EMAT30007 Applied Statistics

Lab 1: Random variables, probability distributions and sampling

Filippo Simini

This lab focusses on the material covered in Lecture 1: Random Variables, PDFs, CDFs and sampling.

1. Random Variables (RVs)

In Lecture1 we encountered Random Variables with various distributions: Binomial, Uniform, Exponential, Gamma.

We used these RVs to model data generated by some real world process. For example, we assumed that the outcome of a coin toss is well described by a Binomial (or Bernoulli) RV.

Let's see how to visualise and sample (i.e. generate numbers) from these distributions in Matlab.

Binomial distribution $X \sim Binomial(n, p)$

The Binomial distribution is a **discrete** distribution with two parameters: $p \in [0, 1]$ and $n \in \mathbb{N}$.

The **domain** (i.e. possible outcomes) is the set of all integers less or equal to n: $h = 0, 1, ..., n$.

The value h can be interpreted as the number of successes in n independent trials, if the probability of success in each trial is p. One example is h = number of heads in n tosses of a (possibly unfair) coin where p is the probability of a head. The *Bernoulli distribution* (e.g. the outcome of one coin toss) is a Binomial distribution with $n = 1$.

In Matlab, you can get information on a probability distribution and its parameters using `makedist` :

```
makedist('Binomial')
```

You can also use `makedist` to create a distribution with specified parameter values :

```
N = 10;
p = 0.5;

pd = makedist('Binomial', 'N', N, "p", p)
```

Theoretical PMF

The Binomial PMF is $P_H(h; n, p) = \binom{n}{h} p^h (1-p)^{n-h}$. In Matlab you can use `pdf(pd, x)` to evaluate the PDF or PMF of a distribution `pd` at the points in `x` .

```
outcomes = 0:1:10
binom_pmf = pdf(pd, outcomes)
clf;  % clear the figure
plot(outcomes, binom_pmf, 'o-b', 'LineWidth', 2)
hold on
binom_mean = mean(pd);  % Let's plot the distribution mean
xline(binom_mean, '--r', 'Linewidth', 3);
legend('theoretical PDF', 'mean', 'Location','northeast')
title('PDF - Binomial') % title for plot
xlabel('outcomes') % x-axis label
ylabel('PDF') % y-axis label
```

Generate a random sample

In Matlab, you can use `random(pd, rows, columns)` to generate a matrix of rows x columns random numbers from a specific distribution `pd` defined using `makedist`.

`binornd(N,p,rows,columns)` also generates random numbers from a binomial distribution specified by parameters N and p.

```
rng('default')  % fix the random seed
S = 10  % sample size
small_sample = random(pd,1,S)
rng('default')
binornd(N,p,1,S)
```

Here, `rng` sets Matlab's random number generator to a specific seed, so that a given identical sequence of (pseudo)random numbers is generated each time a function is called.

Empirical PMF

Given a list of numbers, the empirical PDF or PMF can be computed using `histogram`

```
binom_epmf = histogram(small_sample, 'normalization', 'pdf');  % correctly normalised
binom_epmf.Values
```

Let's draw a larger sample with 10000 observations (and specify the bin edges of the histogram). Note that the empirical PMF of the larger sample is very close to the theoretical PMF (circles).

```
large_sample = random(pd,1,10000);
histogram(large_sample, 'BinEdges', (0.5:1:N+0.5), 'normalization', 'pdf');
```

Theoretical CDF

In Matlab you can use `cdf(pd, x)` to evaluate the Cumulative Distribution Function (CDF) of a distribution `pd` at the points in `x` .

```
binom_cdf = cdf(pd, outcomes)
clf  % clear the plot
stairs(outcomes, binom_cdf, 'o-r', 'LineWidth', 2)
hold on
legend('theoretical CDF', 'Location','southeast')
title('CDF - Binomial') % title for plot
xlabel('outcomes') % x-axis label
ylabel('CDF') % y-axis label
```

Empirical CDF

Given a list of numbers, `data`, the empirical CDF can be computed using `ecdf(data)`

```
[binom_ecdf, x] = ecdf(small_sample);
stairs(x, binom_ecdf, 'o-')
legend('theoretical CDF','empirical CDF')
```

