ISTA 116 Lab: Week 15

Last Revised November 29, 2011

1 Sampling Distributions

- When we sample from a probability distribution, any value that summarizes the sample is a **statistic**.
- Since the samples are random, a particular sampling procedure (e.g., draw 5 people from a population) will produce different samples, and therefore different statistics, every time it is applied.
- As a result, any statistic is itself a random variable.

Take another look at the interactive demo of sampling distributions at onlinestatbook.com

Generate a sampling distribution of means for each of the predefined distributions.

1.1 Examples of Sampling Distribution of Mean

1.1.1 An Aside

We know that the *expectation* when rolling a 6-sided die is 3.5:

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

```
[1] 3.5
```

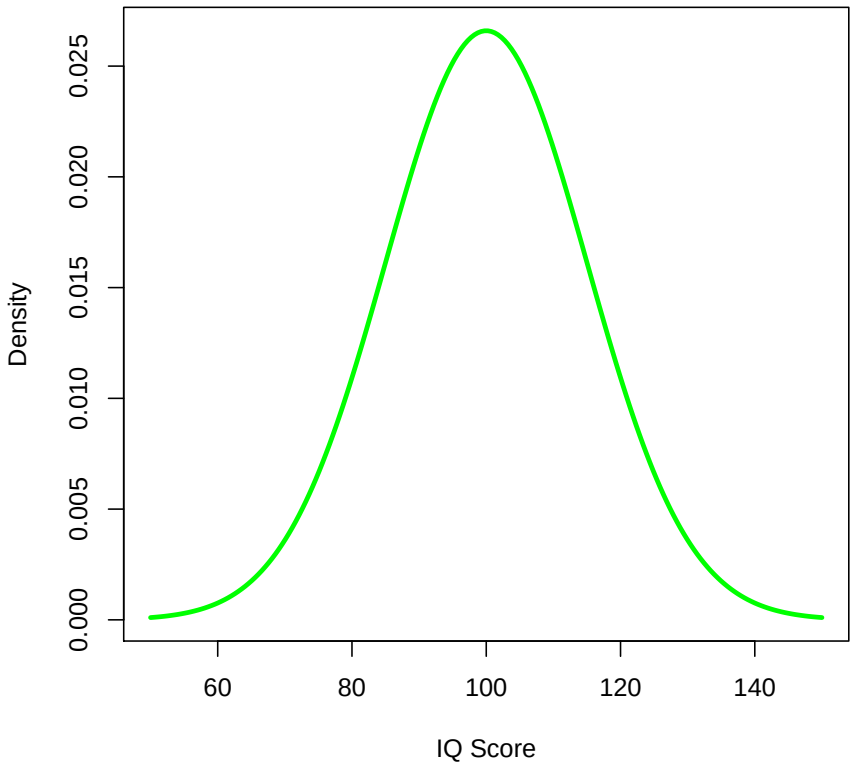
Imagine a very sad game in which you roll n 6-sided dice and find the average of them. You then lose a dollar amount equal to 100 times the absolute difference between the average of your rolls and the expectation. For example, if the mean of your rolls is 1.0 you lose $100 * |1.0 - 3.5| = 100 * 2.5 = \$250$, if your mean is 5.5 you lose \$200.

Would you like to choose the value for n? What would you choose and why? What if you were forced to play this game 10,000 times? Keep this in mind as we go through the next example.

1.1.2 IQ Scores Example

Assume that IQ scores are normally distributed with a standard deviation of 15 points and mean of 100 points.

```
> x = seq(50, 150, length = 200)
> y = dnorm(x, mean = 100, sd = 15)
> plot(x,y,type="l",col="green",xlab="IQ Score",ylab="Density",lwd=3)
```



a.) Use `replicate` and `rnorm` to draw 10,000 samples each of size $n = 1$ from this distribution. Find the means (go ahead and use `mean`, we'll be increasing n next) of each of the 10,000 samples and store them in a variable.

```
> xrange=c(60,140)
> sampleMeans1 = replicate(10000,mean(rnorm(1,100,15)))
```

b.) Plot a histogram of the sampling means collected in part (a). Find the *mean* and *sd* of the 10,000 samples. Draw a vertical line on the histogram at the mean (remember `abline`). Also draw lines at $\pm$ the sd.

```
> hist(sampleMeans1,breaks=30,xlim=xrange)
> mn1 = mean(sampleMeans1)
> sd1 = sd(sampleMeans1)
> abline(v=c(mn1,mn1-sd1,mn1+sd1),col=c("blue","red","red"),lwd=2)
```

c.) Do the same thing using sample sizes $n = 5$, $n = 20$, and $n = 200$. Use *xlim* to give all 4 histograms the same x range.

```
> sampleMeans5 = replicate(10000,mean(rnorm(5,100,15)))
> hist(sampleMeans5,breaks=30,xlim=xrange)
> mn5 = mean(sampleMeans5)
> sd5 = sd(sampleMeans5)
> abline(v=c(mn5,mn5-sd5,mn5+sd5),col=c("blue","red","red"),lwd=2)

> sampleMeans20 = replicate(10000,mean(rnorm(20,100,15)))
> hist(sampleMeans20,breaks=30,xlim=xrange)
> mn20 = mean(sampleMeans20)
> sd20 = sd(sampleMeans20)
> abline(v=c(mn20,mn20-sd20,mn20+sd20),col=c("blue","red","red"),lwd=2)

> sampleMeans200 = replicate(10000,mean(rnorm(200,100,15)))
> hist(sampleMeans200,breaks=30,xlim=xrange)
> mn200 = mean(sampleMeans200)
> sd200 = sd(sampleMeans200)
> abline(v=c(mn200,mn200-sd200,mn200+sd200),col=c("blue","red","red"),lwd=2)
```

