9 Rec 9: Limit Theorems/Confidence Intervals

Directions: Your instructor will spend the the first 40 minutes of the recitation period working some review problems and going over one or more Matlab experiments in the following. During the last 10 minutes of recitation, your proctor will give you a "Lab Form" that your recitation team completes, signs, and turns in. See the last page for an indication of what you will be asked to do on the Lab Form.

Due to time limitations, only a part of the following can be covered during the recitation period. However, you might want in the future to try some of the uncovered experiments on your own. They could give skills useful on some future homework problems and could lend insight into your understanding of the course from an experimental point of view.

This Week's Topics.

- Sums of Independent RV's and Convolution
- Central Limit Theorem (CLT) for a Continuous Sampling Distribution
- CLT for a Discrete Sampling Distribution
- Confidence Intervals
- Variance of Sum of Dependent RV's

9.1 Exp 1: Sums of Independent RV's and Convolution

Let X_1 and X_2 be independent discrete RV's, and let

$$X = X_1 + X_2$$

be the sum. Then

$$p_X(x) = p_{X_1}(x) * p_{X_2}(x), \tag{1}$$

where we are taking convolution in the usual EE 3015 sense. I will eventually prove this result in the class lecture notes. In this experiment, you will verify formula (1) using Matlab. In Matlab, convolution is performed using the function "`conv`".

Example 1. In this example, you let RV's X_1 and X_2 be the numbers which come up in flipping a fair die two times. We know from earlier in the course that the RV

$$X = X_1 + X_2,$$

which is the total of the numbers on the two die flips, has a PMF distributed over the set

$$\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

according to certain probabilities. First, run the following Matlab script, which generates a histogram approximation of this PMF based upon 10000 simulated observations of X:

```
x1=ceil(6*rand(1,10000));
x2=ceil(6*rand(1,10000));
x=x1+x2;
subplot(2,1,1)
bar(2:12,hist(x,2:12)/10000)
```

Now run the following Matlab script, which gives the exact plot of the PMF of X using convolution:

```
PMF1=[1/6 1/6 1/6 1/6 1/6 1/6];
PMF2=[1/6 1/6 1/6 1/6 1/6 1/6];
PMFX=conv(PMF1,PMF2);
subplot(2,1,2)
bar(2:12,PMFX)
```

Compare the two plots you see on your computer screen. Are the two plots about the same?

9.2 Exp 2: CLT for a Continuous Sampling Distribution

Suppose you have a probability distribution governed by a density $f(x)$. Let n be a large positive integer. Independently select n random samples $X_1, X_2, \ldots, X_n$ according to this distribution. Let μ and σ^2 be the mean and variance of the density $f(x)$, given by

$$\mu = \int_{-\infty}^{\infty} x f(x) dx$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$

Assume that the following assumptions hold:

$$-\infty < \mu < \infty, \quad 0 < \sigma^2 < \infty.$$

The central limit theorem (CLT) says that the random variable

$$Z = \frac{(X_1 + X_2 + \ldots + X_n) - n\mu}{\sigma\sqrt{n}} \tag{2}$$

is approximately distributed according to the standard Gaussian distribution, meaning that

$$P(a \le Z \le b) \approx \int_a^b \frac{1}{\sqrt{2\pi}} \exp(-z^2/2) dx.$$

The approximation becomes better as n gets bigger. (In the limit as $n \to \infty$, the approximation becomes exact.)

This is an amazing result if you consider the fact that this approximate Gaussian behavior will occur regardless of what the density $f(x)$ is.

In this experiment, we use Matlab to convince you of the truth of the CLT when the distribution we sample from is a continuous distribution.

