ECEN321 Engineering Statistics

Laboratory Session 2

Probability and Probability Densities

For this Lab 2, you will need to submit an individual Lab Report. Please refer to the Lab marking sheets (and suggested Lab report format) for reference. Check with the Lab instructor the due date of the Lab report, and how / where to submit the report.

In this Lab, we will generate some random quantities and use simple estimators for their properties. Commands you might not have been seen before, but which you might need, are listed after each paragraph.

1 Normal Density

Generate a vector of 1000 normal (i.e. Gaussian) random variable having mean 2.5 and variance 16 (note that we want the variance to be 16, not the standard deviation). **randn ( )**

Find the sample mean and variance of the sample. **mean ( ) var ( ) std ( )**

Generate a histogram of the data using 30 bins. You might want to generate your own grid for the histogram using **linspace ( )** rather than leaving it up to **hist ( )** to decide its own. In this case you probably want three or four times σ on each side of the mean.

We would like to compare the histogram with the true pdf. There are two scaling factors we must apply here:

1. The first is that the sum of entries in the histogram is *N*, whereas the integral of the pdf is 1. So to convert the histogram counts to probabilities we must divide by *N*.

2. The second is that for the pdf to represent a probability, we need to consider some small region ∂x over which an integral is performed. To convert the histogram counts to probability *densities* we need to divide by ∂x.

Recall that the equation for the pdf of a normal random variable is

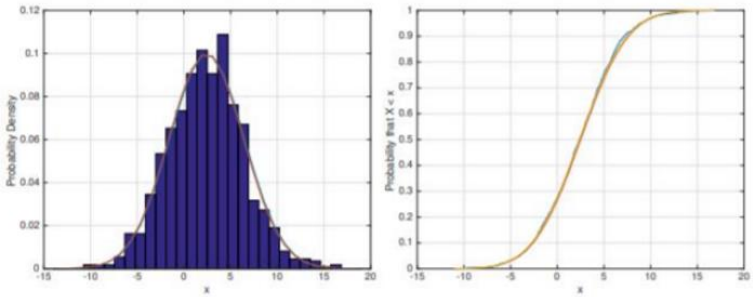
$$p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Overlay these plots on one figure: 1. The histogram, 2. The pdf of the actual variable, and 3. The pdf using $\hat{\mu}$ and $\hat{\sigma}^2$ estimated from the data.

Now, let's compare the empirical cdf of the sample with that of the population. We can generate the empirical cdf by plotting the data versus its rank/*N*. You might want to add an additional point which is slightly smaller than any of the data in the set, so that the probability that $\boldsymbol{X} < x$ is actually zero at the point and/or you might want to offset the rank by 1/(2*N*). **sort ( ) erf ( )**

You'll notice that the sample and theoretical cdfs are much less distinguishable that are the pdfs. You can look more closely at a particular point on the plots by zooming using the mouse.

The plots should look something like this:



2 Student *t*

Now generate a matrix of size *M* x 10,000 of the same ransom variables as before, with *M* = 5.

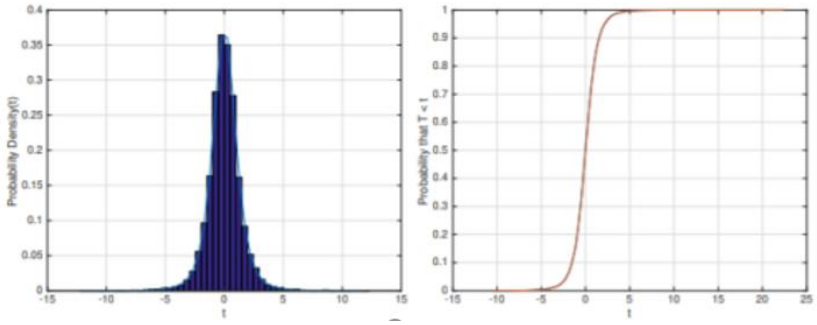
Generate a *t* value for each column by subtracting the true mean from the sample mean for each column, and dividing by the *sample* standard deviation and then multiplying by the square root of *M* (Note that by default, many Matlab functions operate on columns, unless we specify otherwise, *or unless there is only one row*. This last point can be a trap, since, we might sometimes end up with just one row without realising it, so it is a good idea to always specify the dimension you mean.)

Generate a histogram of the data. The *t* distribution for this number of degrees of freedom has heavy tails, so you'll probably want to cover a large range e.g., to ±10 or even more. (What happens to the histogram if the range is too small?)

Compare this with the theoretical pdf of $v = M - 1 = 4$ degrees of freedom. **tpdf ( )** or **gamma ( )** or maybe **gammaln ( )**.

$$p(x) = \frac{\Gamma((v+1)/2)}{\sqrt{v\pi}\Gamma(v/2)(1+x^2/v)^{(v+1)/2}}$$

Plot the sorted data versus the rank, and compare this with the theoretical cdf. **tcdf ( )**



3 Probability and the Birthday Paradox

Background:

How many people do you think it would take to survey, on average, to find two people who share the same birthday? Because there are 365 days in a year (ignoring leap year with 366 days), so you might think that you will need at least 183 people (i.e. ~ half of 365 days) before the odd of finding two people with the same birthday can be 0.5. This seems logical, because 183/365 is close to 0.5. But wait!

Have you ever noticed how sometimes what seems logical turns out to be proved false with a little math? Due to probability, sometimes an event is more likely to occur than we believe it to. In this case, if you survey a random group of just 23 people there is actually about a 50–50 chance that two of them will

have the same birthday. This is known as the birthday paradox. Don't believe it's true? You can test it and see mathematical probability in action!

