Lab 5 Cont: Continuous Random Variables

July 29, 2016

1 Goals

Today we will be reviewing continuous random variables in `R` and also getting a preview of Friday's lecture on the Central Limit Theorem and sampling distributions.

- Review calculation of variance
- `R` commands for continuous random variables
- Examine properties about expectation and variance of linear combinations of random variables

2 Estimating Parameters

First, suppose we have a discrete distribution over the numbers $0, 1, 2, \ldots 10$ where each value has an equal probability of $1/11$. In this case, we could easily calculate the true mean and standard deviation of our distribution.

$$\mu_x = \frac{1}{11}\sum_{i=0}^{10} i = 5$$

$$\sigma_x^2 = \sum_{i=0}^{10}(i-\mu_x)^2 p(i) = \sum_{i=0}^{10}(i-5)^2\frac{1}{11} = \frac{1}{11}\sum_{i=0}^{10}(i-5)^2 = 10$$

Notice, that the last term looks a lot like the formula for variance we learned the first week of class for a set of data. However, it differs slightly in one way. Here, we divided by N instead of $N-1$. Also, we use μ instead of $\bar{x}$. Here, we are calculating the population variance, while before we were calculating a statistic which might be used to estimate the population variance. Why do we divide by N? It essentially boils down to the fact that we know the true mean when we are calculating the population variance.

Let's take a look at the effect of using $N-1$ versus N and $\bar{x}$ instead of μ when trying to estimate the variance from sampled data.

First, let's take a sample from our distribution, since it is just a discrete uniform distribution, we can use the `sample` command which we have seen before. Let's randomly select a set of 5 numbers from our sample. Note that we are sampling with replacement.

```
set.seed(10)
observed.sample <- sample(c(0:10), size = 5, replace = T)
```

Before we consider the N vs $N-1$ issue, let's first tackle μ vs $\bar{x}$.

2.1 Question

- What do you think is typically smaller? $\sum_i (x_i - \mu_x)^2$ or $\sum_i (x_i - \bar{x})^2$

```
var.mu <- sum((observed.sample - 5)^2)
var.x.bar <- sum((observed.sample - mean(observed.sample))^2)
var.mu
```

```
## [1] 34
```

```
var.x.bar
```

```
## [1] 26.8
```

So it looks like in this case, using $\bar{x}$ results in a smaller squared error than μ. Is this always the case? Let's take many samples and take a look. Let's sample data 5000 times, and see how many times the square errors using $\bar{x}$ is smaller than using μ_x

```
set.seed(11)
x.bar.is.smaller <- rep(0, 5000)

for(i in 1:5000){
  observed.sample <- sample(c(0:10), size = 5, replace = T)
  var.mu <- sum((observed.sample - 5)^2)
  var.x.bar <- sum((observed.sample - mean(observed.sample))^2)
  x.bar.is.smaller[i] <- (var.x.bar <= var.mu)
}

mean(x.bar.is.smaller)
```

```
## [1] 1
```

We can see that using $\bar{x}$ is less than or equal to using μ_x every single time. There are a few times, when the sum of squared errors using both procedures is 0, (ie when all the observations in the sample are 5), but that happens only a small percentage of the time. Why does using $\bar{x}$ always result in a smaller sum of squared errors than using the true μ? Well intuitively, we know that $\bar{x}$ adapts to my data. We can show it rigorously using the following proof.

Suppose I could pick any value (not just μ or $\bar{x}$) to make the squared errors as small as possible. Let's call that value a. Using what you've learned in calculus, to minimize an equation, we take the derivative and solve for 0.

$$0 = \frac{\partial \sum_i (x_i - a)^2}{\partial a} = 2 \sum_i (x_i - a^\star)$$

$$0 = \sum_i x_i - \sum_i a^\star = \sum_i x_i - N a^\star$$

$$\Rightarrow a^\star = \frac{1}{N} \sum_i x_i$$

So we can see that $\bar{x}$ actually makes the sum of squared errors as small as possible, so using any value besides $\bar{x}$ results in a larger sum of squared errors.

The real population variance is defined by using μ though. So when we use $\bar{x}$ instead, what will happen to our estimates?

2.2 Questions

- How will this affect bias in estimating the true population variance?

Let's take a look by checking for 5000 simulations where we divide by $N-1$.

```
set.seed(111)
variance.est.x.bar <- rep(0, 5000)
variance.est.mu <- rep(0, 5000)
for(i in 1:5000){
  observed.sample <- sample(c(0:10), size = 5, replace = T)
  variance.est.x.bar[i] <- 1/(5-1) * sum((observed.sample - mean(observed.sample))^2)
  variance.est.mu[i] <- 1/(5-1) * sum((observed.sample - 5)^2)
}


mean(variance.est.x.bar)

## [1] 10.16152

mean(variance.est.mu)

## [1] 12.6423
```

The true value of the variance is 10. As we can see, the estimate using μ is higher than the truth, but the estimate using $\bar{x}$ seems very close to the truth.

