PS 6 R Code

David Zynda

February 23, 2019

Problem 1

Write an R code to implement the following:

(a) Generate 5 i.i.d obs from the uniform random variable on $[-0.5, 0.5]$ and compute the sample average.

```
set.seed(1234)
obs = runif(5, -0.5, 0.5)
mean(obs)
```

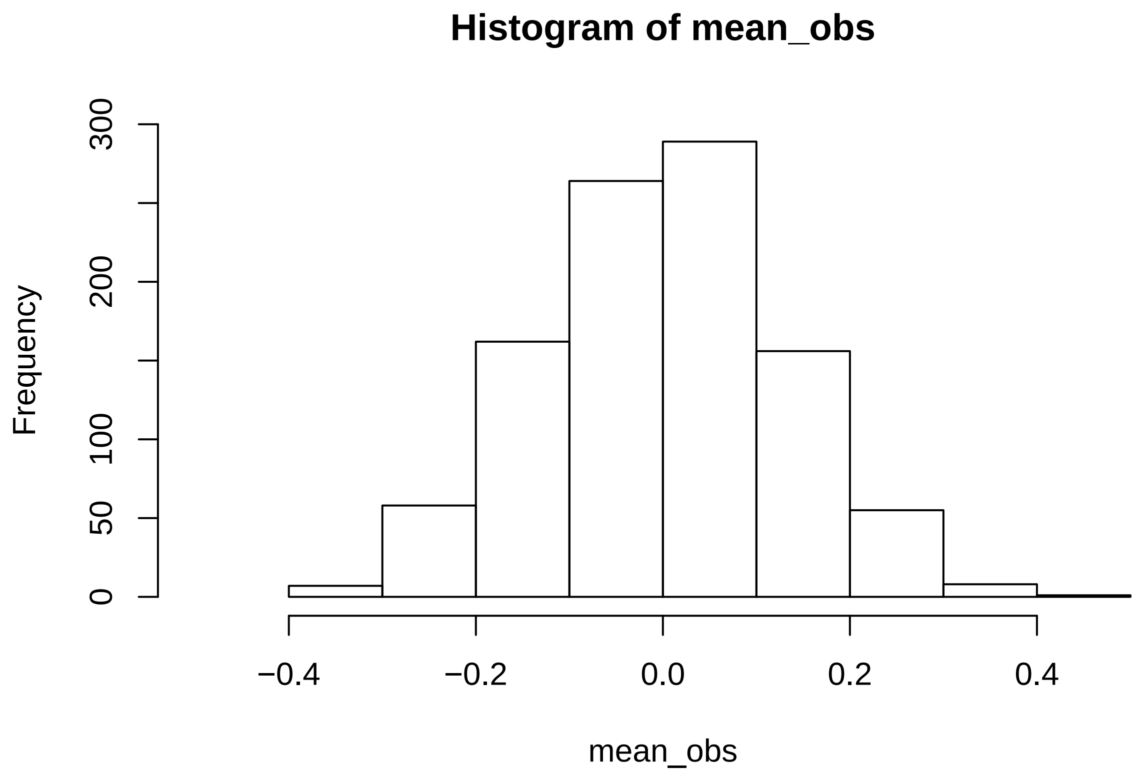
```
## [1] 0.06591447
```

(b) Repeat the above 1000 times, using new observations each time

```
N = 1000
mean_obs = rep(0, N)

for(i in 1:N){
  mean_obs[i] = mean(runif(5, -0.5, 0.5))
}
```

(c) Draw Histogram

```
hist(mean_obs, xlim = c(-0.5, 0.5), ylim = c(0, 300))
```

