Lab 2. Randomization and Normal Random Variables

1. Randomizing treatments using R

As we saw in our first lecture, randomization is an essential part of a designed experiment, as it can minimize the effect of unmeasured confounding factors and allow us to infer a *causal* relationship between the treatments and the results of an experiment. Section 3.2 of Dean and Voss describes an approach for randomly assigning treatments to experimental units, and Section 3.8 gives some example R code to do so. Here I give an example of how to randomly assign treatments to experimental units using R.

Consider randomly assigning 6 plots to be given either treatment A or treatment B. So we have $n = 6$ experimental units and $r_A = 3$ will be assigned to treatment A and $r_B = 3$ will be assigned to treatment B. observations each. Following DV 3.2, we will do the following in R:

1. Make a vector with $r_A = 3$ A's and $r_B = 3$ B's
2. Make a vector with the plot numbers 1 through $n = 6$
3. Draw $n = 6$ random variables
4. Find the ordering of the random variables and take that order to be the experimental unit that will receive the treatments in "labels"

The result is a completely randomized design, where the treatments were randomly assigned to each experimental unit.

Steps 3-4 above essentially result in a random re-shuffling of the vector of treatments. This keeps the correct number of treatments r_A and r_B, but randomly assigns them to experimental units. In R, we can do this reshuffling easily using the "sample" command. Here is a block of R code that will create a completely randomized design (CRD) for this experiment:

```
treatments.not.random=c(rep("A",3),rep("B",3))
plot.ID=1:length(treatments.not.random)
treatments.random=sample(treatments.not.random)
data.frame(plot.ID,treatments.not.random,treatments.random)
```

```
##   plot.ID treatments.not.random treatments.random
## 1       1                     A                 B
## 2       2                     A                 A
## 3       3                     A                 A
## 4       4                     B                 B
## 5       5                     B                 B
## 6       6                     B                 A
```

```
## resulting randomized treatment
CRD=data.frame(plot.ID,treatment=treatments.random)
CRD
```

```
##   plot.ID treatment
## 1       1         B
## 2       2         A
## 3       3         A
## 4       4         B
## 5       5         B
## 6       6         A
```

Note that if our design changes (for example we need to assign a different number of units to the "A" treatment), all that is needed is to change the line where we define the "treatments.not.random" so that it contains the correct number of each treatment. The rest of the code can stay the same. For example, if we had an experiment where there were 4 treatments (A, B, C, and D), and we wanted to assign each treatment to two experimental units, then code to create a CRD for this case could look like this:

```
treatments.not.random=c(rep("A",2),rep("B",2),rep("C",2),rep("D",2))
plot.ID=1:length(treatments.not.random)
treatments.random=sample(treatments.not.random)
CRD=data.frame(plot.ID,treatment=treatments.random)
CRD
```

```
##   plot.ID treatment
## 1       1         C
## 2       2         D
## 3       3         D
## 4       4         A
## 5       5         B
## 6       6         B
## 7       7         C
## 8       8         A
```

2. Simulating Normal Random Variables in R

R has the ability to sample random variables from the Normal, and other, distributions. The following code will let you easily sample N independent random variables $X_1, X_2, \ldots, X_N$ where

