MAGIST

Geog 413

Module 1 Exercise 4: The Normal Probability Distribution

1. R has a great deal of functionality for working with probability distributions of different kinds.

2. We typically separate probability distributions into two types – those dealing with discrete distributions where the data can only take on certain values like integers, and those dealing with continuous distributions where the data can take on any value within a specified range.

3. The most commonly used discrete probability distributions include the Binomial and Poisson distributions. The most commonly used continuous probability distributions are the Normal, Student's t and chi-squared distributions.

4.R uses an abbreviated name for every probability distribution. Norm represents the normal distribution. For each distribution the following prefixes give you different information – dnorm generates the normal density function (sometimes known as the probability mass function), pnorm provides cumulative probabilities associated with the normal distribution, qnorm yields normal quantiles and rnorm gives you observations on normal random variables. We will explore these functions in a moment.

5. As usual you can get help on a probability distribution as

```
> ?Binomial or ?Normal
```

Generating Random Numbers

1. To generate random numbers from a specific probability distribution we can use the prefix r linked to the distribution type. So to generate 10 random numbers from a normal probability distribution with a mean of 100 and a standard deviation of 12

```
> rnorm(10, mean=100, sd=12)
 [1] 122.51887 119.08965  86.60163  85.50060  92.40772 129.02680 106.56703
 [8] 103.20224  97.64806 115.53455
```

If you do not specify the optional values of the mean and standard deviation, the command assumes you are looking for observations from the standard normal distribution which has a mean of 0 and a standard deviation of 1

```
> rnorm(10)
 [1]  0.1135948  0.5316454  1.4075156  0.9693813 -0.7009500  1.0917987
 [7]  1.0299192 -0.3044963  0.5709020  0.2529149
```

2. The binomial distribution is often associated with statistical trials where there are only two outcomes – a success or a failure. The classic example is flipping a coin (heads or tails are the outcomes). But we can also use the binomial distribution in different situations such as rolling a die where success is defined as getting

the number 3 and failure is defined as not getting the number 3. We use the binomial distribution to figure out the probability of a success (or a failure) in repeated trials of an experiment. In the binomial distribution, we assume that the probability of a success is the same from one trial to the next. And we assume that each observation or trial is independent from any other. To generate random variables from the binomial distribution across a series of trials with a given probability, you need to specify the number of values that you want (x), the number of trials of the experiment (size) and the probability of success (probability)

```
> rbinom(x, size, probability)
```

So assume that we are focused on rolling a die and getting the number 6. The probability of a success is 1/6 = .166666. We will perform this experiment 100 times. The value of the random variable represents the number of times we expect a success (getting the number 6) over the 100 trials. So if we specify an x value of 1

```
> rbinom(1, 100, .166666)
 [1] 12
```

we expect to see 12 successes (getting the number 6) in 100 rolls of a die. We can specify different values of x where we are asking to repeat the whole experiment multiple times, generating the number of successes each time

```
> rbinom(10, 100, .166666)
[1] 13 16 19 16 15 18 17 12 17 16
```

The values that we have to specify for each different probability function (the mean and standard deviation for the normal distribution and the number of trials and the probability of success in the binomial distribution) are known as the parameters of the different distributions. They control the shape of the probability distribution.

Calculating Probabilities for Discrete Distributions

We can easily calculate the probability that the discrete random variable (X) takes a specific value. This could be written as P(X=x) where the upper case X represents the random variable and the lower case x represents a specific value of that variable. So this is the same as asking, in 5 coin tosses, what is the probability of observing 3 heads?

```
> dbinom(3, 5, .5)
[1] 0.3125
```

Here d stands for the density function. What about the probability of getting 5 successes in 5 trials?

```
> dbinom(5, 5, .5)
[1] 0.03125
```

Yes, this is much rarer.

Calculating Cumulative Probabilities for Discrete and Continuous Random Variables

1. Note that while R will return a value for dnorm, strictly speaking the probability of a continuous variable assuming a specific value P(X=x) is 0. However, we can readily use the cumulative probability function (pnorm) with both discrete and continuous random variables. The cumulative probability function gives us the probability that $P(X \leq x)$, that the random variable X takes a values equal to or less than the value x.

Let's look at this for the standard normal distribution

> pnorm(1, 0, 1) or just > pnorm(1)
[1] 0.8413447

So there is an 84.13% chance that a value drawn at random from a normal probability distribution with mean = 0 and standard deviation = 1 is less than or equal to 1. Does this make sense? Well you know the area under the normal probability curve to the left of 0 (the center of the standard normal distribution) is 0.5 and you know the area between -1 and +1 standard deviations about the mean is about ~68%. So 0.5 plus half of .68 is roughly 0.84.

Now calculate the probability than the random variable is less than or equal to 0.9, 0.99, 0.999, 0.9999. You will see that this probability approaches .8413. That means the area under the normal curve between 0.99999999 and 1.0 is essentially zero (this is basically P(X=1)). That is what we mean when we say that for continuous distributions, P(X=x) = 0.