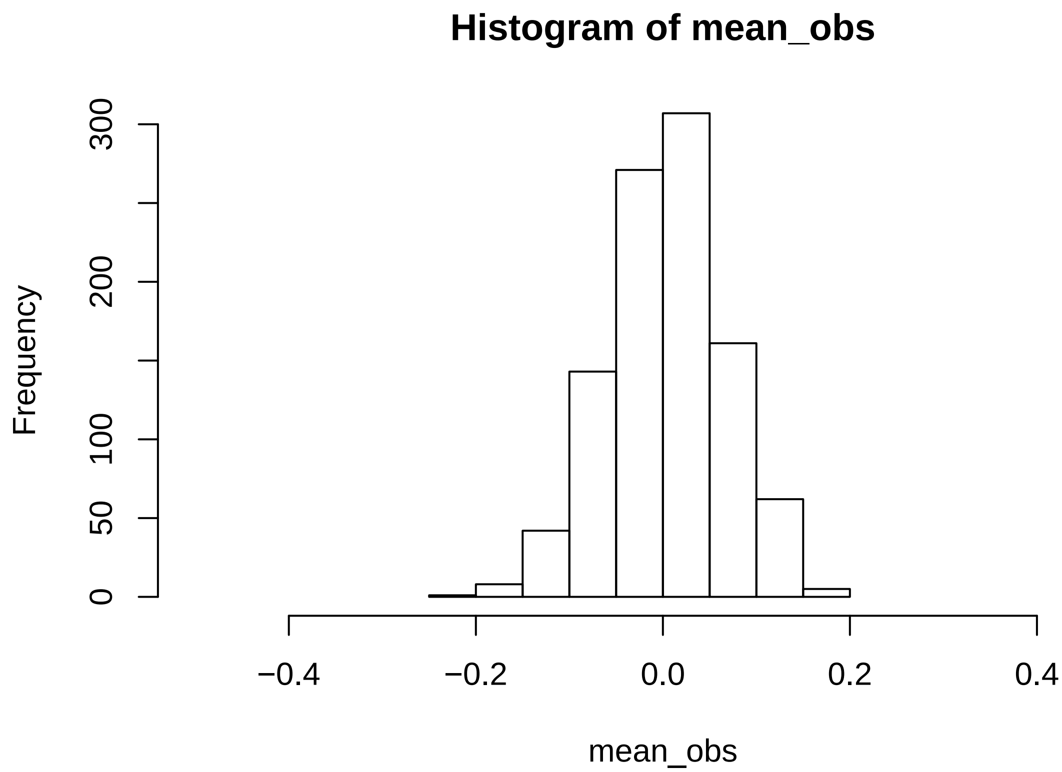


(d) Do parts (a) through (c) now with 20 observations from the same random variable

```
N = 1000
M = 20

obs = runif(M, -0.5, 0.5)
mean(obs)
```

```
## [1] -0.03899555
```

```
mean_obs = rep(0, N)

for(i in 1:N){
  mean_obs[i] = mean(runif(M, -0.5, 0.5))
}

hist(mean_obs, xlim = c(-0.5, 0.5), ylim = c(0, 300))
```