$$X_i \sim N(\texttt{mn}, \texttt{sd}^2)$$

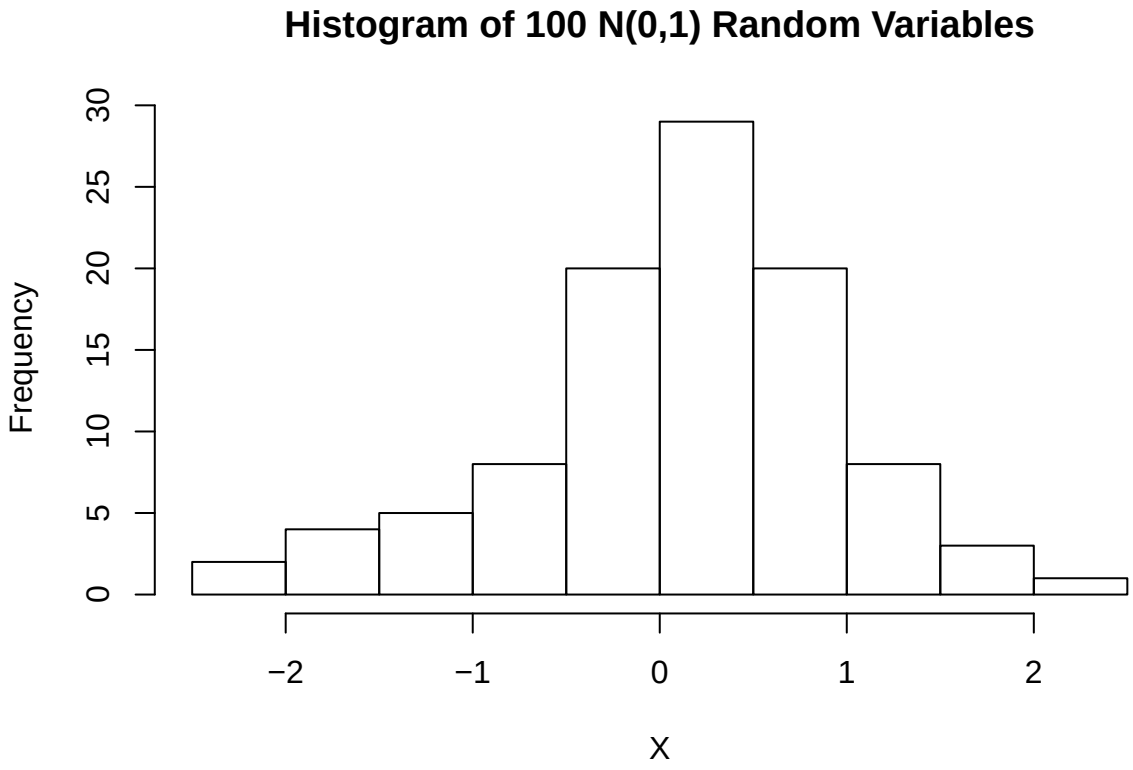
```
N = 100 ## The number of RVs to generate
mn = 0  ## The mean
sd = 1  ## The standard deviation

X=rnorm(N,mean=mn,sd=sd)
X
```

```
##   [1]  0.053267804 -0.046745191  0.234284468 -0.493618148 -1.589538510
##   [6]  0.792626900 -0.368997953  0.863760031  0.786190295  0.075048536
##  [11]  0.642258878 -0.475152931  0.155849435  0.955779313  2.326139434
##  [16] -0.324356927  1.721020853 -0.486255411  0.089253014 -0.332762955
##  [21]  0.089602628 -0.032713070 -0.769334006  0.969295763 -0.849538381
##  [26]  1.037879820 -1.997932565 -1.276023633  0.418968441  0.525969799
##  [31] -0.804313130  1.035283031 -2.114946223  1.674782196  0.690182587
##  [36]  0.587311733 -0.214632634  0.383586748  1.276070605  0.382928361
##  [41] -1.455992640  0.233126120 -0.633861278  0.650472531 -0.936849228
##  [46]  0.402303117  0.446888953  0.029941706 -0.159252256 -0.189987865
##  [51]  0.002662618  1.027551046  0.223779879 -0.124554270 -0.408478250
##  [56]  0.600322026  1.181974028  0.750450661  0.939258710  0.666703957
##  [61]  0.201529826  0.494565484  0.027002325  1.482287591  0.228580799
##  [66]  1.876425296  1.077479148  0.114318930  0.834884703  0.279693261
##  [71]  0.097755989 -0.885302011  0.305273626  0.851995844 -0.474240693
##  [76]  0.437568587 -0.334220966 -0.017853027 -1.494145382  0.119974409
##  [81] -0.030938455  0.548029263  0.189145582 -2.174457701  0.676724376
##  [86] -0.495502150 -1.413010646 -0.695778950 -1.740586483  0.946792429
```

```
##  [91] -0.130086443  0.402687225 -1.616460824 -0.918371323  1.116348246
##  [96]  0.187405722 -1.452518963 -0.422032257  0.614897179  0.410261367
```

```
hist(X,main="Histogram of 100 N(0,1) Random Variables")
```

