

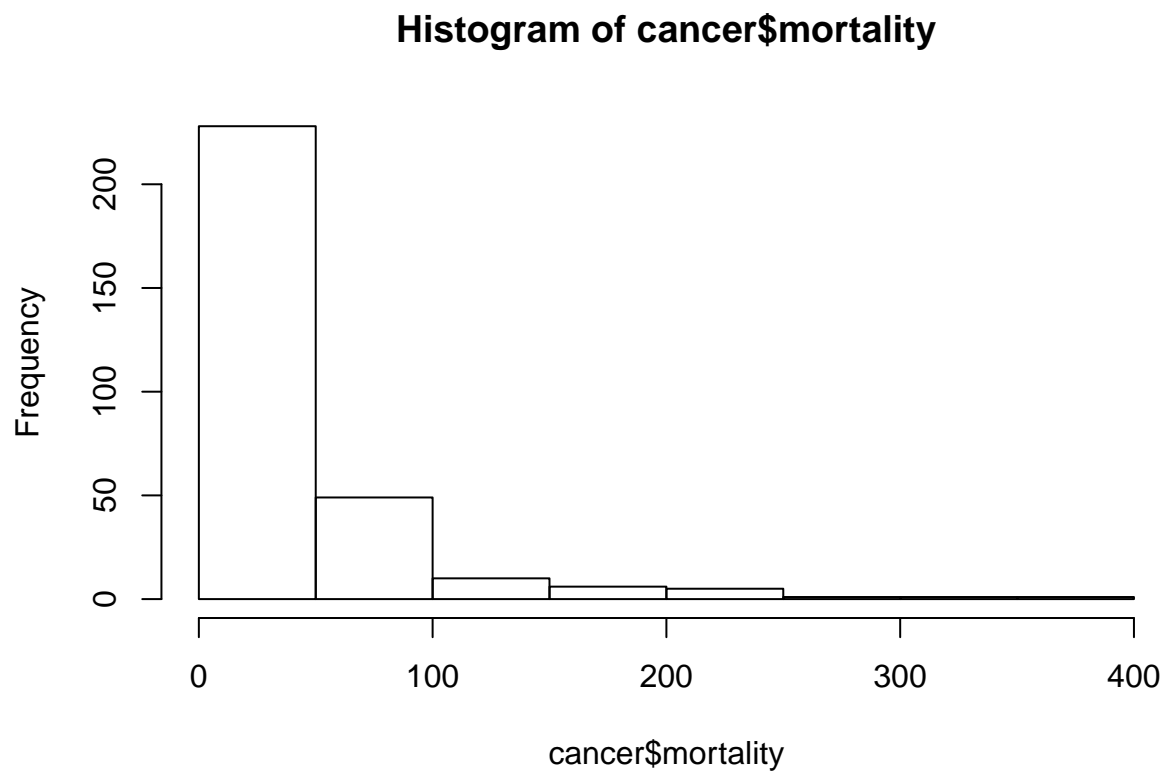
# HW 1 #65

*Arjun Panda*

*October 9, 2016*

Parts a, b, c, d, f

```
# a.  
hist(cancer$mortality, freq = T)
```



```
# b.  
mean(cancer$mortality)
```

```
## [1] 39.85714
```

```
sum(cancer$mortality)
```

```
## [1] 11997
```

```
(length(cancer$mortality)-1)/length(cancer$mortality)*var(cancer$mortality) #population var
```

```
## [1] 2590.103
```

