Lecture 7: Modeling continous data or why we like

Thus far we have played with simple models and discrete—binomially distributed data. Now we will switch to modeling continuous data using the normal distribution and dive into linear regression.

Why go Gaussian?

Imagine 100 of us go and hang out on the 50 yd line in Neyland Stadium. We all begin flipping coins (or globes), and each time it comes up heads we take a step forward; each time we get a tails, we take a step back. Each of us do this for 20 flips and then stop.

Can we predict the proportion of us hanging out at the end on the 50 yd line? what about the south 40 yd line?

We can simulate this without having to sneak on the field. For each person, we generate a list of steps and then add them up. Because everyone has different gaits, we will use a uniform distribution to generate step sizes.

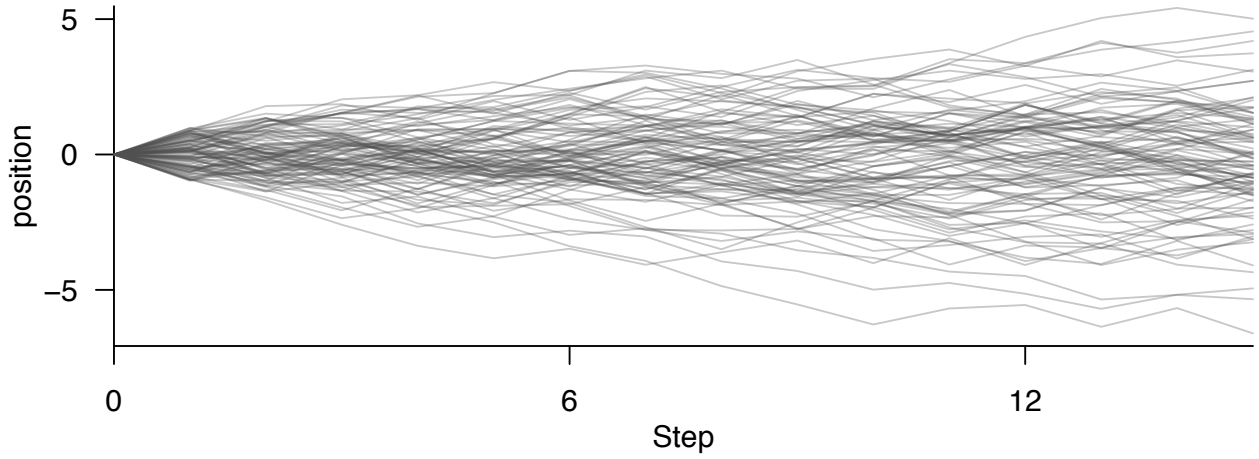
```
pos <- replicate(15, runif(100,-1, 1)) # simulate positions
cumPos <- t(apply(pos,1,cumsum)) # calculate cumulative position at each step
cumPos <- cbind(rep(0,100), cumPos) # add initial step
```

If we plot this out we see that even though we are simulating random walks from a uniform distribution, the familiar Gaussian shape emerges very quickly from the randomness.

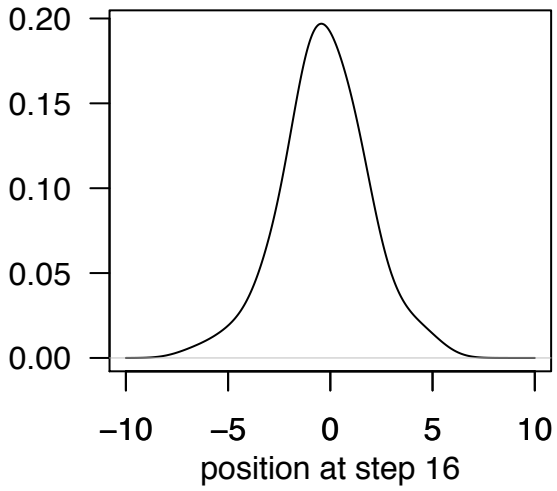
```
par(mar=c(3,3,0.1,0.5))
plot(1:100, cumPos[,16], xlim=c(0,15), type="n", las=1, axes=FALSE, xaxs="i")

axis(1, at=seq(0,24, by=6))
axis(2, at=seq(-10,10, by=5), las=1)
mtext(text = "Step", side=1, line = 2)
mtext(text = "position", side=2, line = 2.1)

for(i in 1:nrow(cumPos)) {
  lines(0:15, cumPos[i,], col="#50505050")
}
```




```
par(mar=c(3,3,0.1,0.5))
plot(density(cumPos[,16],adj=1.5, from=-10, to=10), main="", las=1, xlab="",ylab="")
axis(1, at=seq(-10,10, length=5), las=1)
mtext(text = "position at step 16", side=1, line = 2)
```



Any process that adds together random values from the same distribution (e.g., uniform) converges to a normal given a large enough sample.

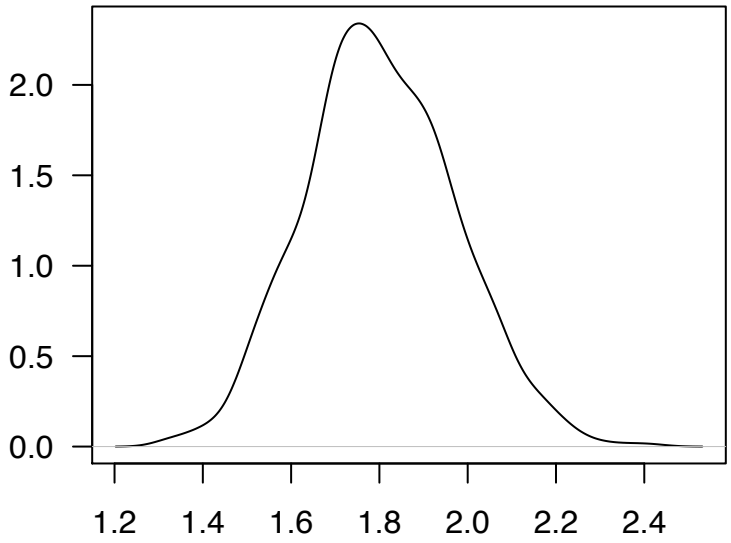
- Each sample can be thought of as a deviation from an average value.
- When those deviations are added together, those fluctuations cancel each other out.
- Eventually the most likely way to realize the sums will be one in which every fluctuation is canceled and sum to 0, relative to the mean.

The same thing happens when small effects are multiplied. For example, suppose the growth rate of an organism is affected by 10 loci, each with small interacting (i.e., multipicative affects). We can sample an individual by:

```
prod(1 + runif(10, 0,0.1))
```

where each of 12 loci has an effect from 1 (no multiplicative effect) to 1.1 (10% increase). If we sample 1,000 indidviduals,

```r
growth <- replicate(1000, prod(1+runif(12,0,0.1)))
plot(density(growth), main="",las=1)
```



we approximate a bell curve.

- Large multiplicative effects are normal on a log scale.

Given the phenomena described above, there are 2 good reasons for using a normal distribution for likelihoods and/or priors:

1. The normal distribution describes widespread patterns such as measurement errors, growth variation, etc.
    - Because the fluctuations of many different processes add together to resemble a normal distribution, we cannot easily identify the underlying process without additional information
    - this doesn't make the normal less useful for modeling.
2. The normal is generally the distribution with maximum entropy (to be elaborated on later).
    - Most natural expression of our ignorance. If all we can do is say there is finite variance, the Gaussian is the shape that realizes this ignorance in the largest number of ways without any additional assumptions.