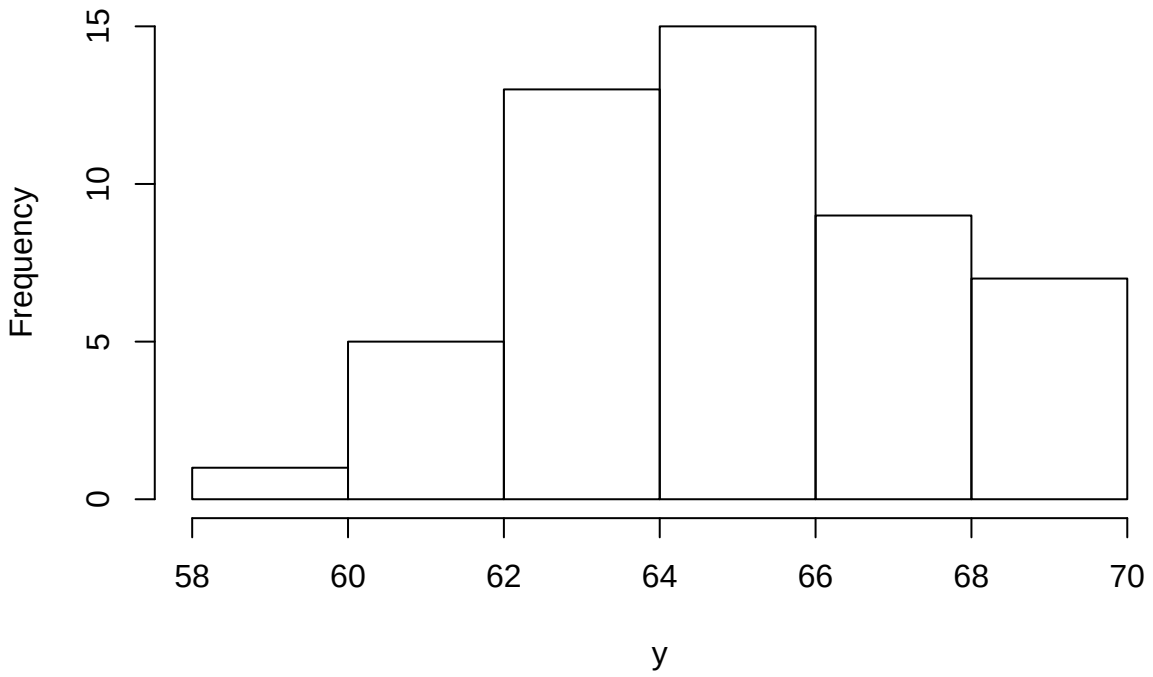


The simulated random variables are in the data in X.

Note that we defined the N, `mn`, and `sd` in this code before the "rnorm" command. We could instead sample the random variables using one line of code. For example, the following code will simulate 50 random normal variables with mean 65 and variance 5.

$$y_i \sim N(65, 5)$$

```
y=rnorm(n=50,mean=65,sd=sqrt(5))
hist(y,main="Histogram of 50 N(65,5) Random Variables")
```



Multiple Normal random variables could be added together by first simulating each, and then adding. For example, let

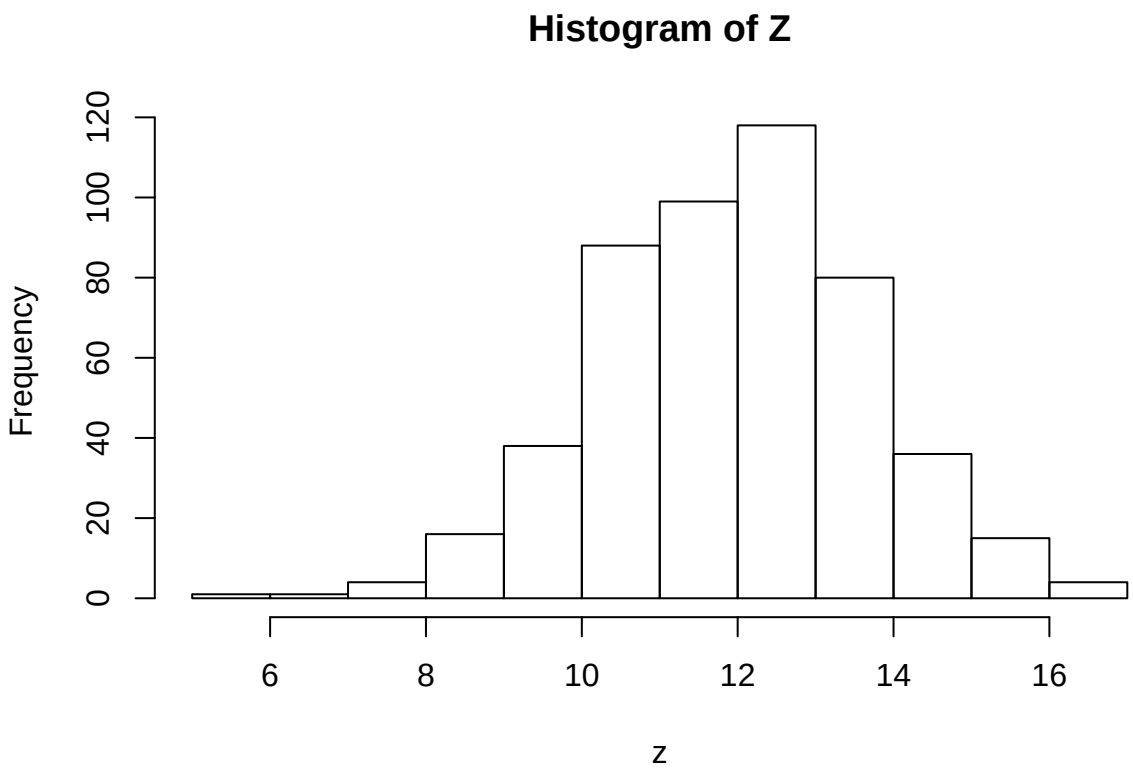
$$x_i \sim N(13, 1), \quad y_i \sim N(-1, 2), \quad x_i \perp\!\!\!\perp y_i$$