Example 2. We take samples from an exponential distribution with mean $\mu = 1$ and variance $\sigma^2 = 1$. Run the following Matlab script, which estimates the PDF of

$$\frac{(X_1 + X_2) - 2\mu}{\sigma\sqrt{2}},$$

where X_1, X_2 are two random samples from our exponential distribution.

```
n=2;
x=sum(-log(rand(n,100000)));
z=(x-mean(x))/std(x);
N=1000;
A=min(z);B=max(z);
Delta=(B-A)/N;
t=A-Delta/2+[1:N]*Delta;
PDFestimate=hist(z,t)/(Delta*100000);
subplot(3,1,1)
bar(t,PDFestimate)
```

Your estimated PDF plot looks kind of skewed, doesn't it? (Not very Gaussian bell-shaped at all!) Now run the following script, which estimates the PDF of

$$\frac{(X_1 + X_2 + \cdots + X_8) - 8\mu}{\sigma\sqrt{8}},$$

where we have now taken 8 independent samples $X_1, X_2, \cdots, X_8$ from our exponential distribution.

```
n=8;
x=sum(-log(rand(n,100000)));
z=(x-mean(x))/std(x);
N=1000;
A=min(z);B=max(z);
Delta=(B-A)/N;
t=A-Delta/2+[1:N]*Delta;
PDFestimate=hist(z,t)/(Delta*100000);
subplot(3,1,2)
bar(t,PDFestimate)
```

Does your estimated PDF plot look more like a Gaussian bell-shaped curve? Now run the following script, which estimates the PDF of

$$\frac{(X_1 + X_2 + \cdots + X_{32}) - 32\mu}{\sigma\sqrt{32}},$$

where we have now taken 32 independent samples $X_1, X_2, \cdots, X_{32}$ from our exponential distribution.

```
n=32;
x=sum(-log(rand(n,100000)));
z=(x-mean(x))/std(x);
N=1000;
A=min(z);B=max(z);
Delta=(B-A)/N;
t=A-Delta/2+[1:N]*Delta;
PDFestimate=hist(z,t)/(Delta*100000);
subplot(3,1,3)
bar(t,PDFestimate)
```

Of the three estimated PDF plots you plotted in this experiment, this last one should look most like a Gaussian bell-shaped curve.

Example 3. In this example, we choose our independent random samples from a Uniform$(-1, 1)$ distribution. According to Appendix A, this distribution has mean $\mu = 0$ and variance $\sigma^2 = 1/3$. Run the following script, which estimates the CDF of

$$Z = \frac{(X_1 + X_2 + \cdots + X_{32}) - 32\mu}{\sigma\sqrt{32}},$$

where we have taken 32 independent samples $X_1, X_2, \cdots, X_{32}$ from our uniform distribution.

```
n=32;
number_of_experiments=100000;
x=2*rand(n,number_of_experiments)-1;
var_x=1/3;
Sn=sum(x);
% Find the PDF of y=Sn/sqrt(n*var_x)
y=Sn/sqrt(n*var_x);
Bins=number_of_experiments/1000;
y_min=min(y);
y_max=max(y);
Delta=(y_max-y_min)/Bins;
t=y_min+Delta/2+[0:Bins-1]*Delta;
P=cumsum(hist(y,t)/number_of_experiments);
u=-4:0.01:4;
P_N01=cdf('norm',u,0,1);
plot(t,P,'b--',u,P_N01,'r-')
axis([-4 4 -0.1 1.1])
title('CDF of CLT Z variable(dashed), CDF of standard Gaussian(solid)')
```

You will see two plots on your computer screen on the same set of axes. One of them (the dashed plot) is the estimated CDF of the central limit theorem Z variable, using 32 samples. The other one (the solid line plot) is the actual CDF of the Gaussian$(0, 1)$ distribution. Do the two plots seems pretty close together? Now see if your terminal is powerful enough for you to re-run the preceding script in which you change the first line to $n = 100$. Since you are now using 100 samples, the two CDF curves might look even closer now.

9.3 Exp 3: CLT for a Discrete Sampling Distribution

In Experiment 2, we sampled from a continuous distribution. In this experiment, you will sample from a discrete distribution and see that the CLT is still true. Unlike Experiment 2, you will use Matlab to find the precise distribution of the normalized sum Z in (2) (instead of estimating this distribution with simulated samples). This precise distribution is found via Matlab function "`conv`" as was done in Experiment 1 in a simpler case.

