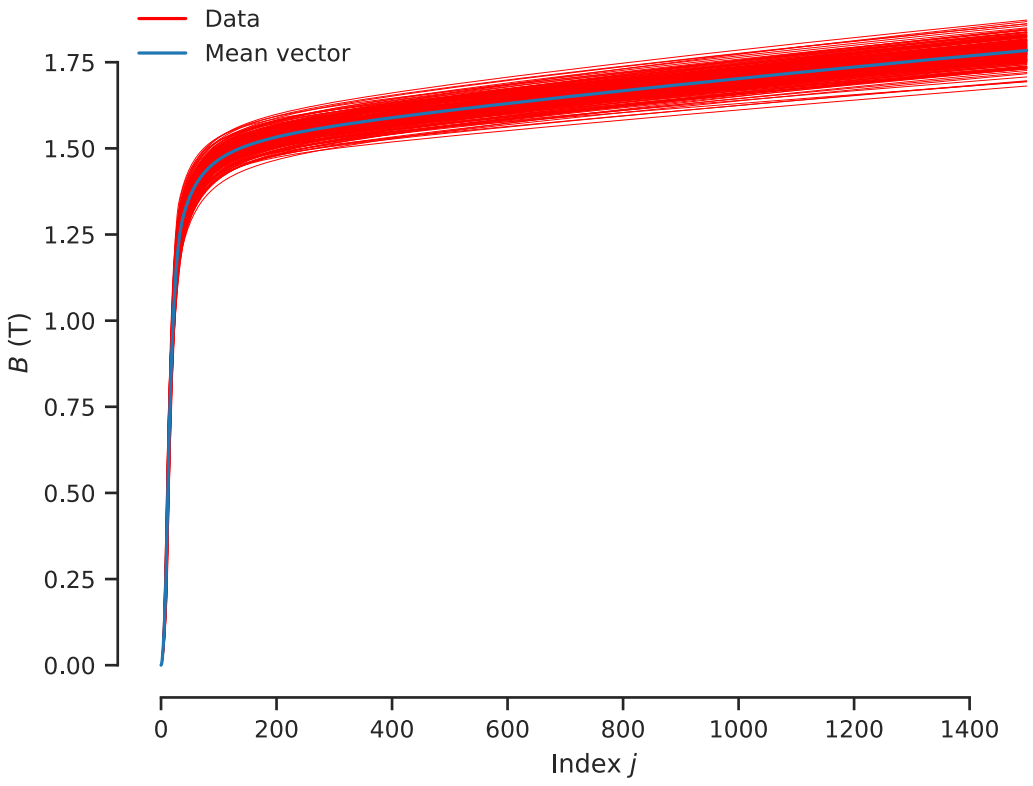
First, let's find the mean of this random vector:

[ ]:
```
B_mu = np.mean(B_data, axis=0)
B_mu
```

[ ]:
```
array([0.        , 0.00385192, 0.01517452, …, 1.78373703, 1.78389267,
       1.78404828])
```

Let's plot the mean on top of all the data we have:

[ ]:
```
fig, ax = plt.subplots()
ax.plot(B_data[:, :].T, 'r', lw=0.1)
plt.plot([],[], 'r', label='Data')
ax.plot(B_mu, label="Mean vector")
ax.set_xlabel(r"Index $j$")
ax.set_ylabel(r"$B$ (T)")
plt.legend(loc="best", frameon=False)
sns.despine(trim=True);
```