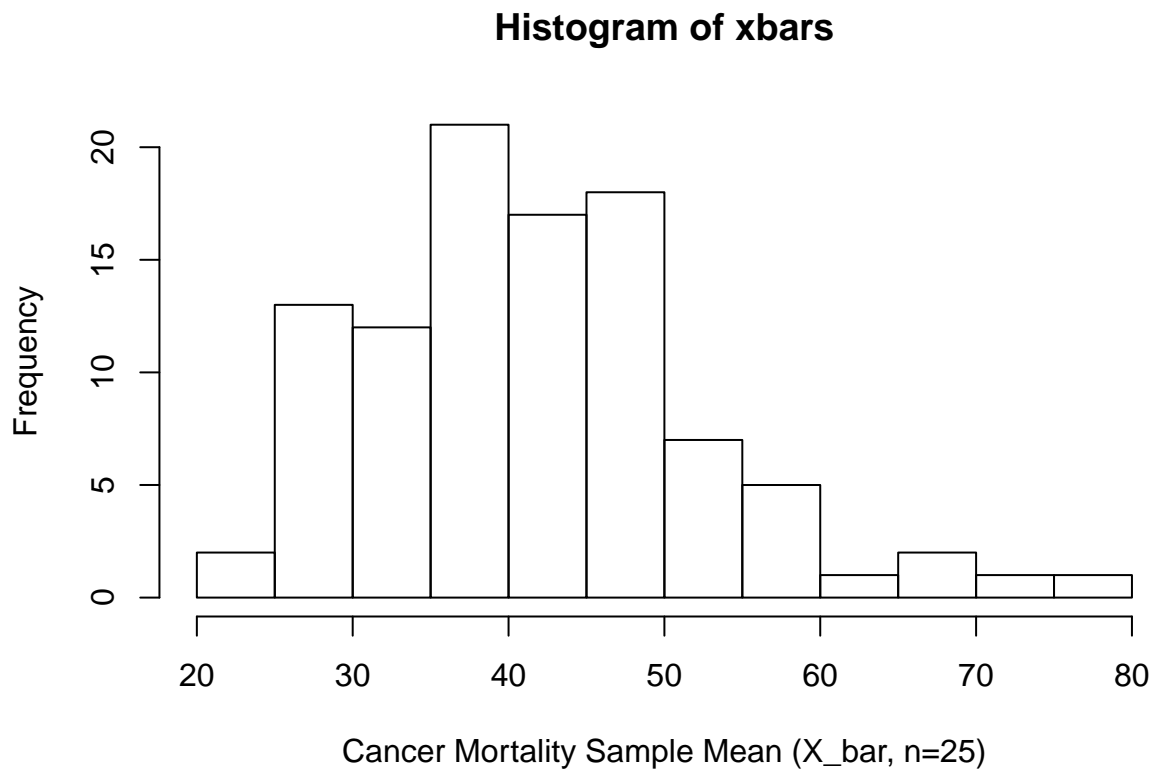
```
(length(cancer$mortality)-1)/length(cancer$mortality)*sd(cancer$mortality) #pop SD
```

```
## [1] 50.80844
```

```
# Note: Since the variance function in R uses the n-1 in the denominator, I modify  
#the result to reflect the actual population variance, not the estimator.
```

```
# c. Simulate sampling distribution of x_bar with n=25
```

```
xbars <- numeric(100)  
for(i in 1:100){  
  xbars[i] <- mean(sample(cancer$mortality, 25, replace = F))  
}  
hist(xbars, freq = T, xlab = "Cancer Mortality Sample Mean (X_bar, n=25)")
```



```
# d.  
x_bar <- mean(sample(cancer$mortality, 25, replace = F)) #n=25  
x_bar #mean cancer mortality estimator
```

```
## [1] 46.6
```

```
total <- x_bar*301 #total cancer mortality estimator
total
```

```
## [1] 14026.6
```

```
# e.
popvar <- (1-(1/301))*var(cancer$mortality)
sqrt(popvar) # pop SD
```

```
## [1] 50.89305
```

```
# f.
var_xbar <- (var(cancer$mortality)/25)*(1-(25/301))
lower_lim <- x_bar - (1.96*sqrt(var_xbar))
upper_lim <- x_bar + (1.96*sqrt(var_xbar))
c(lower_lim, upper_lim) # 95% CI for X_bar
```

```
## [1] 27.46456 65.73544
```

```
#I am 95% confident that the average cancer mortality per county
#in the three states is contained within the values above
```

```
var_total <- (301^2)*var_xbar
lower_lim <- total - 1.96*sqrt(var_total)
upper_lim <- total + 1.96*sqrt(var_total)
c(lower_lim, upper_lim) # 95% CI for T
```

```
## [1] 8266.832 19786.368
```

```
#I am 95% confident that the total cancer for all counties
#in the three states is contained within the values above
```

```
#Conclusion: Yes, the intervals cover the population values.
```

