

# Math 1180:Lab10

Kyle Gaffney

March 19th, 2014

## 1 Normal Distribution

The Normal Distribution is a surprisingly accurate approximation of a general sum of random variables. The exact distribution for any general sum of random variables is often impossible to compute, however under a relatively wide and loose range of assumptions the normal distribution provides an accurate approximation to this unknown exact distribution.

To see this in action let us consider an example problem:

Suppose that a certain plant has 10 different genes affecting its height with an equally likely probability of getting the short and tall allele at each of these 10 loci. We will assume that plants with 2 copies of the short allele gain no height, plants with 1 copy of each allele gain 1 cm, and plants with 2 tall alleles gain 2.5 cm from that locus. Suppose that total height gain is the sum from each locus. The question then is what is the distribution for the heights of these plants.

We could calculate the exact distribution of the heights of the plants by using the Binomial Distribution. The problem is that with 10 different loci the enumeration of all the different cases is quite large. Instead let us treat the height gain from each locus as an independent process and find the expected value and variance of this individual process. Because the probabilities of the short and tall alleles are identical we have that:

$$Pr(H_i = 0) = 0.25$$

$$Pr(H_i = 1) = 0.5$$

$$Pr(H_i = 2.5) = 0.25$$

Then:

$$E(H_i) = 0 * 0.25 + 1 * 0.5 + 2.5 * 0.25 = 0 + 0.5 + 0.625 = 1.125$$

and furthermore:

$$Var(H_i) = 0^2(0.25) + 1^2(0.5) + 2.5^2(0.25) - 1.125^2 = 0.797$$

Now this is the expected height gain and the variance of the height gain from a single locus, and there are 10 independent identical loci with these same statistics. Thus:

$$E(H) = \sum_{i=1}^{10} E(H_i) = 10 * 1.125 = 11.25$$

$$Var(H) = \sum_{i=1}^{10} Var(H_i) = 10 * 0.797 = 7.97$$

The claim is that the Normal Distribution with mean 11.25 and standard deviation=  $\sqrt{7.97}$  is an accurate approximation of the distribution of height of the plants.

### 1.1 Simulation of the Plant Problem

Suppose that we simulate this process. For each locus we need to simulate the 2 alleles, then depending on the results we need to add 0, 1, or 2.5 cm to the height of the plant. Then we need to do this 10 times for each plant.

The alleles are either short or tall with equal (p=0.5) probability. We need to simulate 2 of them, 10 times. We can do this easily with the rbinom command in R which generates random numbers from the binomial distribution. Here we will consider the number to represent the number of tall alleles located at each locus.

```
A=rbinom(10,size=2,prob=1/2)
A
```

Now we need to count up how many 0's, 1's, and 2's we have so that we can convert to height. We have done something similar before:

```
sum(A==0)
```

Then the height of the plant will be:

```
H=0*sum(A==0)+1*sum(A==1)+2.5*sum(A==2)
H
```

Now we have successfully calculated the height of a random single plant. To see if the Normal Distribution is a good approximation though we should simulate a bunch of plants:

```
H=0
for (i in 1:1000){
A=rbinom(10,2,1/2)
```

```
H[i]=0*sum(A==0)+1*sum(A==1)+2.5*sum(A==2)}
head(H)
summary(H)
hist(H)
```

First what kind of distribution does this look like? Let us superimpose the Normal Distribution we calculated earlier with a scaled histogram.

```
hist(H,freq=F,breaks=20)
x=seq(0,25,.5)
lines(x,dnorm(x,mean=11.25,sd=sqrt(7.97)))
```

Why did this happen? Do we still get a nice looking approximation with the normal distribution if we do fewer trials? What about if the plant has fewer loci? What about if the probabilities are not equal for each allele type?

You can explore these questions in R by changing the probabilities in the binomial distribution, changing how many numbers we generate, and by changing how many trials we run in the for loop.

## 2 Assignment for the Week

1. Randomly generate 100 numbers from the exponential distribution with rate=1, now plot a histogram of these values with bins of size 0.2.

```
hist(n,freq=F,breaks=seq(0,max(n)+1,.2))
```

where n is the list of randomly generated numbers.

2. Repeat number 1 but now pick 100 pairs of numbers and average the pairs. Once again plot a histogram of the results with bin size 0.2

Hint:

```
n=0
for (i in 1:100){
n[i]=sum(rexp(2,1))/2}
```

3. Repeat number 2 but now pick 100 sets of 10 numbers and average the sets. Plot the resulting histogram as before.
4. Finally pick 100 sets of 50 random numbers and average the 50 number sets. Plot the resulting list in a histogram like before.
5. Describe what is happening to the histograms. What distribution do they resemble, try to explain why this is happening?