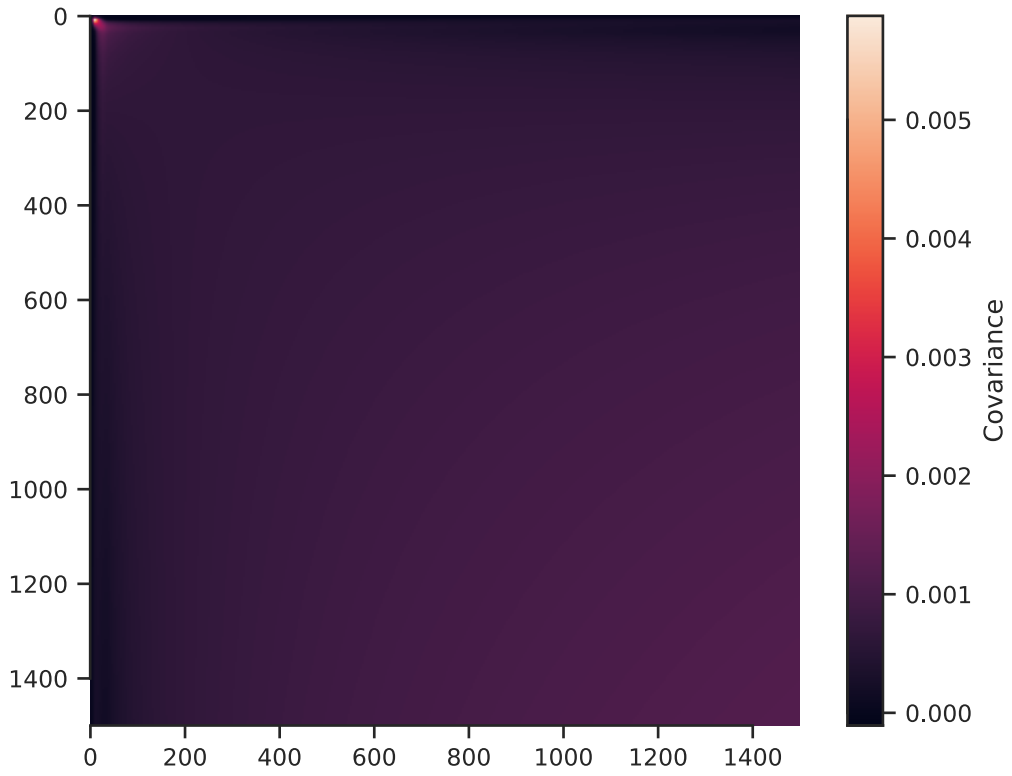
It looks good. Now, find the covariance matrix of **B**. This is going to be a 1500x1500 matrix.

[ ]:
```
B_cov = np.cov(B_data.T)
B_cov
```

[ ]:
```
array([[0.00000000e+00, 0.00000000e+00, 0.00000000e+00, …,
        0.00000000e+00, 0.00000000e+00, 0.00000000e+00],
       [0.00000000e+00, 1.16277948e-06, 4.41977479e-06, …,
        3.18233676e-06, 3.18391580e-06, 3.18549316e-06],
       [0.00000000e+00, 4.41977479e-06, 1.68041482e-05, …,
        1.22832828e-05, 1.22890907e-05, 1.22948922e-05],
       …,
       [0.00000000e+00, 3.18233676e-06, 1.22832828e-05, …,
        1.20268920e-03, 1.20293022e-03, 1.20317114e-03],
       [0.00000000e+00, 3.18391580e-06, 1.22890907e-05, …,
        1.20293022e-03, 1.20317134e-03, 1.20341237e-03],
       [0.00000000e+00, 3.18549316e-06, 1.22948922e-05, …,
        1.20317114e-03, 1.20341237e-03, 1.20365351e-03]])
```

Let's plot this matrix:

[ ]:
```
fig, ax = plt.subplots()
c = ax.imshow(B_cov, interpolation='nearest')
plt.colorbar(c, label="Covariance")
sns.despine(trim=True);
```