## Part g. (repeat for n=100)

```
# d.
x_bar <- mean(sample(cancer$mortality, 100, replace = F)) #n=100
x_bar #mean cancer mortality estimator
```

```
## [1] 34.18
```

```
total <- x_bar*301 #total cancer mortality estimator
total
```

```
## [1] 10288.18
```

```
# e.
popvar <- (1-(1/301))*var(cancer$mortality)
sqrt(popvar) # pop SD
```

```
## [1] 50.89305
```

```
# f.
var_xbar <- (var(cancer$mortality)/100)*(1-(100/301))
lower_lim <- x_bar - (1.96*sqrt(var_xbar))
upper_lim <- x_bar + (1.96*sqrt(var_xbar))
c(lower_lim, upper_lim) # 95% CI for X_bar
```

```
## [1] 26.01508 42.34492
```

```
#I am 95% confident that the average cancer mortality per county
#in the three states is contained within the values above
```

```
var_total <- (301^2)*var_xbar
lower_lim <- total - 1.96*sqrt(var_total)
upper_lim <- total + 1.96*sqrt(var_total)
c(lower_lim, upper_lim) # 95% CI for T
```

```
## [1] 7830.539 12745.821
```

```
#I am 95% confident that the total cancer mortality for all counties
#in the three states is contained within the values above
```

```
#Conclusion: Yes, the intervals cover the population values.
```

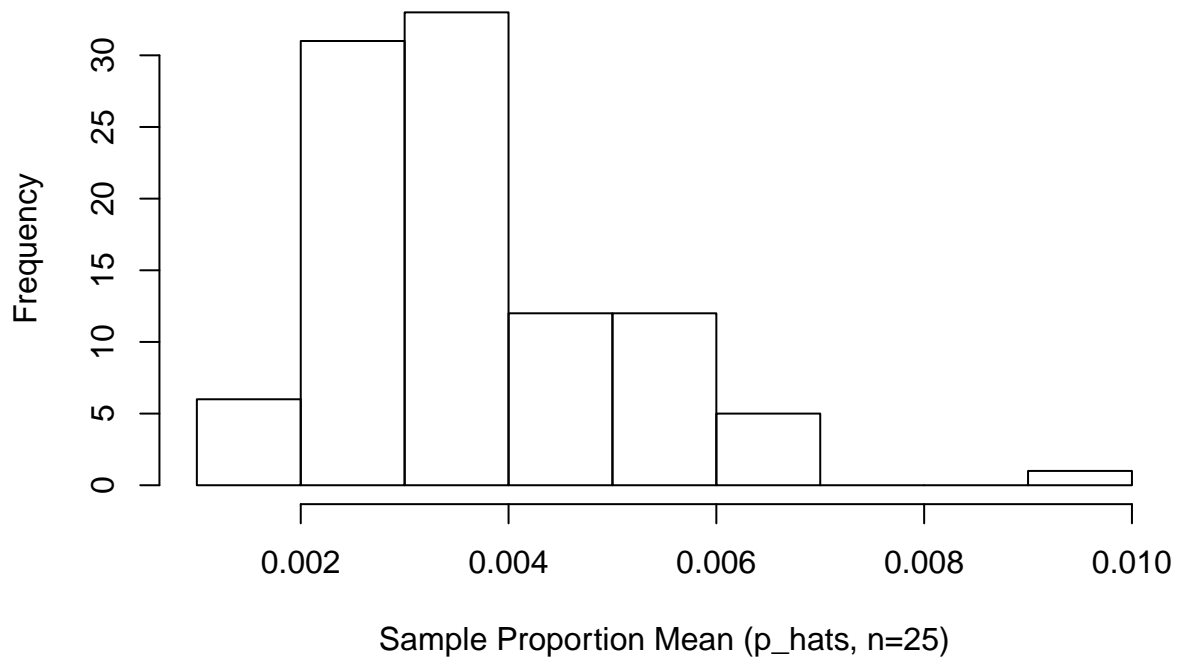
## Part h.