Example 4. In this example, you illustrate the CLT with the following discrete density being the one from which independent samples are summed up:

$$f(x) = (1/2)\delta(x) + (1/2)\delta(x-1)$$

Run the MATLAB program which follows in order to plot the density of the normalized sum Z given in (1) when $n = 1600$. (We chose this value of n because it gave a nice spacing of .05 between the values of the normalized sum in (2).)

```
n=1600;
%generate range of values of sum X1+X2+...+Xn
k=0:n;
%generate prob dist of sum via convolution
p=[.5 .5];
q=p;
for i=1:n-1
q=conv(q,p);
end
Q=q;   %Q is the prob dist of the sum
mean_of_dist = .5;
variance_of_dist = .25;
mean_of_sum = n*mean_of_dist;
standev_of_sum = sqrt(n*variance_of_dist);
%generate range of values of normalized sum
t=(k-mean_of_sum)/standev_of_sum;
density=standev_of_sum*Q;  %approx. values of density of normalized sum
plot(t,density)
axis([-3 3 0 .5])
```

Run tests on the resulting plot to see if it closely approximates the standard Gaussian density function $\exp(-x^2/2)/\sqrt{2\pi}$. (Does the curve have the right peak value? Does it have the right value at $x = \pm 1$?) Try to modify the code so that you get the approximate Gaussian CLT plot on the same set of axes as the actual standard Gaussian density curve.

Example 5. In this example, the CLT normalized sum Z (2) is based on $n = 1000$ X_i samples from the discrete probability distribution in which the values $1, 2, 3$ are taken on with probabilities $1/3, 1/3, 1/3$. The following program finds the CDF of the RV Z and then computes the exact value of $P(Z \leq 1)$:

```
clear;
n=1000;
p=[1/3 1/3 1/3];
q=p;
for i=1:n-1
q=conv(q,p);
end
CDF=cumsum(q); %these are the values of CDF of Z
mu = 2;
sigma = sqrt(2/3); %check that this number is right
z=((1000:3000)-n*mu)/(sqrt(n)*sigma); %these are the values of Z
Probability=CDF(max(find(z<=1)))
```

By the CLT, the distribution of Z should be approximately standard Gaussian. On page 123 of your textbook, look up the probability $P[Z \leq 1]$ for a standard Gaussian Z. Is the figure given by last line of above program correct to two decimal places? Change the last line of the program in order to obtain $P[Z \leq 1.5]$ and then compare to the cumulative prob you get from page 123.

9.4 Exp 4: Confidence Intervals

Suppose you take n independent samples $X_1, X_2, \cdots, X_n$ from a Gaussian distribution with unknown mean μ and known standard deviation σ; these X_i's form a so-called "random sample of size n". The sample mean $\bar{X}$ based on this random sample is defined by

$$\bar{X} \triangleq \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

We want to take an interval centered at $\bar{X}$ which will be highly likely to contain μ. For example, we can take this interval to be

$$[\bar{X} - 1.645\sigma/\sqrt{n}, \bar{X} + 1.645\sigma/\sqrt{n}], \tag{3}$$

which is called a 90% confidence interval for μ because μ is inside this interval with probability 0.90, that is,

$$P[\bar{X} - 1.645\sigma/\sqrt{n} < \mu < \bar{X} + 1.645\sigma/\sqrt{n}] = 0.90.$$

This means that if we determine a large number of confidence intervals by taking many different random samples of size n, we can expect that about 90% of these confidence intervals will contain μ. The interval

$$[\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n}] \tag{4}$$

is the 95% confidence interval for μ, using this same Gaussian distribution—about 95% of a large number of conficence intervals should contain μ.

