MATH 232: Engineering Mathematics
Reading Guide for LAS, Section 5.5: Normal Distributions and CLT

Goals:
1. To learn relationships between distributions, expected values and variances (standard deviations) when a new random variable such as

\[ Y = \sum_{i=1}^{n} X_i \quad \text{or} \quad \overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \]

is created out of existing random variables \( X_1, \ldots, X_n \) that have common mean \( \mu \) and common variance \( \sigma^2 \).

2. To learn what is meant by i.i.d. random variables, and how to recognize settings in which it is reasonable to assume we have such variables.

Read: Section 5.5 of LAS

Terms to know:
- normal random variable (one having \( \text{Norm}(\mu, \sigma) \) as its distribution), standard normal distribution, 68-95-99.7% rule, i.i.d. random variables (or i.i.d. random sample of size \( n \)), sample mean \( \overline{X}_n \) of \( n \) random variables \( X_1, \ldots, X_n \) (itself a random variable), Central Limit Theorem

Questions you should be able to answer:

1. Suppose we have \( n \) different random variables \( X_1, \ldots, X_n \), and from these we create a random variable \( Y = \sum_i X_i \). Without making any additional assumptions about the \( X_i \), do we know anything about \( E(Y) \) or \( \text{Var}(Y) \)? What if we know each of the \( X_i \) has a normal distribution? What if we assume the \( X_i \) are independent of each other? What if we make all of these assumptions?

2. If \( X_1, \ldots, X_n \) are i.i.d. random variables with common mean 5 and common variance 12, what distribution would we attribute to \( \overline{X}_n \) when \( n \) is large?

R examples in the vein of those from class:
To get the student survey dataset:

```r
> ss = read.table('http://www.calvin.edu/~scofield/data/tab/ssurv.txt', header=T, sep='	')
```

Record 94 (row 94 in the data frame `ss`) is missing haircut data. To create a new vector containing all existing haircut data:

```r
> haircut = ss$haircut[c(1:93,95:280)]
```
Treating these 279 haircut fees as our population, we may view the distribution via:

```r
> library(lattice)
> histogram(haircut, n=20)
```

To sample 4 students at random (i.e., an SRS) from this population and calculate the mean amount paid by those 4 students for a haircut:

```r
> mean(sample(haircut, 4))
```

If you repeat the previous command multiple times, the mean should change (because the sampled students will change). To repeat it a hundred thousand times, storing the results in a vector:

```r
> x = replicate(100000, mean(sample(haircut, 4)))
```

The distribution of sample means which may be calculated from samples of size 4 from our population is called the sampling distribution of the sample mean with \( n = 4 \). I do not know an easy way to ask R to plot this sampling distribution. But we can get a good feel for its appearance by looking at a histogram of the (many) values in \( x \):

```r
> x11()  # creates a new empty figure without erasing old figure
> histogram(x, n=20)
```

Suppose we take an SRS of size \( n \) from our population of haircut costs, letting \( X_1 \) be the first sampled value, \( X_2 \) the second, and so on. The central limit theorem applies to i.i.d. random samples and, strictly speaking, unless we are sampling with replacement, the variables \( X_1, \ldots, X_n \) are neither independent nor identically distributed. But, so long as \( n \) is kept small relative to the size of the population (which here is 279), the i.i.d. assumption is not badly violated. Thus, by the corollary to the Central Limit Theorem (Corollary 5.5.6), we should not be surprised that the sampling distribution for samples of size \( n = 30 \) is roughly normal, looking like \( \text{Norm}(13.87, 2.82) \). (Here \( \mu, \sigma \), the mean and s.d. in the population, have values 13.87 and 15.43 respectively. So 2.82 \( \approx \) 15.43/ \( \sqrt{30} \).) The following commands produce a histogram simulating this sampling distribution, and then overlays, for the purpose of comparison, the normal distribution \( \text{Norm}(13.87, 2.82) \):

```r
> x = replicate(100000, mean(sample(haircut, 30)))
> hist(x, breaks=40, freq=F)
> y = seq(5, 25, .1)
> lines(y, dnorm(y, 13.87, 2.82))