# ISTA 116 Lab: Week 14

Last Revised November 21, 2011

## 1 Sampling Distributions

- A sampling distribution is the probability distribution of a statistic of a random sample.
- **Example**: Rolling a fair 6-sided die. What is the mean of 5 random rolls? 50 rolls? 500 rolls?
- Here the **statistic** we're interested is the mean of the random sample.
- The **parent distribution** is the Discrete Uniform Distribution, with parameter $n = 6$
- We can use Expectation to find what the theoretical mean should be

```
> x <- 1:6
> p <- rep(1/6, times=6)
> sum(x*p)
```

```
[1] 3.5
```

- This is fine, but what happens if we really try this experiment? We will find out by drawing random samples from the parent distribution and finding the mean of the random samples. In other words, let's roll a die 5 times and see what the mean is. Will it be the same as the Expected value?

```
> s <- sample(x, size=5, prob=p)
> mean(s)
```

```
[1] 4
```

- What if we repeat this a few times?

```
> mean(sample(x, size=5, prob=p))
[1] 3.6
> mean(sample(x, size=5, prob=p))
[1] 4
> mean(sample(x, size=5, prob=p))
[1] 3.4
```

- The mean of 5 rolls is itself a random variable with its own distribution called the **sampling distribution**.
- We can find sampling distributions for other statistics as well, such as `median`, `sd`, `var`, etc.

# 2 Sampling Distributions in R

- `R` has built-in functions to draw random samples from common distributions, such as Binomial and Normal.
- **Example**: flipping 3 fair coins simultaneously and counting the number of heads. What is the mean number of heads if we repeat the experiment 5 times? 50 times? 500 times?
- Here the **statistic** we're interested is the mean number of heads.
- The **parent distribution** is the Binomial with $n = 3$ and $p = 0.5$.
- We can use Expectation to find what the theoretical mean number of heads should be

```
> x <- 0:3
> p <- dbinom(x, size=3, prob=.5)
> sum(x*p)
[1] 1.5
```

- To draw random samples from the parent distribution, we will use the `rbinom` function, which takes 3 arguments:

  - `n`: The number of random samples we want to draw.
  - `size`: The number of Bernoulli trials (coin flips) that describes the parent distribution (3 in our example case).
  - `prob`: The probability of success that describes the parent distribution (0.5 for a fair coin).

```
> rbinom(5, size=3, prob=0.5)
[1] 3 0 0 1 2
```

- When we repeat the process of drawing random samples and calculating the mean, we will get different numbers for the mean:

```
> mean(rbinom(5, size=3, prob=0.5))
[1] 1.4
> mean(rbinom(5, size=3, prob=0.5))
[1] 1.6
> mean(rbinom(5, size=3, prob=0.5))
[1] 1.8
> mean(rbinom(5, size=3, prob=0.5))
[1] 1.2
```

- The distribution of these calculated mean values is the sampling distribution for the mean statistic.
- The larger the sample size, the closer we will get to the theoretical mean:

```
> mean(rbinom(5000, size=3, prob=0.5))
[1] 1.4714
> mean(rbinom(5000, size=3, prob=0.5))
[1] 1.4974
> mean(rbinom(5000, size=3, prob=0.5))
[1] 1.4954
> mean(rbinom(5000, size=3, prob=0.5))
```

```
[1] 1.5032
```

- In addition to `rbinom`, `R` can draw random samples from a Normal distribution as well, using `rnorm`.
- `rnorm` takes 3 arguments: `n`, `mean`, and `sd`.

```
> rnorm(5, mean=0, sd=1)
[1]  1.8830844  0.6857073  0.7493470 -0.4231100 -1.9999216
```

---

**Example**: Interactive demo of sampling distributions at onlinestatbook.com

**Exercise**: Generate 10 random samples from a Normal distribution with `mean` 10 and `sd` 3, and find the mean and standard deviation of the samples.