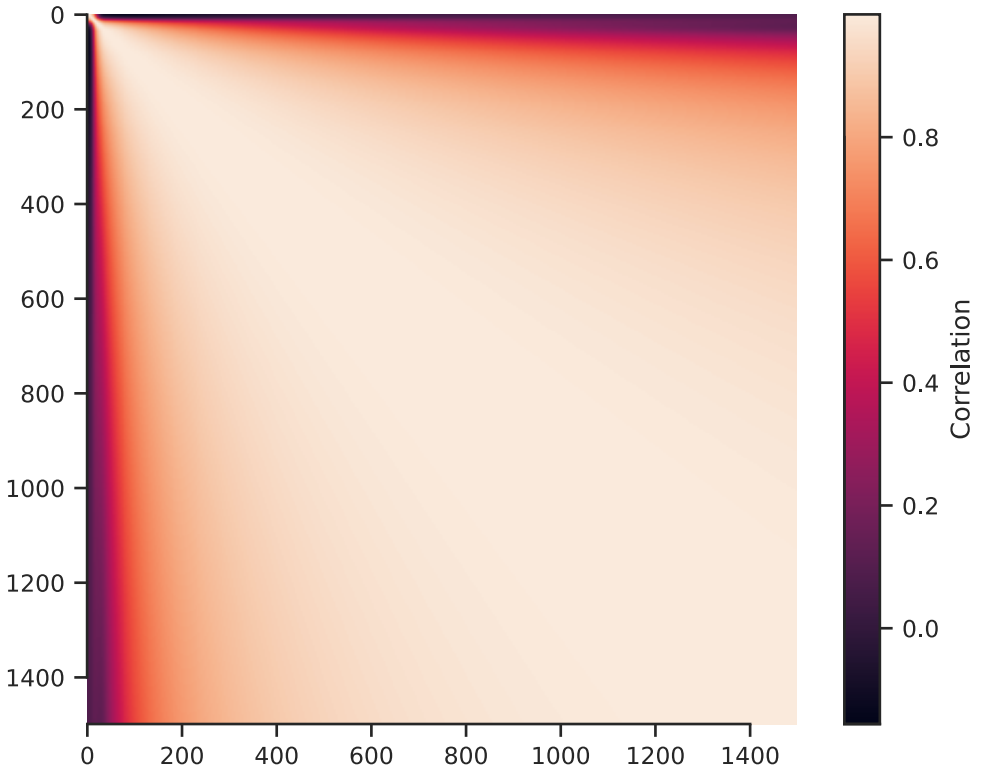
The numbers are very small. This is because the covariance depends on the units of the variables. We need to do the same thing we did with the correlation coefficient: divide by the standard deviations of the variables. Here is how you can get the correlation coefficients:

[ ]:
```
# Note that I have to remove the first point because it is always zero
# and it has zero variance.
B_corr = np.corrcoef(B_data[:,1:].T)
B_corr
```

[ ]:
```
array([[1.        , 0.99986924, 0.99941799, …, 0.08509827, 0.08512344,
        0.08514855],
       [0.99986924, 1.        , 0.99983894, …, 0.08640313, 0.08642667,
        0.08645015],
       [0.99941799, 0.99983894, 1.        , …, 0.08782484, 0.08784655,
        0.08786822],
       …,
       [0.08509827, 0.08640313, 0.08782484, …, 1.        , 0.99999998,
        0.9999999 ],
       [0.08512344, 0.08642667, 0.08784655, …, 0.99999998, 1.        ,
        0.99999998],
       [0.08514855, 0.08645015, 0.08786822, …, 0.9999999 , 0.99999998,
        1.        ]])
```

Here is the correlation visualized:

[ ]:
```
fig, ax = plt.subplots()
c = ax.imshow(B_corr, interpolation='nearest')
plt.colorbar(c, label="Correlation")
sns.despine(trim=True);
```



The values are quite a bit correlated. This makes sense because the curves are all very smooth and look very much alike.

