Assignment 2

Jiarong Ye

September 13, 2018

Q1

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.

```
In [119]: random_units = sample(c(rep('none', 5), rep('low', 5), rep('medium', 5), rep('high', 5
          assigned_dataset1 = data.frame(1:20, random_units)
          colnames(assigned_dataset1) = c('id', 'experimental_units')
          assigned_dataset1
```

id	experimental_units
1	none
2	medium
3	medium
4	high
5	none
6	low
7	high
8	none
9	medium
10	high
11	none
12	low
13	high
14	low
15	medium
16	low
17	low
18	high
19	medium
20	none

Q2

2.Repeat question 1 to obtain a second experimental design assigning the 20 units to the 4 levels of the treatment.

```
In [120]: random_units = sample(c(rep('none', 5), rep('low', 5), rep('medium', 5), rep('high', 5
          assigned_dataset2 = data.frame(1:20, random_units)
          colnames(assigned_dataset2) = c('id', 'experimental_units')
          assigned_dataset2
```

id	experimental_units
1	low
2	low
3	low
4	high
5	none
6	none
7	none
8	low
9	high
10	medium
11	medium
12	medium
13	high
14	high
15	medium
16	none
17	medium
18	high
19	low
20	none

Q3

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.

```
In [121]: random_units = sample(c(rep('r1', 3), rep('r2', 5), rep('r3', 5)))
          assigned_dataset3 = data.frame(1:13, random_units)
          colnames(assigned_dataset3) = c('id', 'experimental_units')
          assigned_dataset3
```

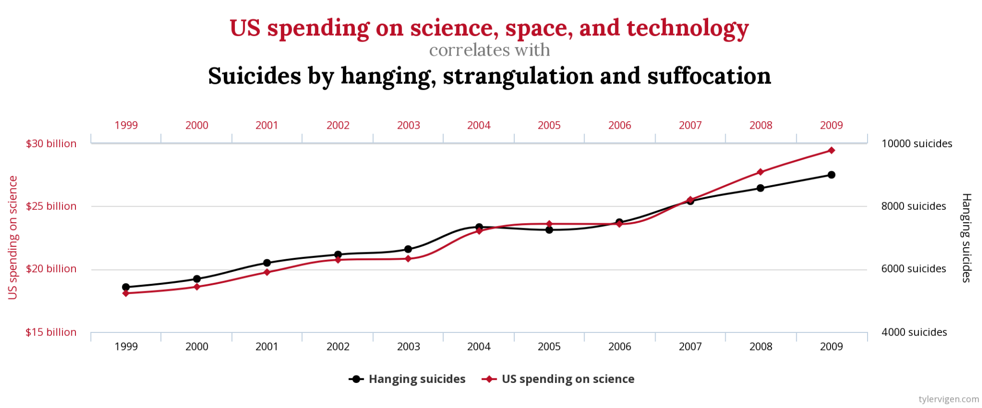
id	experimental_units
1	r3
2	r2
3	r2
4	r2
5	r3
6	r3
7	r1
8	r1
9	r3
10	r3
11	r1
12	r2
13	r2

Q4

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.

Example:



- Although these two factors(Us spending on science, space and technology & Suicides by hanging, strangulation and suffocation) are strongly correlated (99.79%), there's unlikely to have actual causality exists between these two factors. Because suicide rate and science and technology funding are basically two distinct fields, the occurrence of neither one could be applied to explain why the other's changes. The increase of Us spending on science does not contribute to the ascending trend of hanging suicide rate, and vice versa.

- **The US annual GDP growth** has a causative link with **US spending on science and technology**. With a more promising economic growth, the funding poured into the field of science research will be more likely to increase. And with significant amount of spending on technological research, the revolutionary outcomes could contribute back to the country's economic development, therefore creating a virtuous circle.

Q5

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.