The empirical CDF of the larger sample with 10000 observations generated earlier is very close to the theoretical CDF.

```
[binom_ecdf, x] = ecdf(large_sample);
stairs(x, binom_ecdf, 's--k', 'LineWidth', 2)
legend('theoretical CDF','small sample', 'large sample')
```

Quantile function

The quantile function is the inverse of the CDF. In Matlab, you can use `icdf(pd, x)` to evaluate the inverse of distribution `pd` at points `x` .

```
icdf(pd, 0.1)
icdf(pd, 0.2)
% evaluate at more points within 0 and 1
pp = 0: 0.01: 1;

clf;
stairs(pp, icdf(pd, pp), '-')
hold on
legend('theoretical quantile function', 'Location','southeast')
title('Inverse CDF - Binomial') % title for plot
xlabel('outcomes') % x-axis label
ylabel('CDF') % y-axis label
```

Use Matlab's function `quantile(data, x)` to compute the empirical quantile of `small_sample` evaluated at points `p` .For the large sample, the empirical quantiles should be close (or identical) to the theoretical ones.

```
quantile(large_sample, 0.1)
quantile(large_sample, 0.2)

% evaluate at more points within 0 and 1
stairs(pp, quantile(large_sample, pp), '--')
legend('theoretical quantile function', 'large sample empirical quantile function', 'Location','southeast')
```

Q-Q plot

Q-Q plots are used to visually assess if empirical quantiles are close to theoretical ones, or if the quantiles of two distributions are similar. The quantiles of one distribution are plotted against the theoretical ones (or against the quantiles of the other distribution): if the points are aligned along the $y = x$ line, then the empirical quantiles is similar to the theoretical ones (or the two distributions are similar). Let's use a Q-Q plot to see if the quantiles of the small and large samples are similar to the quantiles of the theoretical Binomial distribution:

```
clf;

theor_q = icdf(pd, pp);
empir_q = quantile(data, pp);
empir_q_ls = quantile(large_sample, pp);

plot(theor_q, empir_q, 'o')
hold on
plot(theor_q, empir_q_ls, 's', 'MarkerSize', 10)
hold on
plot(theor_q, theor_q, '-','HandleVisibility','off')
legend('small sample', 'large sample', 'Location','southeast')
```

```
title('Q-Q plot - Binomial') % title for plot
xlabel('theoretical quantiles') % x-axis label
ylabel('empirical quantiles') % y-axis label
```

As expected, the quantiles of the large sample are closer to the theoretical ones than the quantiles of the small sample. The empirical quantiles of discrete distributions can be inaccurate around discontinuity points. The Q-Q plot works better for continuous distributions.

Common distributions

Repeat the previous analysis for the following distributions: Uniform, Exponential, Multinomial, Gamma, Normal.

For each distribution, anwser the following questions:

1. Is the RV discrete or continuous?
2. What is the domain of the RV's PDF or PMF ?
3. Describe the parameters of the distribution: how many are they? What values can they take?
4. Give one example of real world phenomenon or data that could be modelled using the distribution.

and do the following:

1. Plot the (theoretical) PDF (for contiuous RVs) or PMF (for discrete RVs) and the mean
2. Draw a sample of $S = 1000$ random outcomes.
3. Plot the empirical PDF or PMF of these samples on the same plot created in 1.
4. Plot the (theoretical) CDF
5. Plot the empirical CDF of the samples generated in 2 on the same plot created in 4.
6. Use a Q-Q plot to show the similarity between empirical and theoretical quantiles.

Uniform

1. Is the RV discrete or continuous?
2. What is the domain of the RV's PDF or PMF ?
3. Describe the parameters of the distribution: how many are they? What values can they take?
4. Give one example of real world phenomenon or data that could be modelled using the distribution.

```
% Theoretical PDF

% Mean

% Empirical PDF

% Sample

% Theoretical CDF

% Empirical CDF

% Q-Q plot
```

Exponential

1. Is the RV discrete or continuous?
2. What is the domain of the RV's PDF or PMF ?
3. Describe the parameters of the distribution: how many are they? What values can they take?
4. Give one example of real world phenomenon or data that could be modelled using the distribution.

```
% Theoretical PDF

% Mean

% Empirical PDF

% Sample

% Theoretical CDF

% Empirical CDF

% Q-Q plot
```

Multinomial (sum of two fair dices)