2. OK so let's generate the cumulative probability that the normally distributed random variable with mean = 3.0 and standard deviation = 1.0 takes a value less than 2.

> pnorm(2, 3, 1)
[1] 0.1586553

This makes sense, because you know the area under the curve (the probability distribution) to the left of the mean = 0.5 and you know the area between -1 sd and +1 sd = .68. So the area to the left of -1 sd under the curve ~ 0.5-.34 ~ 0.16. Don't forget the value x=2 is 1 sd less than the mean.

What if you wanted to know the probability that the normally distributed random variable takes a value greater than 2 in this case? $P(X > 2)$ = 1-0.1586553 = 0.84134.

Now, what about the probability of getting 0, 1 or 2 heads ($P(X \leq 2)$ in 3 flips of a fair coin? Here we have to make use of the binomial probability function again

> pbinom(2, 3, 0.5)
[1] 0.875

Well the probability of getting 3 heads in 3 flips = 0.5*0.5*0.5 = 0.125 (dbinom(3, 3, 0.5) and so this probability also looks ok.

Finally, what about the cumulative probability that a random variable from the Student's t-distribution takes a value less than 1.96? (Think about the normal distribution. Getting a value of $X \leq 1.96$ from a standard normal distribution would be 0.975 (ie 2.5% of the area under the curve is to the right of 1.96).) As the sample size and degrees of freedom increase in the Student's t-distribution, that distribution looks more and more like the normal distribution. The parameters for the Student's t-distribution are x and the degrees of freedom

> pt(x=1.96, df=100) or just pt(1.96, 100)
[1] 0.9736105

And if we increase the degrees of freedom (n-1)
> pt(1.96, 1000)
[1] 0.9748634

We get close to 0.975.

Quantiles

1. Given a probability (p) and a distribution, you might want to calculate the quantile for p, that is the value of x for which $P(X = x) = p$. So given a standard normal distribution, what is the value of X, for which p=0.5. This is easy it should be X=0. Check by finding the quantile for the normal distribution given p=0.5

> qnorm(.5, 0, 1)
[1] 0

What about the quantile for a standard normal distribution and the value p=0.8415? Think about it, this should be ~ X=1.0

> qnorm(.8415, 0, 1)
[1] 1.000642

And the quantile value for a normal distribution with mean=5.0 and standard deviation=0.8 given the probability 0.025?

> qnorm(.025, 5, 0.8)
[1] 3.432029

This value X=3.432 is about 2 standard deviations left of the mean, which should leave only about 2.5% of the area under the curve to the left.

***And don't forget, you can easily convert values for any normal distribution into standardized values for the standard normal distribution, what we call Z-scores

$$Z = (X_i - \mu)/\sigma$$

So in this example, Z = (3.432-5)/0.8 = -1.96875

And then,

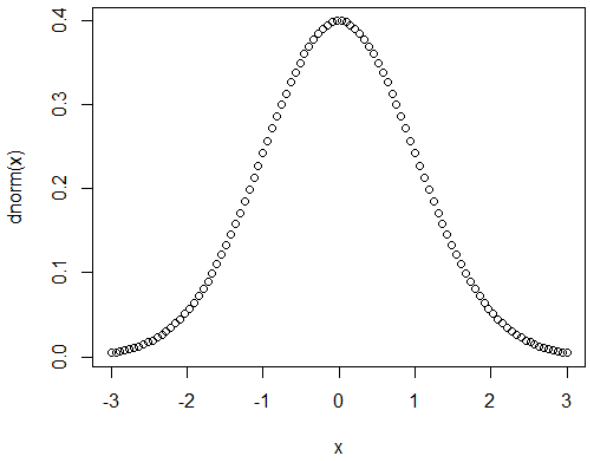
```
> qnorm(.025)
[1] -1.959964
```

Pretty close to our guess.

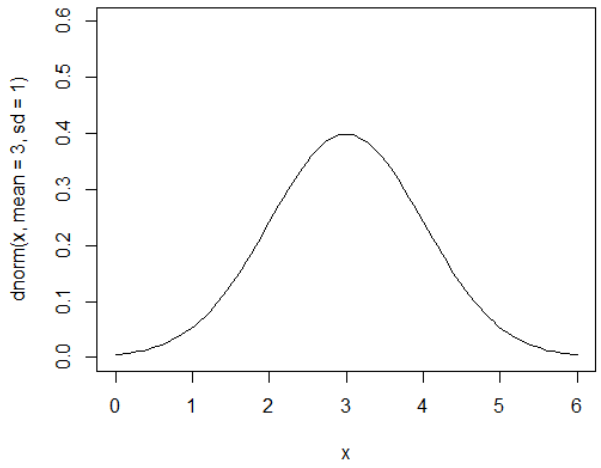
Plotting a Density Function

Let's generate a normal probability distribution. Note that here we use the density function (dnorm), even though P(X=x) = 0 for a continuous variable.

```
> x <- seq(from=-3, to=+3, length.out=100)
> plot(x,dnorm(x))
```