- (a) Since $X \sim N(2,6), Y \sim N(-3,2), Z \sim N(0,1)$,

 and the rule of variance:

$$Var(x+y) = Var(x) + Var(y) + 2cov(x,y)$$

 according to what's denoted in the question, all three random variables are independent of each other,

$$\therefore Var(x+y) = Var(x) + Var(y)$$

$$\therefore W \sim N(2-3+0, 6+2+1) = N(-1,9)$$

$$\therefore E(W) = -1, Var(W) = 9$$

- (b) Since $Y \sim N(-3,2)$,

 and the rule of mean:

$$E(cX) = cE(X)$$

 the rule of variance:

$$Var(cX) = c^2 Var(X)$$

$$\therefore Q \sim N(2 \cdot (-3), 2^2 \cdot 2)$$

$$Q \sim N(-6,8)$$

- (c) according to the rules discussed in part (b), we know that:

$$E(cX) = cE(X)$$

$$Var(cX) = c^2 Var(X)$$

$$\therefore (-2X) \sim N((-2) \cdot 2, (-2)^2 \cdot 6)$$

$$\therefore (-2X) \sim N(-4, 24)$$

and the rule of mean:

$$E(X + c) = E(X) + c$$

the rule of variance:

$$Var(X + c) = E[(X + c)^2] - [E(X + c)]^2 = E(X^2) - [E(X)]^2 = Var(X)$$

$$\therefore P = -2X + 4 \sim N(-4 + 4, 24)$$

$$\therefore P = -2X + 4 \sim N(0, 24)$$

- (d) according the rules of mean and variance discussed in all parts above

$$M = a + bX \sim N(2b + a, 6b^2)$$

if $M$is distributed as a standard Normal distribution, then M is a normal distribution with a mean of 0 and a standard deviation of 1

$$\begin{cases} 2b + a = 0 \\ 6b^2 = 1 \end{cases} \tag{1}$$

therefore:

$$\begin{cases} a = -\frac{\sqrt{6}}{3} \\ b = \frac{\sqrt{6}}{6} \end{cases} \tag{2}$$

or

$$\begin{cases} a = \frac{\sqrt{6}}{3} \\ b = -\frac{\sqrt{6}}{6} \end{cases} \tag{3}$$

Q6

6.Do the following

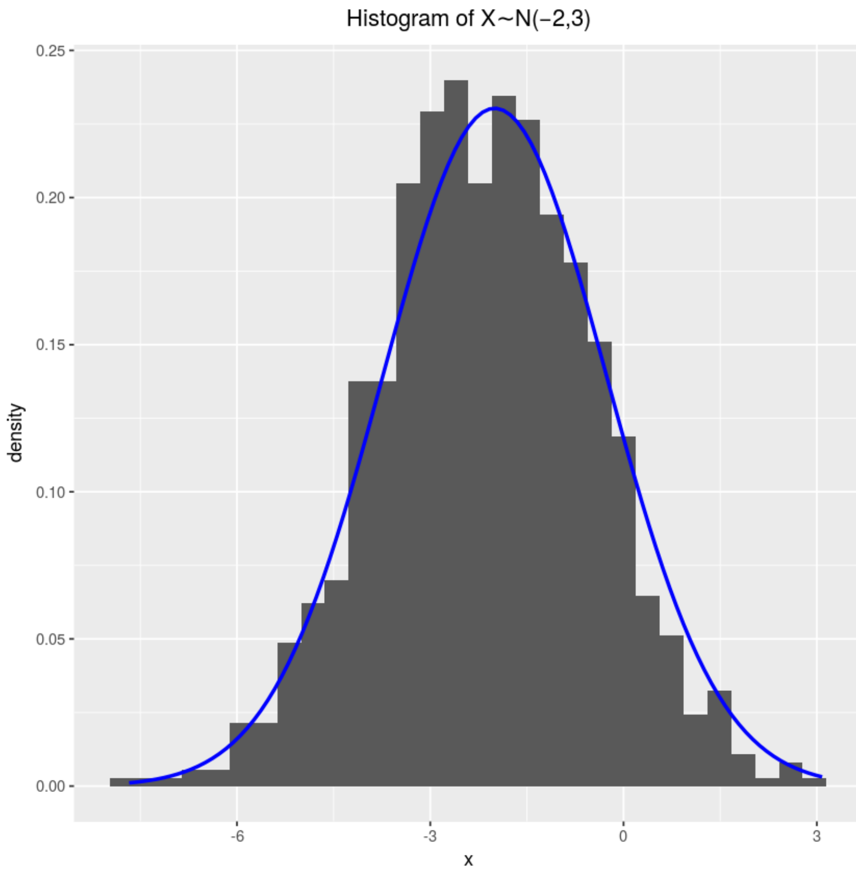
- (a) Use R to simulate 1000 iid random variables X_i with $Xi \sim N(-2, 3)$. Plot a histogram of your simulated values.
- (b) Also simulate 1000 iid random variables Y_i with $Y_i \sim (3, 1)$. Plot a histogram of your simulated values.
- (c) Finally, plot a histogram of Zi, where $Z_i = X_i + Y_i$.
- (d) Is Z_i independent of X_i? Explain your answer.
- (e) Find the sample mean and variance of the Zis you simulated, and compare them with the true, theoretical mean and variance.