1. Is the RV discrete or continuous?
2. What is the domain of the RV's PDF or PMF ?

3. Describe the parameters of the distribution: how many are they? What values can they take?
4. Give one example of real world phenomenon or data that could be modelled using the distribution.

```
% sum of two dices
outcomes = (2 : 1: 12);
th_pdf = [1 2 3 4 5 6 5 4 3 2 1] / 36;

% Theoretical PDF

% Mean

% Empirical PDF

% Sample

% Theoretical CDF

% Empirical CDF

% Q-Q plot
```

Gamma (diameter of a tennis ball)

1. Is the RV discrete or continuous?
2. What is the domain of the RV's PDF or PMF ?
3. Describe the parameters of the distribution: how many are they? What values can they take?
4. Give one example of real world phenomenon or data that could be modelled using the distribution.

```
% Theoretical PDF

% Mean

% Empirical PDF

% Sample

% Theoretical CDF

% Empirical CDF

% Q-Q plot
```

Normal

1. Is the RV discrete or continuous?
2. What is the domain of the RV's PDF or PMF ?
3. Describe the parameters of the distribution: how many are they? What values can they take?
4. Give one example of real world phenomenon or data that could be modelled using the distribution.

```
% Theoretical PDF

% Mean

% Empirical PDF

% Sample

% Theoretical CDF

% Empirical CDF

% Q-Q plot
```

2. Estimate the probability of events

The probability to observe an outcome within a contiguous range of values, for example $[x_0, x_1]$, can be computed using the CDF function $F_X(x)$ as

$Pr(x_0 \leq X < x_1) = F_X(x_1) - F_X(x_0)$

Compute the probabilities of these events:

1. That $x \in [5, 6]$, for RV X with Exponential distribution $P_X(x) = e^{-x/\mu}/\mu$ and parameter $\mu = 4$.
2. That the outcome of a Standard Normal RV with $\mu = 0$ and $\sigma = 1$ is within $-\sigma$ and σ: $x \in [-\sigma, \sigma]$. Does your result depend on σ?
3. You arrive at a bus stop and you read that there is a bus every hour. If you arrived at a random time, what is the probability that you'll wait between 5 and 20 minutes before the next bus arrives? And between 30 and 45 minutes? And between 55 and 70 minutes?

```
% 1. Exponential

% 2. Normal

% 3. Uniform
```

3. Compute the empirical CDF from data

The CDF is defined in Eq.(1) of Lecture 1 as

$F_X(x) = Prob(X \leq x)$

The empirical CDF can be computed following the steps described in Lecture 1:

1. Sort data in ascending order, x.
2. Create an array, y, of increasing integers from 1 to the length of x.
3. Divide each element of y by the length of x, so that the max of y is 1.
4. If there are $(x_i,\ y_i)$ pairs with identical x-value, keep only the pair with the largest y-value. $y = F_X(x)$.

Use the above steps to compute the empirical CDF of the following dataset `data` and compare your result with the output of Matlab's `ecdf` function

```
data = [0.9838    1.1704    6.8989    1.7893    1.5168    0.1136    6.3209    1.6271    3.8187    3.5633];
```

```
% sort data in ascending order, x

% create array y of increasing integers of size length(x)

% divide y by length(x)

% Matlab's CDF
```

It is assumed that the ECDF values of all outcomes smaller than `min(data)` are zero and the ECDF values of all outcomes larger than `max(data)` are one.

4. Compute the empirical PDF from data

The empirical PDF of a data sample can be computed following the steps described in Lecture 1:

1. Divide the range of data, `[min(data) max(data)]`, into bins, with bin edges b.
2. Compute the mid-point of each bin, $x_i = (b_{i+1} + b_i)/2$.
3. Count the number of data elements that fall in each bin, $c_i = |\{d_j : b_i \leq d_j < b_{i+1}\}|$.
4. Normalise. Rescale each element of c: $y_i = c_i/(|d|(b_{i+1} - b_i))$, so that the PDF integral is 1. $y = P_X(x)$.