Let's see what would happen when we use N instead of $N-1$

```
set.seed(111)
variance.est.x.bar <- rep(0, 5000)
variance.est.mu <- rep(0, 5000)
for(i in 1:5000){
  observed.sample <- sample(c(0:10), size = 5, replace = T)
  variance.est.x.bar[i] <- 1/(5) * sum((observed.sample - mean(observed.sample))^2)
  variance.est.mu[i] <- 1/(5) * sum((observed.sample - 5)^2)
}

mean(variance.est.x.bar)

## [1] 8.129216

mean(variance.est.mu)

## [1] 10.11384
```

Again, the true value of the variance is 10. However, the story has changed this time, the estimate using μ is seems pretty good on average, but the estimate using $\bar{x}$ seems biased to be smaller than the truth.

So what's the punchline of all of this? If you know the true mean, then you can divide by N and get a better estimate of the true population variance. However, in most cases, when we don't know the true population mean, and use $\bar{x}$ instead, it seems that using $N-1$ results in a better estimate.

3 Continuous Distributions in R

The PDF of the normal distribution is given by: $f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$

The function that computes the density $f(x)$ at a given point x is `dnorm(x,mean,sd)`. Again, 'norm' stands

for "Normal", and the 'd' stands for "density". Let's calculate the density at $x = 0$ for the Standard Normal. Note that we need to specify the mean and standard deviation (sd) of the normal distribution in the function.

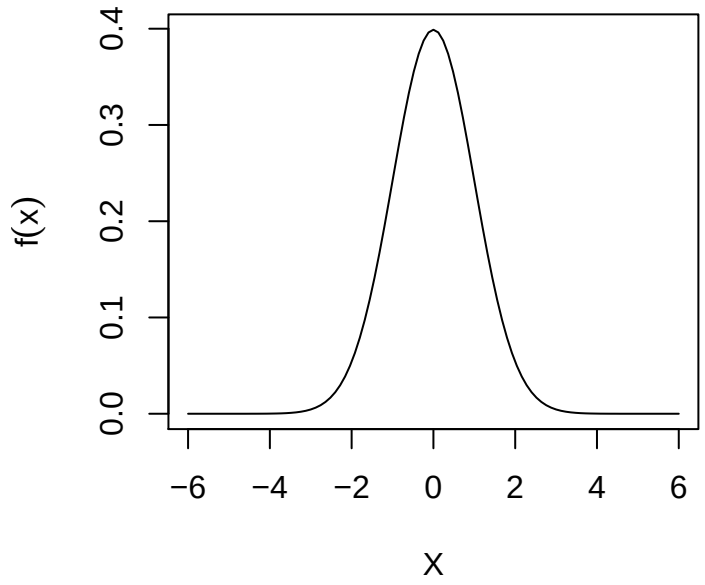
```
dnorm(0, mean = 0, sd = 1)
```

```
## [1] 0.3989423
```

If we don't specify mean or sd in the function, they are set to 0 and 1 respectively by default (the Standard Normal).

In fact, you can draw the entire distribution using the curve function to draw it:

```
curve(dnorm(x), xlim = c(-6, 6), xlab = expression(X),
      ylab = expression(f(x)))
```



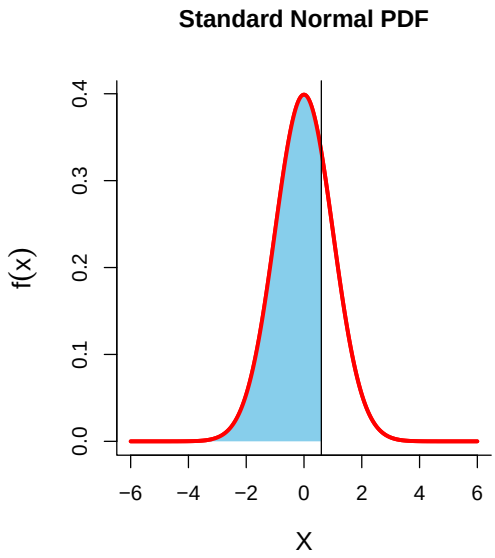
For the CDF, R has two functions. One allows you to calculate the probability that a normal random variable X falls below a certain value, and the other allows you to calculate the inverse, the value X below which a certain fraction of the data lie.