Example 6. In this example, you use Matlab to verify that (4) indeed is the 95% confidence interval for the mean μ when you sample from a Gaussian distribution. In the following

Matlab script, you can enter in on the first 3 lines whatever mean μ and standard deviation σ you want for your Gaussian sampling distribution, as well as the number of samples n that you want to take. The script then computes what percentage of 50000 confidence intervals contain μ.

```
mu =    ; %enter in the mean that you want
sigma =     ; %enter in the standard deviation that you want
n =        ; %enter in the sample size that you want
x=sigma*randn(n,50000)+mu;
sample_means=mean(x); %gives 50000 sample means
C=1.96;
p=mean(mu<sample_means+C*sigma/sqrt(n) & mu>sample_means-C*sigma/sqrt(n));
percentage=round(100*p) %gives percentage of conf intervals containing mu
```

Run the preceding script with $\mu = 0$, $\sigma = 1$, and $n = 10$ several times. Most of the time, does it appear that you are getting 95% of the confidence intervals to contain μ? Now try $n = 15$, $\mu = 1$, $\sigma = 2$. Do you reach the same conclusion?

Example 7. In this example, you provide a Matlab verification that (3) is a 90% confidence interval for the mean μ when you sample from a Gaussian distribution. The Matlab script for verifying this is now

```
mu =     ; %enter in the mean that you want
sigma =       ; %enter in the standard deviation that you want
n =        ; %enter in the sample size that you want
x=sigma*randn(n,50000)+mu;
sample_means=mean(x); %gives 50000 sample means
C=1.645;
p=mean(mu<sample_means+C*sigma/sqrt(n) & mu>sample_means-C*sigma/sqrt(n));
percentage=round(100*p) %gives percentage of conf intervals containing mu
```

Run the preceding script with $\mu = 0$, $\sigma = 1$, and $n = 10$ several times. Most of the time, does it appear that you are getting 90% of the confidence intervals to contain μ? Now try $n = 15$, $\mu = 1$, $\sigma = 2$. Do you reach the same conclusion?

Example 8. Suppose you form sample means based on samples of size $n = 3$ from a uniform distribution with mean μ and variance σ^2. In this example, you will verify that

$$[\bar{X} - 0.95\sigma, \bar{X} + 0.95\sigma]$$

is an approximate 90% confidence interval. The following Matlab script simulates 50000 sample means for samples of size 3 from a Uniform(a, b) distribution. It then computes what percentage of the 50000 confidence intervals contain μ:

```
a  ; %enter in the value of a
b  ; %enter in the value of b
mu=(a+b)/2;
sigma=(b-a)/sqrt(12);
```

```
x=(b-a)*rand(3,50000)+a;
sample_means=mean(x); %gives 50000 sample means
p=mean(mu < sample_means + .95*sigma & mu > sample_means - .95*sigma);
percentage=round(100*p) %gives percentage of conf intervals containing mu
```

Run the above script a few times with $a = 0, b = 1$. Do most of the percentages seem to be 90%? Now run the above script a few times with $a = 1, b = 5$. Again, do most of the percentages seem to be 90%? If you like, run the script with some other choice of a, b chosen by you.

Example 9. Suppose you form sample means based on samples of size $n = 5$ from an exponential distribution with mean $\mu = 1$. It is claimed that

$$[\bar{X} - 0.602, \bar{X} + 0.602]$$

is an approximate 85% confidence interval in this situation. Write a Matlab program which will compute 50000 of these confidence intervals, and will check to see what percentage of them contain $\mu = 1$. See whether you get about 85% of them to work out right. (Hint: Recall that "`-log(rand(1,n))`" simulates n samples from this exponential distribution. Using this fact, you can modify the Matlab script in Example 8 to obtain a Matlab script that will work for this Example.

Exercise. Here is something for you to think about. Consider again the 90% confidence interval (3) for the mean of a Gaussian distribution. What happens to the width

$$\frac{2(1.645)\sigma}{\sqrt{n}}$$

of the confidence interval when you double the number of samples n? If n samples gives width w, how many samples would you need in order to squeeze the width of the confidence interval down to $w/2$? (Note that the trade-off between the width of the confidence interval and the number of samples required to achieve this width is important because as the width of the confidence interval gets smaller, the confidence interval estimate of μ gets better.)