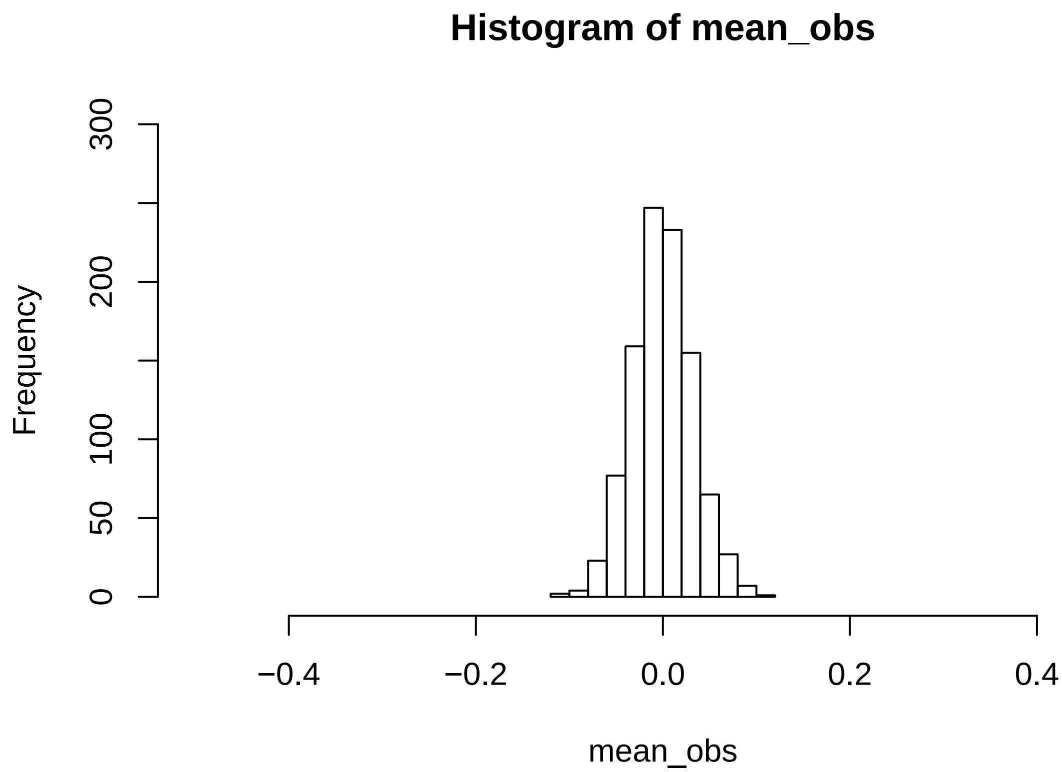


(e) Do it now with 80 obs

```
N = 1000
M = 80

obs = runif(M, -0.5, 0.5)
mean(obs)
```

```
## [1] -0.0784535
```

```
mean_obs = rep(0, N)

for(i in 1:N){
  mean_obs[i] = mean(runif(M, -0.5, 0.5))
}


hist(mean_obs, xlim = c(-0.5, 0.5), ylim = c(0, 300))
```



(f) When you draw the histograms to scale, what do you see? Explain what you observe using the LLN.

The distribution tightens and begins to converge as sample size increases.

Problem 2

Write out code to implement the follows:

(a) Generate 5 i.i.d. observations from the uniform random variable on [-0.5, 0.5] and compute the sample average.

```
dat_gen <- function(N, M){
  # Let N be mean sample size, M be obs
  obs = runif(M, -0.5, 0.5)

  mean_obs = rep(0, N)

  for(i in 1:N){
    mean_obs[i] = mean(runif(M, -0.5, 0.5))
  }
  return(mean_obs)
}


dat_gen(1, 5)
```