Let's check if the covariance is indeed positive definite:

[ ]:
```
print("Eigenvalues of B_cov:")
print(np.linalg.eigh(B_cov)[0])
```

```
Eigenvalues of B_cov:
[-3.59054959e-16 -1.22898540e-16 -8.73546355e-17 …  4.66244763e-02
  1.16644070e-01  1.20726782e+00]
```

Notice that several eigenvalues are negative, but they are too small. Very close to zero. This happens often in practice when you are finding the covariance of large random vectors. It arises from the fact that we use floating-point arithmetic instead of real numbers. It is a numerical artifact. If you tried to use this covariance to make a multivariate average random vector using `scipy.stats` it would fail. Try this:

```
[ ]: B = st.multivariate_normal(mean=B_mu, cov=B_cov)
```

```
---------------------------------------------------------------------------
LinAlgError                               Traceback (most recent call last)
Cell In[179], line 1
----> 1 B = st.multivariate_normal(mean=B_mu, cov=B_cov)

File /nix/store/1wm7609fxdhzpznn6yx97sy81a0mxqdj-python3-3.10.12-env/lib/python3.
 ↪10/site-packages/scipy/stats/_multivariate.py:398, in multivariate_normal_gen
 ↪__call__(self, mean, cov, allow_singular, seed)
    393 def __call__(self, mean=None, cov=1, allow_singular=False, seed=None):
    394     """Create a frozen multivariate normal distribution.
    395
    396     See `multivariate_normal_frozen` for more information.
    397     """
--> 398     return multivariate_normal_frozen(mean, cov,
    399                                       allow_singular=allow_singular,
    400                                       seed=seed)

File /nix/store/1wm7609fxdhzpznn6yx97sy81a0mxqdj-python3-3.10.12-env/lib/python3.
 ↪10/site-packages/scipy/stats/_multivariate.py:839, in␣
 ↪multivariate_normal_frozen.__init__(self, mean, cov, allow_singular, seed,␣
 ↪maxpts, abseps, releps)
    796 """Create a frozen multivariate normal distribution.
    797
    798 Parameters
   (…)
    835
    836 """
    837 self._dist = multivariate_normal_gen(seed)
    838 self.dim, self.mean, self.cov_object = (
--> 839     self._dist._process_parameters(mean, cov, allow_singular))
    840 self.allow_singular = allow_singular or self.cov_object._allow_singular
    841 if not maxpts:

File /nix/store/1wm7609fxdhzpznn6yx97sy81a0mxqdj-python3-3.10.12-env/lib/python3.
 ↪10/site-packages/scipy/stats/_multivariate.py:422, in multivariate_normal_gen
 ↪_process_parameters(self, mean, cov, allow_singular)
    415 dim, mean, cov = self._process_parameters_psd(None, mean, cov)
    416 # After input validation, some methods then processed the arrays
    417 # with a `_PSD` object and used that to perform computation.
    418 # To avoid branching statements in each method depending on whether
    419 # `cov` is an array or `Covariance` object, we always process the
    420 # array with `_PSD`, and then use wrapper that satisfies the
    421 # `Covariance` interface, `CovViaPSD`.
--> 422 psd = _PSD(cov, allow_singular=allow_singular)
    423 cov_object = _covariance.CovViaPSD(psd)
    424 return dim, mean, cov_object
```

```
File /nix/store/1wm7609fxdhzpznn6yx97sy81a0mxqdj-python3-3.10.12-env/lib/python
 ↪10/site-packages/scipy/stats/_multivariate.py:177, in _PSD.__init__(self, M,
 ↪cond, rcond, lower, check_finite, allow_singular)
    174 if len(d) < len(s) and not allow_singular:
    175     msg = ("When `allow_singular is False`, the input matrix must be "
    176            "symmetric positive definite.")
--> 177     raise np.linalg.LinAlgError(msg)
    178 s_pinv = _pinv_1d(s, eps)
    179 U = np.multiply(u, np.sqrt(s_pinv))

LinAlgError: When `allow_singular is False`, the input matrix must be symmetric
 ↪positive definite.
```