- (a)

In [122]:
```
X=rnorm(1000,mean=-2,sd=sqrt(3))
```

In [142]:
```
library(ggplot2)
X_df = data.frame(x = X)
ggplot(X_df, aes(x = x)) +
    geom_histogram(aes(y =..density..)) +
    ggtitle("Histogram of X~N(-2,3)")+
    theme(plot.title = element_text(hjust = 0.5)) +
    stat_function(fun = dnorm, args = list(mean = -2,
                                        sd = sqrt(3)), col='blue', lwd=1)
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

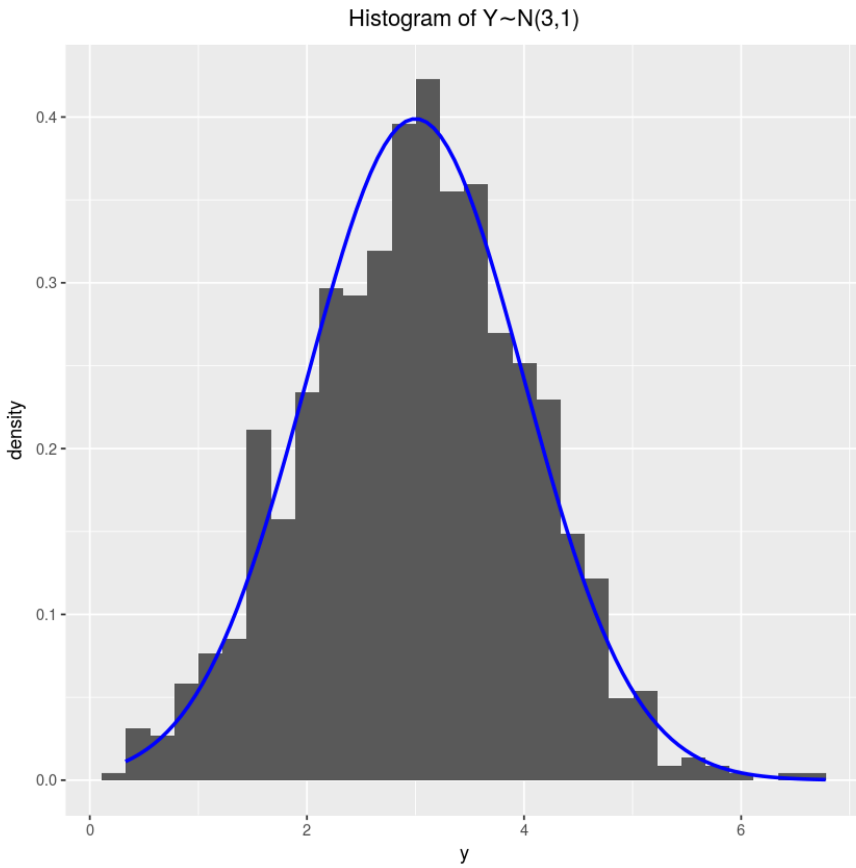


- (b)