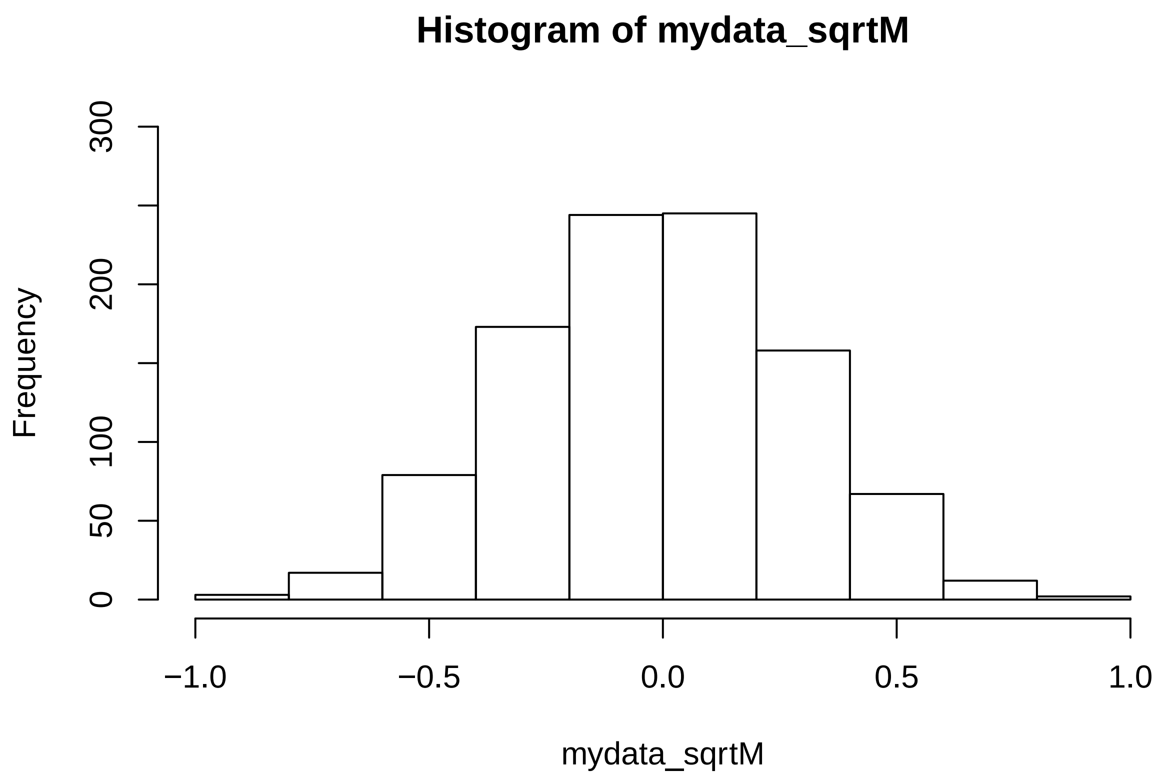
```
## [1] -0.3103529
```

(b) Repeat (a) 1000 times, using new observations each time

```
mydata <- dat_gen(1000, 5)
```

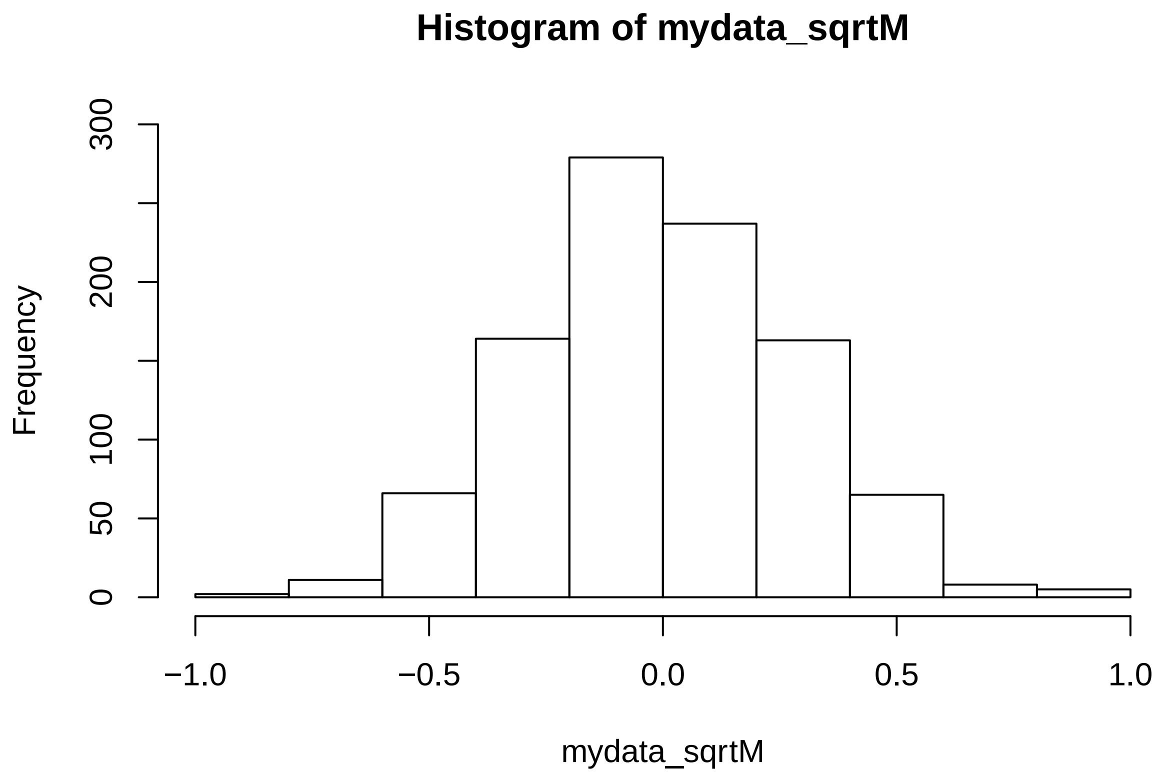
(c) Historgram of 1000 averages times sqrt of 5

```
mydata_sqrtM = mydata * sqrt(5)
hist(mydata_sqrtM, xlim = c(-1, 1), ylim = c(0, 300))
```



