PROBLEM SET #2: SURVEY SAMPLING (Ch. 7)

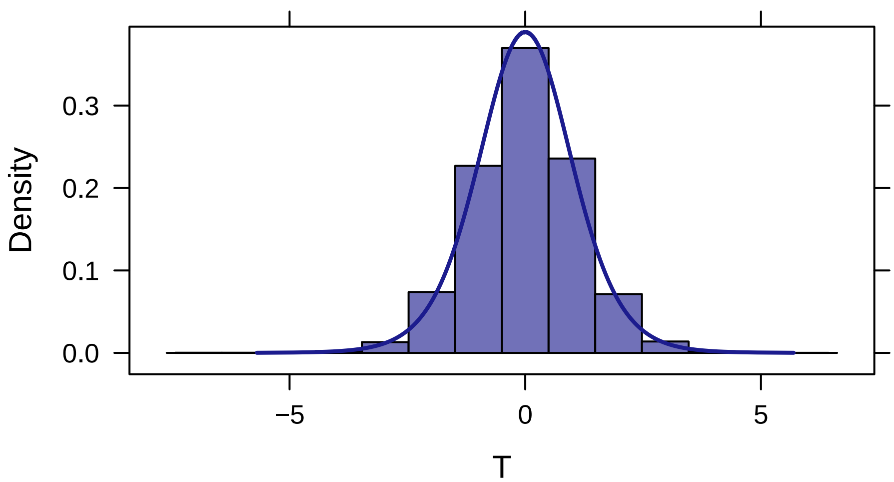
Group 6: Tasheena, Tim, Meredith

02/07/16

Proposition A

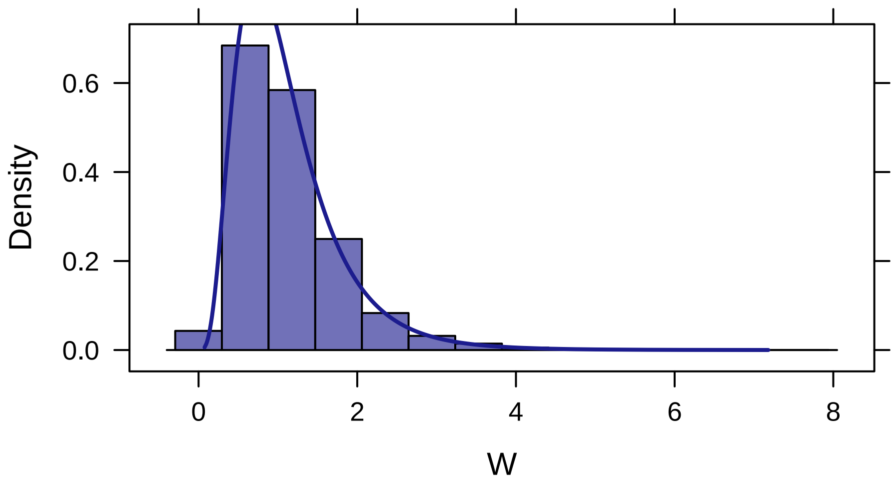
```
numsim <- 10000

m <-10
Z <- rnorm(10000, 0, 1)
U <- rchisq(10000, m)
T <- Z/sqrt(U/m)
histogram(T)
plotDist("t", m, add=TRUE)
```



Proposition B

```
m <- 10
n <- 20
U <- rchisq(10000, m)
V <- rchisq(10000, n)
W <- (U/m)/(V/n)
histogram(W)
plotDist("f", df1=10, df2=20, add=TRUE) #why does df not show up with variable names
```



7.1

- Consider a population consisting of five values: 1, 2, 2, 4, and 8. Find the population mean and variance.
- Generate all possible samples of size 2. Calculate the mean and variance of the sampling distribution.
- Compare results to Theorems A and B in Section 7.3.1.