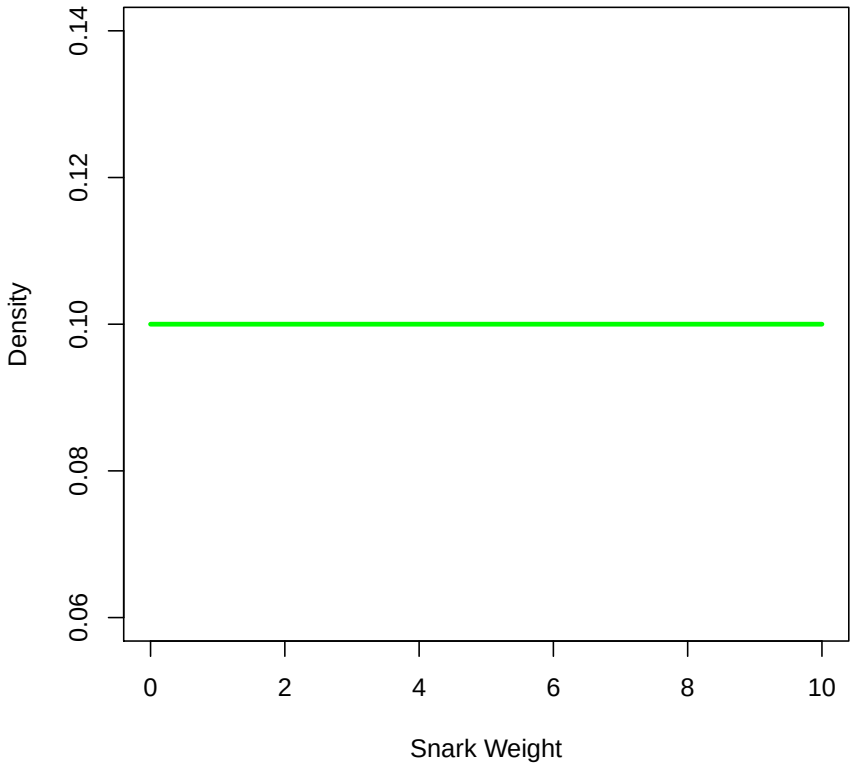
d.)
- What happens to the histogram as we increase n?
- Does this make intuitive sense?
- What if we made a histogram of your losses at the game we described earlier when $n = 1$ compared to a histogram of losses with the value you chose for n?

e.) For each of the 4 sample sizes calculate the sd of the sampling distribution of the mean. Show the mean and sd of the samples for comparison (i.e. the x-values you plotted as lines in your histograms.)

1.1.3 Uniform Distribution Example

Assume that the weight of snarks is uniformly distributed from 0.0 to 10.0 pounds.

```
> x = seq(0.0, 10.0, length = 200)
> y = dunif(x, min=0.0,max=10.0)
> plot(x,y,type="l",col="green",xlab="Snark Weight",ylab="Density",lwd=3)
```



a.) Use `replicate` and `rnorm` to draw 10,000 samples each of size $n = 1$ from this distribution. Find the means (go ahead and use `mean`, we'll be increasing n next) of each of the 10,000 samples and store them in a variable.

```
> xrange=c(0.0,10.0)
> sampleMeans1 = replicate(10000,mean(runif(n=1,min=0.0,max=10.0)))

> hist(sampleMeans1,breaks=30,xlim=xrange)
> mn1 = mean(sampleMeans1)
> sd1 = sd(sampleMeans1)
> abline(v=c(mn1,mn1-sd1,mn1+sd1),col=c("blue","red","red"),lwd=2)
```

b.) Do the same thing using sample sizes $n = 5$, $n = 20$, and $n = 200$. Use *xlim* to give all 4 histograms the same x range.

```
> sampleMeans5 = replicate(10000,mean(runif(n=5,min=0.0,max=10.0)))
> hist(sampleMeans5,breaks=30,xlim=xrange)
> mn5 = mean(sampleMeans5)
> sd5 = sd(sampleMeans5)
> abline(v=c(mn5,mn5-sd5,mn5+sd5),col=c("blue","red","red"),lwd=2)

> sampleMeans20 = replicate(10000,mean(runif(20,0.0,10.0)))
> hist(sampleMeans20,breaks=30,xlim=xrange)
> mn20 = mean(sampleMeans20)
> sd20 = sd(sampleMeans20)
> abline(v=c(mn20,mn20-sd20,mn20+sd20),col=c("blue","red","red"),lwd=2)

> sampleMeans200 = replicate(10000,mean(runif(200,0.0,10.0)))
> hist(sampleMeans200,breaks=30,xlim=xrange)
> mn200 = mean(sampleMeans200)
> sd200 = sd(sampleMeans200)
> abline(v=c(mn200,mn200-sd200,mn200+sd200),col=c("blue","red","red"),lwd=2)
```

c.) **Exercise:** Imagine a snark hunt in which you capture n snarks.

- What is the probability that the mean weight of your n snarks is less than 4 pounds if $n = 200$?
- What if you only catch 20? 5? 1?
- **Can we really do this? Are we making any bad assumptions?**

d.) **Exercise:** Assume normality in the sampling distributions where $n > 1$. What is the probability of a mean weight between 4.5 and 5.5 pounds if we bag 200 snarks? 20 snarks? 5?