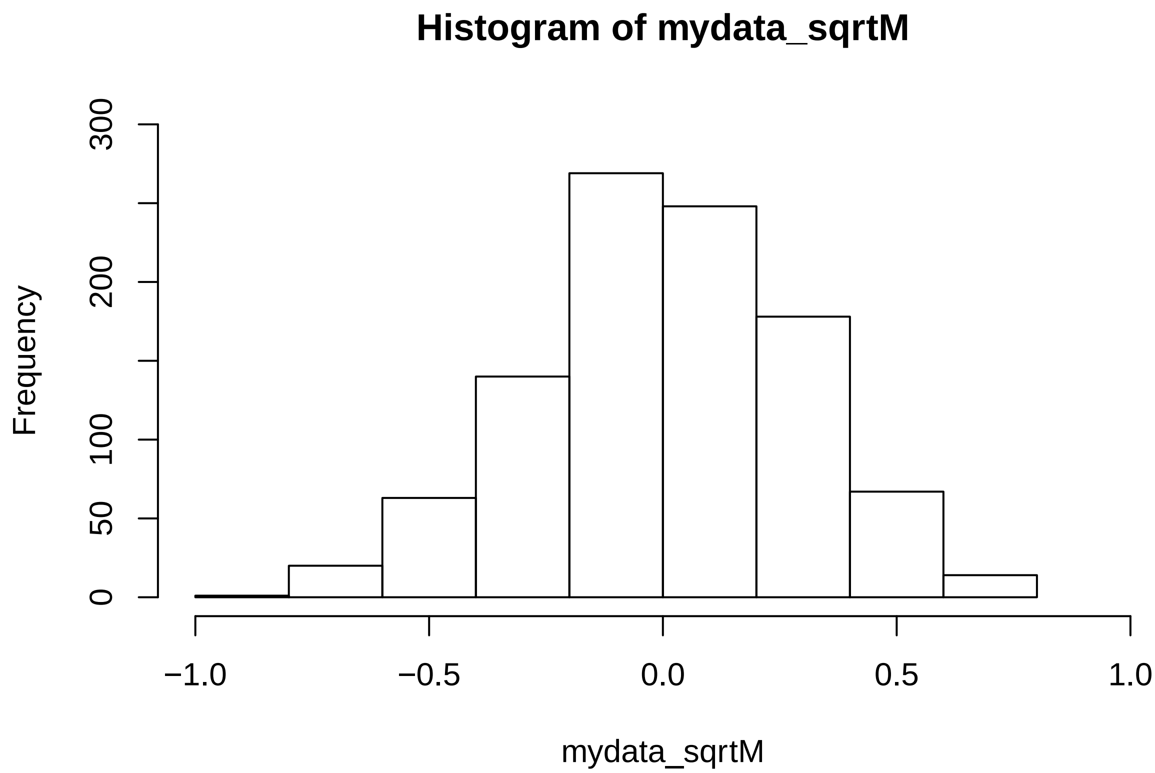
(d) Do (a) - (c) using 20 obs and multiply by sqrt 20

```
mydata = dat_gen(1000, 20)
mydata_sqrtM = mydata * sqrt(20)
hist(mydata_sqrtM, xlim = c(-1, 1), ylim = c(0, 300))
```



(e) Do (a) - (c) using 20 obs and multiply by sqrt 20

```
mydata = dat_gen(1000, 80)
mydata_sqrtM = mydata * sqrt(80)
hist(mydata_sqrtM, xlim = c(-1, 1), ylim = c(0, 300))
```



(f) What do you see?

It appears that the images are becoming more symmetric.