Use the above steps to compute the empirical PDF of the following dataset `data` uysing bin edges `bin_edges` and compare your result with the output of Matlab's `histogram` function

```
data = [0.9838    1.1704    6.8989    1.7893    1.5168    0.1136    6.3209    1.6271    3.8187    3.5633];
bin_edges = 0: 3: 12;
% bin_edges = [0 0.75 1.5 3 5 8 12];
```

```
% compute bins' mid points

% count the number of data points that fall in each bin

% normalise

% PDFs
```

Change the size of the bin edges and check that the PDF is correctly normalised. Is it normalised even if bins have varying sizes, like `bin_edges = [0 0.75 1.5 3 5 8 12]`?

5. Inverse Probability Integral Transform (IPIT) sampling

To sample from RV X with inverse CDF (i.e. the quantile function) F_X^{-1} :

1. Generate a number u from the uniform distribution $U(0,1)$.
2. Compute $x = F_X^{-1}(u)$.

Use the IPIT to draw 1000 independent observations from a Normal distribution with $\mu = 5$ and $\sigma = 3$.

Plot the empirical PDF (histogram) of the generated data and the theoretical PDF on the same graph.

```
% generate uniform random numbers
% compute the inverse CDF of the desired distribution

% PDFs
```

Use the IPIT to draw samples from a multinomial distribution representing the sum of two fair dices.

For discrete distributions like this one, you can use the following code to get the quantile value ($F_X^{-1}(u)$) of number `u`: `outcomes(min(find(th_cdf > u)))`, where `th_cdf` is the RV's CDF.

Plot the empirical PDF (histogram) of the generated data and the theoretical PDF on the same graph.

```
% sum of two dices
outcomes = (2 : 1: 12);
th_pdf = [1 2 3 4 5 6 5 4 3 2 1] / 36;
% th_cdf = ...

% generate uniform random numbers

% compute the inverse CDF of the desired distribution

% PDFs
```

6. Functions of Random Variables & sampling

Equations (5) and (6) in Lecture 1 can be used to determine the CDF and PDF of a strictly increasing or decreasing function of a RV, $Y = g(X)$ with $h = g^{-1}$:

$$F_Y(y) = \begin{cases} F_X(h(y)) & \text{if } h \text{ and } g \text{ increasing} \\ 1 - F_X(h(y)) & \text{if } h \text{ and } g \text{ decreasing} \end{cases}$$

$$P_Y(y) = P_X(h(y)) \cdot \left| \frac{dh(y)}{dy} \right|$$

Remember to transform the domain of the RV: $Y \in [g(x_{min}), g(x_{max})]$ if g increasing or $Y \in [g(x_{max}), g(x_{min})]$ if g decreasing.

Exercise

RV X has the continuous PDF $P_X(x) = (3/2)x^{1/2}$ defined for $x \in [0, 1]$.

1. Sample 1000 random numbers from P_X.
2. Plot the theoretical and the empiricals PDFs on the same graph.
3. Apply the following transformation to the sample data: $Y = e^X$ and calculate the empirical PDF of Y. Compute the PDF $P_Y(y)$ analytically and compare it against the empirical one.
4. Repeat the analysis of point 3. for the transformation $Z = e^{-X}$.

To sample using the IPIT we need to compute the inverse CDF of X...

```
% Sample


% PDFs
```

The theoretical PDF of $Y = e^X$ is $P_Y(y) = ...$

```
% Transformation Y = e^X

% remeber to transform the domain as well
```

The theoretical PDF of $Z = e^{-X}$ is $P_Z(z) = ...$

```
% Transformation Z = e^{-X}

% remeber to transform the domain as well
```

7. Donations

(Function of RVs and probability of events)

You collect donations for a good cause. The amount of each donation, X, is a RV following an exponential distribution with mean $\mu = 5$ pounds.

1. Assume that $N = 10$ people donate in total. Generate $S = 10000$ samples of the RV $Y = \sum_{i=1}^{N} X_i$, the total amout collected, and plot the empirical PDF of Y. What assumption(s) are you making to derive your answer?

2. What is the expected value (mean) of Y, the overall donations? Compute the theoretical and an empirical value of your sample and compare theory and simulation results.
3. Use sampling to estimate the probability that with $N = 10$ donors you will collect more than a target amount of $T = 60$ pounds. How does your answer change as a function of S?

```
% Distribution of one donation


% Samples of N donations


% PDF
```

Using the property of the Expected value of a sum of RVs, the theoretical mean is ...

The probability to collect more than T pounds is given by the CCDF of Y... Let's estimate this value from the samples.

```
% Probability to collect more than 60 pounds
```