9.5 Exp 5: Variance of Sum of Dependent RV's

The variance of a sum of independent RV's is the sum of the separate variances. However, we learned in class a few lectures ago that this property may not hold if the random variables you are adding up are statistically dependent. In this experiment, you use Matlab to estimate the variance of the sum of possibly dependent RV's. You then attempt to verify the estimate by an exact computation of the variance of the sum.

Example 10. Let Z_1, Z_2, Z_3, Z_4 be independent Gaussian(0,1) RV's and let X_1, X_2, X_3 be the RV's

$$\begin{aligned} X_1 &= Z_1 + Z_2 \\ X_2 &= Z_2 + Z_3 \\ X_3 &= Z_3 + Z_4 \end{aligned}$$

Run the following Matlab script, which estimates the variance of $X_1 + X_2 + X_3$ and the sum of the variances of the X_i's:

```
z1=randn(1,50000);
z2=randn(1,50000);
z3=randn(1,50000);
z4=randn(1,50000);
x1=z1+z2; x2=z2+z3; x3=z3+z4;
%estimate the  variance of X1 + X2 + X3 as follows
var(x1+x2+x3)
%now estimate the sum of the separate variances as follows
var(x1) + var(x2) + var(x3)
```

Look at the estimate for

$$Var(X_1 + X_2 + X_3) \tag{5}$$

and the estimate for

$$Var(X_1) + Var(X_2) + Var(X_3) \tag{6}$$

which Matlab printed out on your computer screen. On the basis of these estimates, do you believe that

$$Var(X_1 + X_2 + X_3) \neq Var(X_1) + Var(X_2) + Var(X_3) \tag{7}$$

is true? To finish this example, take pencil and paper and see if you can use EE 3025 theory to compute the exact values of (5) and (6). Are they the same? (Hint: We have

$$X_1 + X_2 + X_3 = Z_1 + 2Z_2 + 2Z_3 + Z_4,$$

and the terms on the right side are independent.)

Example 11. Let Z_1, Z_2, Z_3, Z_4 be independent Gaussian(0,1) RV's and let X_1, X_2, X_3, X_4 be the RV's

$$\begin{aligned} X_1 &= Z_1 \\ X_2 &= Z_2 + X_1/2 \\ X_3 &= Z_3 + X_2/2 \\ X_4 &= Z_4 + X_3/2 \end{aligned}$$

Construct and run a Matlab script to estimate

$$Var(X_1 + X_2 + X_3 + X_4). \tag{8}$$

Then try to compute the variance (8) by hand.

Final Remarks. With the filtering operations used in this experiment, we have seen instances illustrating how *independent* RV's Z_i are converted at the filter output into *dependent* RV's X_i. But, even though the X_i's exhibit dependence, it still might be true that statements like the following hold:

$$\lim_{n\to\infty} P\left[\frac{\sum_{i=1}^n (X_i - E[X_i])}{\sqrt{Var(\sum_{i=1}^n X_i)}} \leq z\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp(-t^2/2)dt \tag{9}$$

$$P\left[\lim_{n\to\infty}\frac{X_1+X_2+\cdots+X_n}{n}=C\right]=1,\ \text{ some constant } C. \tag{10}$$

$$P\left[\lim_{n\to\infty}\frac{X_1^2+X_2^2+\cdots+X_n^2}{n}=D\right]=1,\ \text{ some constant } D. \tag{11}$$

Statement (9) is what the CLT becomes in the dependent case. Statements (10)-(11) are *laws of large numbers.*

EE 3025 S2007 Recitation 9 Lab Form

Name and Student Number of Team Member 1:

Name and Student Number of Team Member 2:

Name and Student Number of Team Member 3:

****

Study Experiment 5 carefully. In the examples in this experiment, you filter *independent* RV's Z_i to obtain *dependent* RV's X_i. I will have you do similar filtering and then examine the behavior of averages like

$$\frac{X_1^2 + X_2^2 + \cdots + X_{n-1}^2 + X_n^2}{n}$$

for large n. You will see that such averages can converge to some fixed quantity as n becomes large even though the terms are dependent.