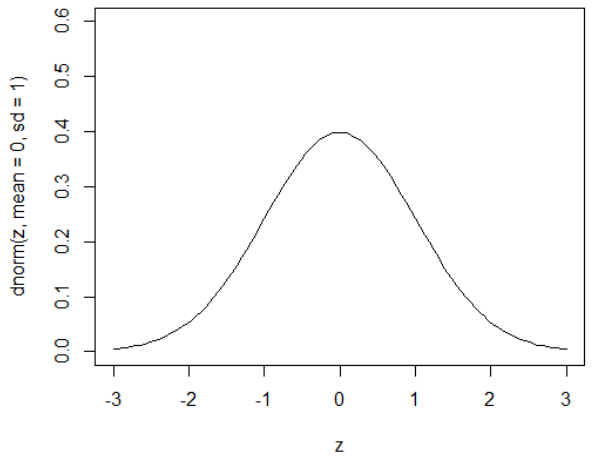
To generate a normal plot for a random variable with a mean of 3 and a standard deviation of 1, first generate 100 observations over the interval 0 to 6, then plot the values of x against the density function for a normal probability distribution with a mean of 3 and a standard deviation of 1. Type="l" specifies a solid line, and ylim=ylim specifies that the y-axis should run between the values 0 and 0.6.




```
> x <- seq(from=0, to=6, length.out=100)
> ylim <- c(0, 0.6)
> plot(x, dnorm(x, mean=3, sd=1), type="l", ylim=ylim)
```

Notice how we can convert this normal probability distribution with a mean of 3 and a standard deviation of 1, to the standard normal distribution with mean=0 and sd=1

```
> z <- (x-3)/1
> plot(z,dnorm(z,mean=0,sd=1), type="l", ylim=ylim)
```



Finally, on plotting, it is relatively easy to combine different plot types. The dissolved oxygen (DO) data discussed earlier for our sample of 50 lakes is replicated below (It can also be read into R from the dodata.csv file under Module 1).

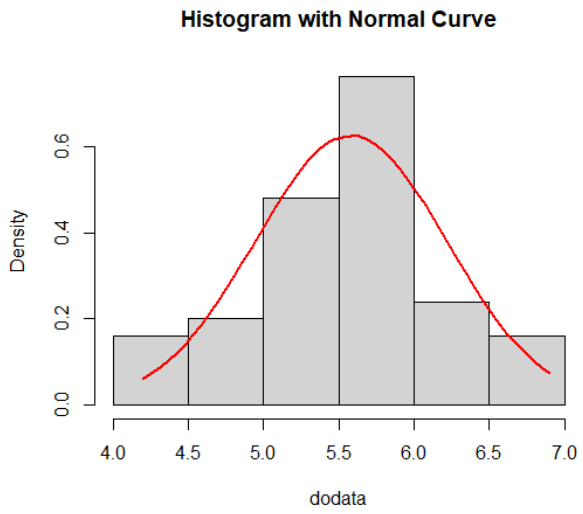
```
> dodata <- c(4.2, 4.3, 4.4, 4.5, 4.7, 4.7, 4.8, 4.8, 4.9, 5.1, 5.1, 5.1, 5.2, 5.3, 5.3, 5.4, 5.4, 5.4, 5.4, 5.5, 5.5, 5.6, 5.6, 5.6, 5.6, 5.6, 5.6, 5.7, 5.7, 5.7, 5.7, 5.8, 5.8, 5.8, 5.8, 5.9, 5.9, 5.9, 5.9, 6.0, 6.1, 6.2, 6.3, 6.4, 6.4, 6.4, 6.6, 6.7, 6.8, 6.9)

> hist(dodata, prob=TRUE, main="Histogram with Normal Curve")

> x <- seq(min(dodata), max(dodata), length=50)

> y <- dnorm(x, mean=mean(dodata), sd=sd(dodata))

> lines(x,y, col="red", lwd=2)
```



Just two final things:

Combinations Function

You often want to perform combinations calculations, think, how many distinct subsets of size k can be drawn from a n items. Use the choose function

```
> choose(n=5, k=3)
```

[1] 10

Sample function

Let's say we want to sample 10 observations from the dissolved oxygen dataset (the object dodata). Here we employ the sample function

```
> sample(dodata, 10)
```

[1] 5.6 5.1 5.4 5.1 5.2 5.9 5.6 4.3 5.5 6.0

This is sampling without replacement (ie we don't replace the observations already selected as we repeatedly select new ones). If you want to sample with replacement

```
> sample(dodata, 10, replace=TRUE)
```