Then we can simulate $z_i = x_1 + y_i$ by the following code

```
x=rnorm(500,mean=13,sd=1)
y=rnorm(500,mean=-1,sd=sqrt(2))
z=x+y
alldata=data.frame(x,y,z)
## look at data
head(alldata)
```

```
##          x          y        z
## 1 11.88368 -0.3134240 11.57026
## 2 13.62018 -1.5717420 12.04844
## 3 12.36432 -0.6625851 11.70174
## 4 10.74953 -0.9929059  9.75662
## 5 13.00359 -2.0667119 10.93688
## 6 13.21375 -2.7739887 10.43976
```

```
## plot data
hist(z,main="Histogram of Z")
```


Histogram of Z
Frequency
0
20
40
60
80
100
120
6
8
10
12
14
16
z

Homework Assignment

1. Suppose that you are planning to run an experiment with one treatment factor having four levels: "none", "low", "medium", and "high", and you have the resources to conduct the experiment on 20 experimental units. Assign at random 20 experimental units to the 4 levels of the treatment, so that each treatment is assigned 5 units. Your answer should include your R code used.
2. Repeat question 1 to obtain a second experimental design assigning the 20 units to the 4 levels of the treatment.
3. Suppose that you are planning to run an experiment with one treatment factor having three levels. It has been determined that $r_1 = 3$, $r_2 = r_3 = 5$. Assign at random 13 experimental units to the 3 treatments so that the first treatment is assigned 3 units and the other two treatments are each assigned 5 units.
4. Visit http://www.tylervigen.com/spurious-correlations (or some other website of your choosing) and find an example of two observed quantities that are correlated, but you think are not causally related. Clearly show the data (you could download an image), and describe why you think the two quantities are not causally related. Give an example of another factor (not measured) which you think could have a causative link with one or both of the quantities shown. Give some explanation for why this not measured factor could be causally linked to one or both of the quantities.
5. Let $X \sim N(2, 6)$ and $Y \sim N(-3, 2)$ and $Z \sim N(0, 1)$. All three random variables are independent of each other. Do the following. Show all work.
   (a) What is the distribution of $W = X + Y + Z$? What are $E(W)$ and $Var(W)$?
   (b) What is the distribution of $Q = 2Y$?
   (c) What is the distribution of $P = -2X + 4$?
   (d) Find a and b so that $M = a + bX$ is distributed as a standard Normal distribution.
6. Do the following
   (a) Use R to simulate 1000 iid random variables $\{X_i\}$ with $X_i \sim N(-2, 3)$. Plot a histogram of your simulated values.
   (b) Also simulate 1000 iid random variables $\{Y_i\}$ with $Y_i \sim (3, 1)$. Plot a histogram of your simulated values.
   (c) Finally, plot a histogram of $\{Z_i\}$, where $Z_i = X_i + Y_i$.
   (d) Is Z_i independent of X_i? Explain your answer.
   (e) Find the sample mean and variance of the Z_is you simulated, and compare them with the true, theoretical mean and variance.