```
In [143]: Y = rnorm(1000,mean=3,sd=sqrt(1))
```

```
In [144]: Y_df = data.frame(y = Y)
          ggplot(Y_df, aes(x = y)) +
              geom_histogram(aes(y =..density..)) +
              ggtitle("Histogram of Y~N(3,1)")+
              theme(plot.title = element_text(hjust = 0.5)) +
              stat_function(fun = dnorm, args = list(mean = 3,
                                                      sd = sqrt(1)), col='blue', lwd=1)
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



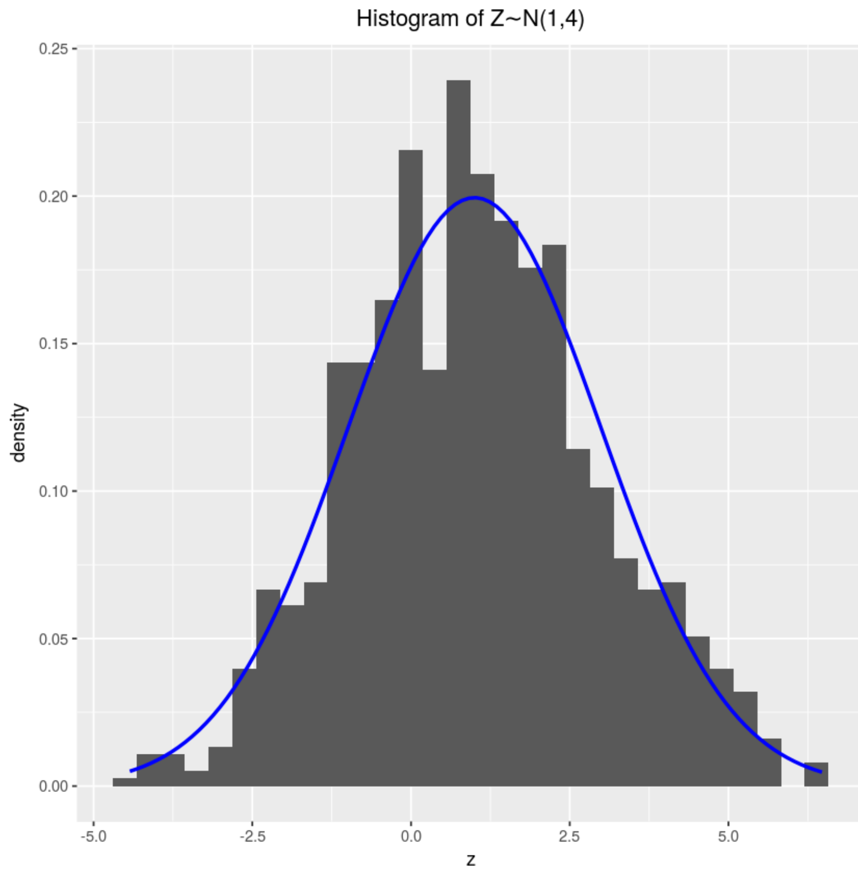
- (c)

Since $Z = X + Y \sim N(-2+3, 3+1)$, thus $Z \sim N(1,4)$

```
In [145]: Z = rnorm(1000,mean=1,sd=sqrt(4))
```

```
In [146]: Z_df = data.frame(z = Z)
          ggplot(Z_df, aes(x = z)) +
              geom_histogram(aes(y =..density..)) +
              ggtitle("Histogram of Z~N(1,4)")+
              theme(plot.title = element_text(hjust = 0.5)) +
              stat_function(fun = dnorm, args = list(mean = 1,
                                                     sd = sqrt(4)), col='blue', lwd=1)
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



- (d) Yes, Z_i is independent of X_i.

Because if two variables are independent of each other, then the covariance should be close to 0. We already know that X and Y are randomly generated and independent, so:

```
In [147]: cov(X, Y)
```

0.0160952090678688
Since X and Y are 1000 simulated numbers, we could consider $cov(X, Y) \approx 0$

```
In [148]: cov(rnorm(10000,mean=-2,sd=sqrt(3)), rnorm(10000,mean=3,sd=sqrt(1)))
```

0.010238213251291