Statistical Theory and Analytical Results:

As we have learnt in Lectures, instead of doing an actual survey (by conducting a simple random sampling), we can rely on simulation using Matlab to help us. But before we do that, let's analyse the mathematical theory behind.

The birthday paradox, also known as the birthday problem, states that in a random group of 23 people, there is about a 50 percent chance that two people have the same birthday. Is that really true? There are multiple reasons why this seems like a paradox. One is that when in a room with 22 other people, if a person compares his or her birthday with the birthdays of the other people it would make for only 22 comparisons — only 22 chances for people to share the same birthday.

But when all 23 birthdays are compared against each other, it makes for much more than 22 comparisons. How much more? Well, the first person has 22 comparisons to make, but the second person was already compared to the first person, so there are only 21 comparisons to make. The third person then has 20 comparisons, the fourth person has 19 and so on. If you add up all possible comparisons (22 + 21 + 20 + 19 + … +1) the sum is 253 comparisons, or combinations. Consequently, each group of 23 people involves 253 comparisons, or 253 chances for matching birthdays. This is known as the *counting* problem.

When comparing probabilities with birthdays, it can be easier to look at the probability that people do not share a birthday. A person's birthday is one out of 365 possibilities (excluding February 29 birthdays). The probability that a person does not have the same birthday as another person is 364 divided by 365 because there are 364 days that are not a person's birthday. This means that any two people have a 364/365, or 99.726027 percent, chance of not matching birthdays.

As mentioned before, in a group of 23 people, there are 253 comparisons, or combinations, that can be made. So, we're not looking at just one comparison, but at 253 comparisons. Every one of the 253 combinations has the same odds, 99.726027 percent, of not being a match. If you multiply 99.726027 percent by 99.726027 253 times, or calculate (364/365)253, you'll find there's a 49.952 percent chance that all 253 comparisons contain no matches. Consequently, the odds that there is a birthday match in those 253 comparisons is 1 – 49.952 percent = 50.048 percent, or just over half! The more trials you run, the closer the actual probability should approach 50 percent.

Here is a Matlab function that calculate the above probability for a group of people from 1 to n. Try it out. Note: for a Matlab function, you have to save it as the same name as the function, in this case we save it as birthday.m. Try n = 100.

birthday.m

```
function [A]=birthday(n)
  A=ones(n,1);
  p=1;
  for i=1:n
    A(i)=1-p;
    p=p*(365-i)/365;
  end

```

Question a: With the understanding from the explanation of the above Statistical theory, describe what the Matlab function is doing.

Question b: Plot the probability vs n (we will call this one the analytical result). Is it true that 23 people will give a probability of about 0.5?

By Simulation:

Now we are ready to use Matlab to simulate the sampling process to answer this question: If there is a group of n people in a room, what is the probability that two or more of them having the same birthday?

Hint: You can use randi ( ) to generate randomly a set of 365 days for n people and determine if there are two or more days (i.e. numbers) that are the same. If yes, increment the count by 1. Repeat the iteration for 1000 times, then find the fraction of count/1000, and this will give you the probability.

Write a function that determines the answer to this question by Matlab simulation. The program you write can take n as the input and prints out the probability that two or more of n people will have the same birthday for n=2,3,4…. 100. For example, when "25" is entered, your Matlab simulation will produce:

"The number of people: 25

The probability that two or more of 25 people

will have the same birthday equals 0.580000"

Plot the probability vs n (we will call this one the simulation results). Compare it with the analytical results – do they agree with each other?

321: Additional Materials beyond Lab 2

Now that you have completed Lab 2 and the report, let's investigate one more density distribution.
Note: This is optional and will not be marked.

Chi-squared Distribution

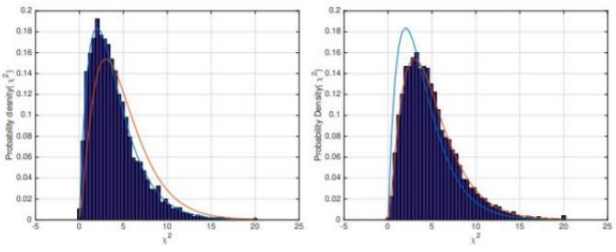
Beside the Normal and Student t density distributions that you have seen in Lab 2, chi-squared is another commonly used density distribution. The chi-squared distribution (also known as the chi-square or $\chi 2$-distribution) with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables. The chi-squared distribution is a special case of the gamma distribution and is one of the most widely used probability distributions in inferential statistics, notably in hypothesis testing and in construction of confidence intervals. Try explore more about the chi-squared distribution by following the following steps from Part 2 of Lab 2.

Generate a chi squared random variable by multiplying the sample variance for each column (of the matrix from Section 2) by $\frac{(5-1)}{\sigma^2}$. Generate a histogram for this also.

Compare this with the theoretical pdf for $\nu = M - 1 = 4$ degrees of freedom `chi2pdf()`:

$$p(x) = \frac{(1/2)^{\nu/2}}{\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}$$

Also generate a chi squared random variable from the sum of differences from the *true mean*, divided by σ^2. Compare this with the theoretical pdf for $\nu = M = 5$ degrees of freedom:



If you have finished chi-squared distribution…

If you have finished investigating chi-squared density distribution and still have time, go and try out the materials on Page 28 to 30 of Lab 1_2 Statistics Using Matlab.

If you still have time…

Go ahead and look at Page 30 to 38 “Simple Linear Regression in MATLAB” of Lab 1_2 Statistics Using Matlab.