

## Summary of Lab Week 6

Sampling distributions

Let's work through how to construct a sampling distribution. We'll use the distribution of sample means as an example, starting with a case where the population is Normally distributed. Recall that the idea is to pretend we know the population exactly and to pretend we are drawing a large (or infinite) set of samples from that population.

The first step is to draw a sample from the idealized population distribution. This can be done in R using random number generation. R generates random numbers from a Standard Normal distribution using the `rnorm()` function. Choose a sample size  $n$  and generate one sample.

```
> X = rnorm(n)
> X
> mean(X)
```

Since all we care about is the mean of the sample (not the raw scores themselves), we can be a little more efficient and combine the above commands:

```
> mean(rnorm(n))
```

Now let's get a large number of sample means. We use a loop to do this.

```
> for(i in 1:10000) M[i] = mean(rnorm(n))
```

This command tells R to execute `M[i] = mean(rnorm(n))` 10000 times, once for each value of  $i$ . Each time, a random Normal sample is drawn and its mean is stored as the  $i^{\text{th}}$  component of the vector `M`.

`M` is now a large vector of sample means. Each entry of `M` is the mean of a different sample.

Taken together, these entries give a good approximation to the theoretical distribution of sample means. (The reason it's an approximation is that we didn't get an infinite number of samples, but we got a large enough number that the approximation is very good.)

```
> hist(M, breaks = 30)
```

Observe how the sampling distribution is Normal, just like the population. Now check what the mean and standard deviation of this distribution are. Recall that for the population,  $\mu = 0$  and  $\sigma = 1$ . Recall also that the Central Limit Theorem predicts the distribution of sample means will have a mean of  $\mu$  and a standard deviation of  $\sigma/\sqrt{n}$ . How well did those predictions work out?

Now get a new distribution of sample means using a new value of  $n$  (this size of each sample).

Do this a few times, using values of  $n$  ranging from near 1 to much larger values. Notice how the shape of the sampling distribution is always Normal, even when the sample size is very small. This is because the original population is Normal; with other populations, the sampling distribution doesn't become Normal until  $n$  is large enough. Notice also that the mean always matches  $\mu$  (0), and the standard deviation gets smaller as the sample size gets larger.

Uniform distribution and CLT

We can do the same kind of "simulation experiment" for other types of populations. Let's try the Standard Uniform distribution, which has scores evenly distributed between 0 and 1.

```
> X = runif(n)
```

Get a large number of sample means from the Standard Uniform. Now make a histogram of your distribution of sample means. Does the distribution look Normal? The answer will depend on what sample size you used. Try a value of  $n$  close to 1, then try a value greater than or equal to 30.

How well do the mean and standard deviation of the sampling distribution match the predictions

of the CLT? To answer this, first we need to know  $\mu$  and  $\sigma$  for the original population. You should be able to figure out  $\mu$  yourself. Calculating  $\sigma$  is a bit harder, so I give you the answer below, but it's a nice exercise if you want to brush up on integrals.

```
> mu = .5  
> sigma = sqrt(1/12)  
> sigma_M = sigma/sqrt(n)
```

Compare these predictions to the actual values from your set of sample means.