```
In [149]: cov(rnorm(100000,mean=-2,sd=sqrt(3)), rnorm(100000,mean=3,sd=sqrt(1)))
```

-0.00922166054983182

```
In [150]: cov(X, Z)
```

0.0603978617660555
So similarly, $cov(X, Z) \approx 0$, so it's safe to conclude that Z_i is independent of X_i.

- (e)

```
In [151]: Z_sample_mean = mean(Z)
          Z_sample_var = var(Z)
```

```
In [152]: Z_sample_mean
```

1.00402874538632

$$Z \sim N(1,4)$$

$$1.00402874538632 \approx 1$$

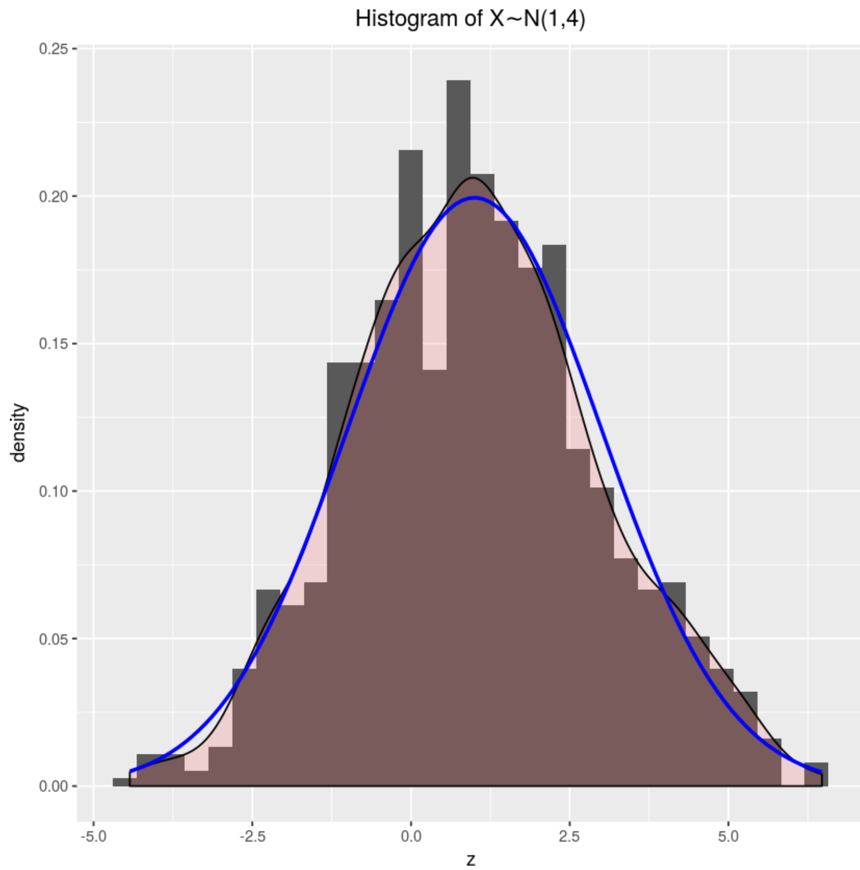
```
In [154]: Z_sample_var
```

3.74829970236857

$$3.74829970236857 \approx 4$$

```
In [155]: ggplot(Z_df, aes(x = z)) +
              geom_histogram(aes(y =..density..)) +
              geom_density(alpha=0.2, fill="#FF6666")+
              ggtitle("Histogram of X~N(1,4)")+
              theme(plot.title = element_text(hjust = 0.5)) +
              stat_function(fun = dnorm, args = list(mean = 1, sd = sqrt(4)),
                            col='blue', lwd=1, type='area')
```

```
Warning message:
“Ignoring unknown parameters: type”`stat_bin()` using `bins = 30`. Pick better value with `binwi
```


Histogram of X~N(1,4)
density
0.25
0.20
0.15
0.10
0.05
0.00
-5.0
-2.5
0.0
2.5
5.0
z