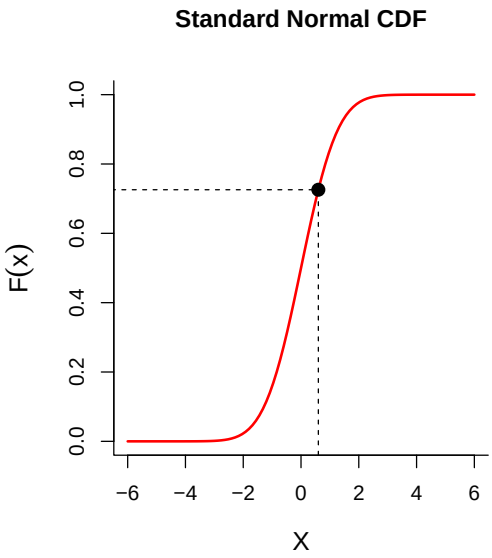
Let's assume X is a standard normal random variable. Then the probability that $X \leq 0.6$ can be found using the function pnorm

```
pnorm(0.6, mean = 0, sd = 1)
```

```
## [1] 0.7257469
```

PDF and CDF together:





The quantile function gives you the inverse of this (the inverse CDF is often referred to as Φ^{-1}). Use this when you want to find x such that $P(X \leq x) = p$. The function is called qnorm(). Again, let's assume X is a standard normal random variable. Then to find the x such that $P(X \leq x) = 0.975$, we run the following command:

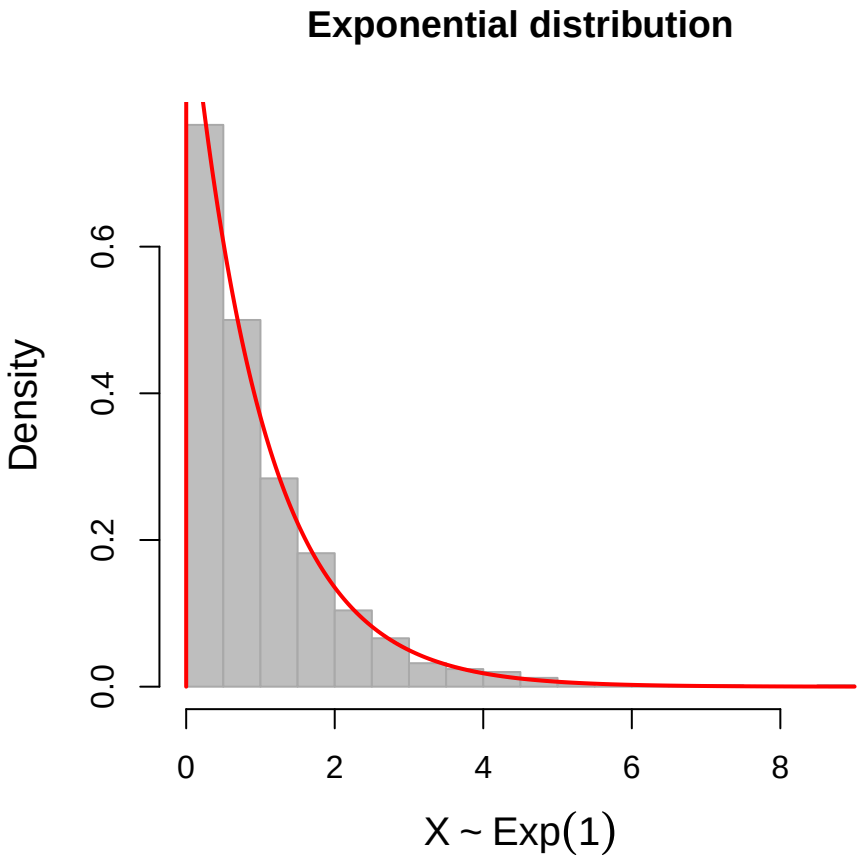
```
qnorm(0.975, mean = 0, sd = 1)
```

```
## [1] 1.959964
```

The exponential distribution is used to model waiting time between events, for example, the time between customer arrivals in a busy restaurant. We can get all the same functions by using the "*exp" commands.

Compare the histograms of random samples drawn from this distribution, with the theoretical distribution (superimposed curves):

```
par(bty="l", cex.lab=1.25)
hist(rexp(1000, rate=1), col="grey", xlab = expression(X %~% Exp(1)), freq=FALSE, breaks=20,
bor="darkgrey", main='Exponential distribution')
curve(dexp(x, rate=1), lwd=2, col=2, add=TRUE, n=10000)
```



3.1 Lab 5 Assignment

Complete the lab assignment document on the catalyst website.