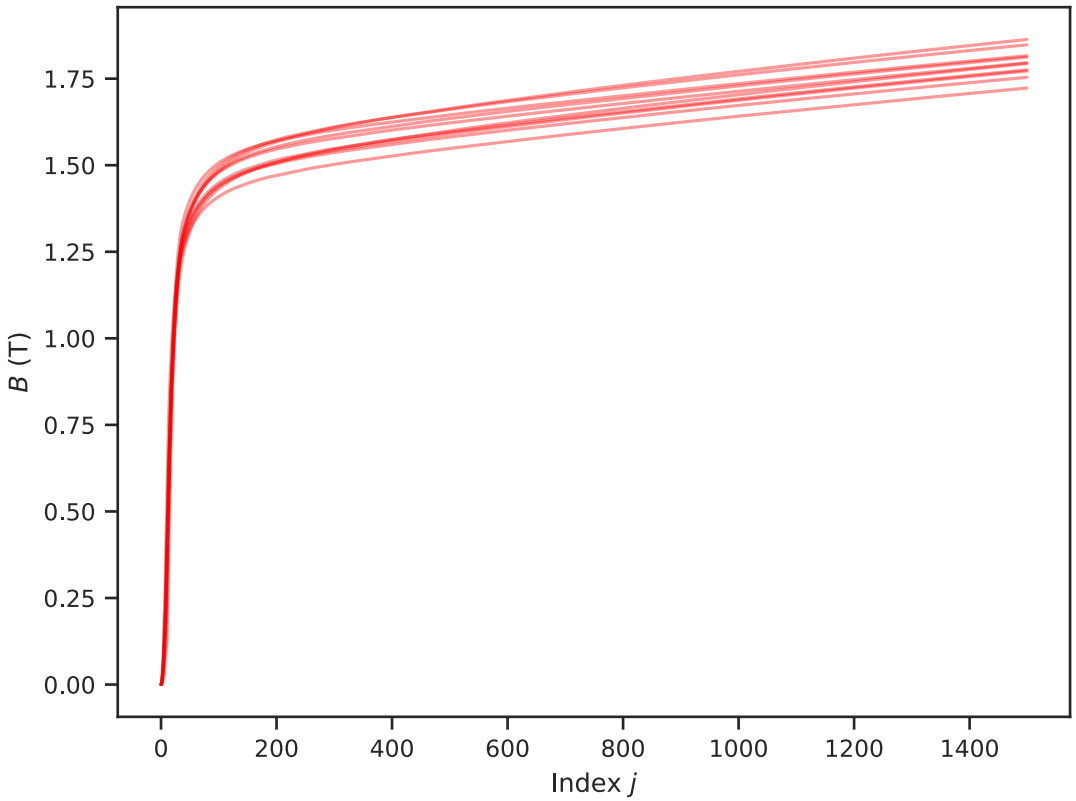
The way to overcome this problem is to add a small positive number to the diagonal. This needs to be very small so that the distribution stays mostly the same. It must be the smallest possible number that makes the covariance matrix behave well. This is known as the *jitter* or the *nugget*. Find the nugget playing with the code below. Every time you try, multiply the nugget by ten.

[ ]:
```
# Pick the nugget here
nugget = 1e-12 * 1000
# This is the modified covariance matrix
B_cov_w_nugget = B_cov + nugget * np.eye(B_cov.shape[0])
# Try building the distribution:
try:
    B = st.multivariate_normal(mean=B_mu, cov=B_cov_w_nugget)
    print('It worked! Move on.')
except:
    print('It did not work. Increase nugget by multiple of 10.')
```

```
It worked! Move on.
```

H. Now, you have created your first stochastic model of a complicated physical quantity. By sampling from your newly constructed random vector $\mathbf{B}$, you have essentially quantified your uncertainty about the $B - H$ curve as induced by the inability to control steel production perfectly. Take ten samples of this random vector and plot them.

[ ]:
```
fig, ax = plt.subplots()
ax.plot(B.rvs(10).T, 'r', alpha=0.4);
ax.set_xlabel(r"Index $j$")
ax.set_ylabel(r"$B$ (T)");
```



Congratulations! You have made your first stochastic model of a physical field quantity. You can now sample $B - H$ curves in a way that honors the manufacturing uncertainties. This is the first step in uncertainty quantification studies. The next step would be to propagate these samples through Maxwell's equations to characterize the effect on the performance of an electric machine. If you want to see how that looks, look at {cite}`sahu2020` and {cite}`beltran2020`.