Problem 5

(a) Generate 100 iid data from the model in the statement and regress y_i on the constant term x_i and z_i to verify that there is the attenuation bias. State a way you can simulate data to verify the direction of inconsistency proved in class

```
library(MASS)

# Generate data
r = 0
varcov = matrix(c(1,r,0, r,1,0, 0,0,1), nrow = 3, ncol = 3)
mus = matrix(c(rep(0,3)), nrow = 3)

regressors = mvrnorm(100, mus, varcov)
x_star = regressors[,1]
z_i = regressors[,2]

y = rep(0,100)
y = 1 + x_star + z_i + regressors[,3]

v = rnorm(100, 0, 1)
x_i = regressors[,1] + v
```

```r
# Regress yi on constant term, xi (not xi*) and zi

my_lm = lm(y ~ x_i + z_i)
summary(my_lm)
```

```
##
## Call:
## lm(formula = y ~ x_i + z_i)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2235 -0.9307 -0.0502  0.8569  2.5457
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.03055    0.12292   8.384 4.08e-13 ***
## x_i          0.50204    0.09203   5.455 3.74e-07 ***
## z_i          0.95606    0.12700   7.528 2.66e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.213 on 97 degrees of freedom
## Multiple R-squared:  0.4594, Adjusted R-squared:  0.4482
## F-statistic: 41.21 on 2 and 97 DF,  p-value: 1.11e-13
```

As can be seen from the summary generated above, x_i is underestimated - negatively biased.

(b) What will happen to size of inconsistency if σ_v^2 increases? Verify with $\sigma_v^2 = 2$

```r
# Generate data
r = 0
varcov = matrix(c(1,r,0, r,1,0, 0,0,1), nrow = 3, ncol = 3)
mus = matrix(c(rep(0,3)), nrow = 3)

regressors = mvrnorm(100, mus, varcov)
x_star = regressors[,1]
z_i = regressors[,2]

y = rep(0,100)
y = 1 + x_star + z_i + regressors[,3]

v = rnorm(100, 0, sqrt(2))
x_i = regressors[,1] + v


# Regress yi on constant term, xi (not xi*) and zi

my_lm = lm(y ~ x_i + z_i)
summary(my_lm)
```

```
##
## Call:
## lm(formula = y ~ x_i + z_i)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1736 -0.8957  0.0325  0.9092  4.5327
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0275     0.1421   7.231 1.11e-10 ***
## x_i           0.3606     0.0853   4.228 5.35e-05 ***
## z_i           0.9566     0.1345   7.111 1.97e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.414 on 97 degrees of freedom
## Multiple R-squared:  0.3955, Adjusted R-squared:  0.3831
## F-statistic: 31.73 on 2 and 97 DF,  p-value: 2.496e-11
```

As variance increases, the slope becomes more negatively biased.

(c) What will happen if ρ increases? Verify this by simulating with $\rho = 0.5$

```r
r = 0.5
varcov = matrix(c(1,r,0, r,1,0, 0,0,1), nrow = 3, ncol = 3)
mus = matrix(c(rep(0,3)), nrow = 3)

regressors = mvrnorm(100, mus, varcov)
x_star = regressors[,1]
z_i = regressors[,2]

y = rep(0,100)
y = 1 + x_star + z_i + regressors[,3]

v = rnorm(100, 0, 1)
x_i = regressors[,1] + v


# Regress yi on constant term, xi (not xi*) and zi

my_lm = lm(y ~ x_i + z_i)
summary(my_lm)
```

```
##
## Call:
## lm(formula = y ~ x_i + z_i)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8344 -0.5573  0.1490  0.6771  2.2944
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.14893    0.10711  10.726  < 2e-16 ***
## x_i          0.42702    0.08607   4.961 2.98e-06 ***
## z_i          1.21911    0.12470   9.776 4.08e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.061 on 97 degrees of freedom
## Multiple R-squared:  0.5976, Adjusted R-squared:  0.5893
## F-statistic: 72.02 on 2 and 97 DF,  p-value: < 2.2e-16
```

It becomes more negatively biased