This is effective because the ratio estimates achieve comparable precision with a much smaller

## Parts i, j, k

```
# i.
cancer$prop <- cancer$mortality/cancer$pop
ratios <- numeric(100)
for(i in 1:100){
  y_bar <- mean(sample(cancer$mortality, 25, replace = F))
  x_bar <- mean(sample(cancer$pop, 25, replace = F))
  ratios[i] <- y_bar/x_bar
}
hist(ratios, freq = T, xlab = "Sample Proportion Mean (p_hats, n=25)")
```

## Histogram of ratios



*#Compared to part c, the range of the distribution is more constrained.*

```
# j.
y_sample <- sample(cancer$mortality, 25, replace = F)
x_sample <- sample(cancer$pop, 25, replace = F)
y_bar <- mean(y_sample)
x_bar <- mean(x_sample)
R <- y_bar/x_bar #ratio estimator

y_bar_R <- R*mean(cancer$pop)
y_bar_R
```

```
## [1] 62.83246
```

```
T_R <- R*sum(cancer$pop)
T_R
```

```
## [1] 18912.57
```

*# These estimates are comparable to the values formulated  
# in part d.*

*# k. confidence intervals*

```

var_ybar_R <- (1/25)*(1-((24)/300))*((R^2)*var(x_sample)) + var(y_sample)
               - 2*R*cov(x_sample,y_sample))
lower <- y_bar_R - (1.96*sqrt(y_bar_R))
upper <- y_bar_R + (1.96*sqrt(y_bar_R))
c(lower,upper) # 95% CI for Ybar_R

```

```
## [1] 47.29614 78.36877
```

*# I am 95% confident that the average number of cancer mortalities in each county  
# falls within the above values.*

```

var_T_R <- (301^2)*var_ybar_R
lower <- T_R - (1.96*sqrt(var_T_R))
upper <- T_R + (1.96*sqrt(var_T_R))
c(lower,upper) # 95% CI for T_R

```

```
## [1] 10949.25 26875.89
```

*#I am 95% confident that the total cancer mortality for all counties  
#in the three states is between the above values.*

## Parts l, m, n

```

# l. four strata
cancer$group <- ntile(cancer$pop, n=4)
n1 <- mean(sample(cancer$mortality[cancer$group==1], size = 6))
n2 <- mean(sample(cancer$mortality[cancer$group==2], size = 6))
n3 <- mean(sample(cancer$mortality[cancer$group==3], size = 6))
n4 <- mean(sample(cancer$mortality[cancer$group==4], size = 6))
W <- 75/301
xbar_s <- W*(n1+n2+n3+n4)
xbar_s #pop mean mortality estimate

```

```
## [1] 42.9402
```

```

T_s <- 301*xbar_s
T_s #pop total mortality estimate

```

```
## [1] 12925
```

*# m.*

*#sampling fraction for proportional allocation is simply 1/4th for each strata,  
#since each strata is equally sized.*

*#sampling fraction for optimal allocation:*

```

denom <-
  W*var(cancer$mortality[cancer$group==1]) + W*var(cancer$mortality[cancer$group==2]) +

```

```

W*var(cancer$mortality[cancer$group==3]) + W*var(cancer$mortality[cancer$group==4])

n_1 <- var(cancer$mortality[cancer$group==1])/denom
n_2 <- var(cancer$mortality[cancer$group==2])/denom
n_3 <- var(cancer$mortality[cancer$group==3])/denom
n_4 <- var(cancer$mortality[cancer$group==4])/denom

c(W*n_1, W*n_2, W*n_3, W*n_4) #optimal allocation sampling fraction

```

```
## [1] 0.003474048 0.009457559 0.032221686 0.954846707
```

```
# Comparing variances.
```

```

sig_bar <- W*(sqrt(n_1)) + W*(sqrt(n_2)) + W*(sqrt(n_3)) + W*(sqrt(n_4))

VarXsp_by_VarXso <-
  1 + (W*((sqrt(n_1)-sig_bar)^2 + (sqrt(n_2)-sig_bar)^2 +
    (sqrt(n_3)-sig_bar)^2 + (sqrt(n_4)-sig_bar)^2)
    / ((W*(sqrt(n_1) + sqrt(n_2) + sqrt(n_3) + sqrt(n_4)))^2)
  )

VarXsp_by_VarXso

```

```
## [1] 2.325144
```

Here we conclude that the variance of the mean proportional allocation is 232.5144143 percent of the variance under optimal allocation. Therefore, optimal allocation is clearly preferred.