# 3 Visualizing a sampling distribution with `R`

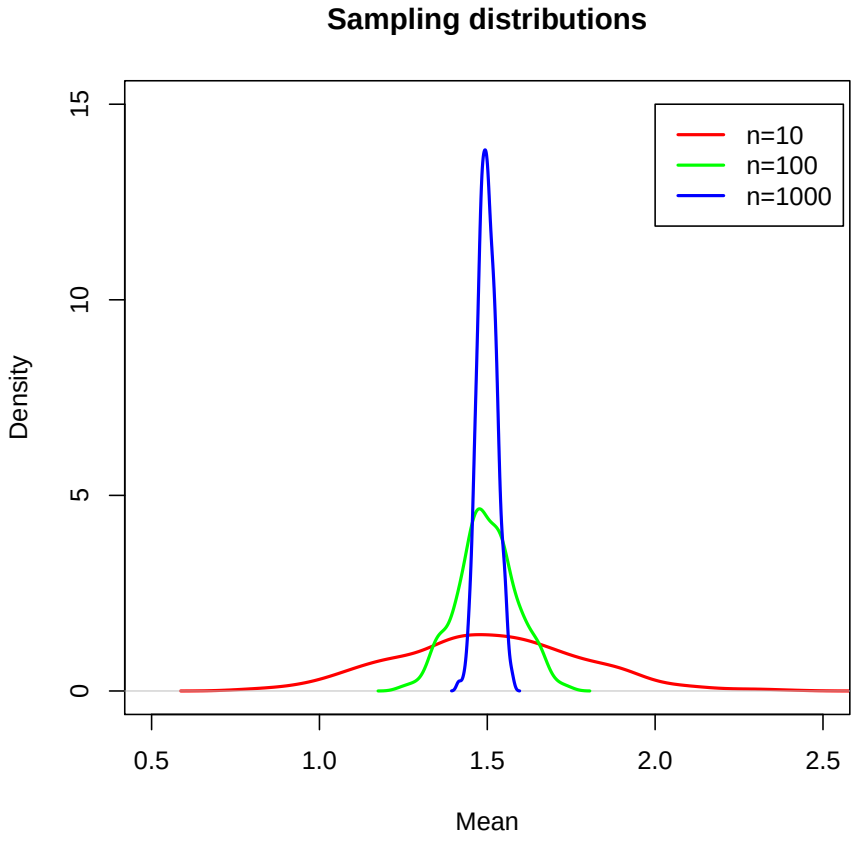
- To visualize the sampling distribution, we need to repeatedly draw random samples, compute the statistic we're interested in, save the results, and then draw a histogram after many repetitions. Doing this by hand would get really boring.
- We can use `R`'s `replicate` function to repeat a task many times automatically. We need to give `replicate` 2 arguments:
  - `n`: The number of times to repeat the task.
  - `expr`: The command to repeat `n` times.
- In the section above, drew samples using `rbinom`, and found the mean of each random sample. Repeating this task 5 times using `replicate` is done as follows, resulting in the mean values of each random sample:

```
> replicate(5, mean(rbinom(5000, size=3, prob=0.5)))
[1] 1.5266 1.5028 1.4972 1.4906 1.5104
```

- Using `replicate`, we can generate thousands of random samples and look at the histogram or density curve of the sampling distribution.

- Let's compare the sampling distributions for the means of random samples of size 10, 100, and 1000 for a binomial distribution with parameters $n = 3$ and $p = 0.5$.

```
> # We'll sample 500 times for each case
> means10 <- replicate(500, mean(rbinom(10, size=3, prob=0.5)))
> means100 <- replicate(500, mean(rbinom(100, size=3, prob=0.5)))
> means1000 <- replicate(500, mean(rbinom(1000, size=3, prob=0.5)))
> plot(density(means10), col='red', lwd=2, xlim=c(0.5,2.5), ylim=c(0,15),
+ xlab="Mean", main="Sampling distributions")
> lines(density(means100), col='green', lwd=2)
> lines(density(means1000), col='blue', lwd=2)
> legend(2,15, legend=c('n=10', 'n=100', 'n=1000'),
+ col=c('red','green','blue'), lwd=2)
```



**Exercise**: Visualize the sampling distributions of the mean and median calculated using random samples from a Normal distribution of mean 10 and standard deviation 3. Use a sample size of 100, and sample 500 times. Plot the density curves of each sampling distribution.