Analytical Solution

$$\mu = \frac{1+2+2+4+8}{5} = \frac{17}{5}$$
$$\sigma^2 = \frac{(1-\frac{17}{5})^2 + (2-\frac{17}{5})^2 + (2-\frac{17}{5})^2 + (4-\frac{17}{5})^2 + (8-\frac{17}{5})^2}{5}$$
$$\sigma^2 = 6.24$$
$$\sigma = 2.498$$

$$All\ possible\ combinations\ of\ sample\ size\ 2:$$
$$(1,2),(1,2),(1,4),(1,8),(2,2),(2,4),(2,8),(2,4),(2,8),(4,8)$$
$$p(1.5) = 0.2,\ p(2) = 0.1,\ p(2.5) = 0.1,\ p(3) = 0.2,\ p(4.5) = 0.1,\ p(5) = 0.2,\ p(6) = 0.1$$
$$E(\bar{X}) = 0.2(1.5) + 0.1(2) + ... + 0.2(5) + 0.1(6) = 3.4$$
$$E(\bar{X^2}) = 0.2(1.5^2) + 0.1(2^2) + ... + 0.2(5^2) + 0.1(6^2) = 13.9$$
$$Var(\bar{X}) = 13.9 - 3.4^2 = 2.34$$

The distribution parameters could have been calculated by calculating every possible combination or by using the simplified theorems. They are Theorem A and Theorem B in 7.3.1. See calculation below.

$$E(\bar{X}) = \mu = \frac{17}{5}$$
$$Var(\bar{X}) = \frac{6.24}{2}(1 - \frac{2-1}{5-1}) = 2.34$$

Empirical Solution

- Since R defaults to the sample variance, we created a function that fixes the R default and, thus, allows us to calculate the population variance.

```
popvar = function(v) {
    n = length(v)
    correction = (n-1)/n
    popvariance = var(v) * (correction)
    return(popvariance)
}
```

```
x <- c(1,2,2,4,8)
mean(x) #population mean
```

```
## [1] 3.4
```

```
popvar(x) #population variance
```

```
## [1] 6.24
```

```
sampDist <- combn(x, m=2, simplify = TRUE); sampDist
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]    1    1    1    1    2    2    2    2    2     4
## [2,]    2    2    4    8    2    4    8    4    8     8
```

```
sampDist <- cbind(sampDist[1,], sampDist[2,])
frame <- data.frame(sampDist)

sampmean <- mean(~ (X1+X2)/2, data=frame, format="proportion") ;sampmean #sample mean
```

```
## [1] 3.4
```

```
sampvar <- popvar((frame$X1+frame$X2) / 2) ;sampvar #sample variance
```

```
## [1] 2.34
```

7.2

- Suppose that a sample of size n = 2 is drawn from the same population as 7.1. For each sample, record the proportion of sample values that are greater than 3.
- Find the sampling distribution of this statistic by listing all possible samples.
- Find the mean and variance of this distribution.

```r
sampDist2 <- combn(x, m=2, simplify = TRUE)
extraCol <- matrix(, nrow = 10, ncol = 1)
sampDist2 <- cbind(sampDist2[1,], sampDist2[2,], extraCol)
frame <- data.frame(sampDist2)

frame$X3 <- ifelse((frame$X1<3 & frame$X2 < 3), 0, frame$X3)
frame$X3 <- ifelse((frame$X1<3 & frame$X2 > 3), 0.5, frame$X3)
frame$X3 <- ifelse((frame$X1>3 & frame$X2 > 3), 1, frame$X3)

samp2mean <- mean(frame$X3) ;samp2mean
```

```
## [1] 0.4
```

```r
samp2var <- popvar(frame$X3) ;samp2var
```

```
## [1] 0.09
```

```r
frame
```

```
##    X1 X2  X3
## 1   1  2 0.0
## 2   1  2 0.0
## 3   1  4 0.5
## 4   1  8 0.5
## 5   2  2 0.0
## 6   2  4 0.5
## 7   2  8 0.5
## 8   2  4 0.5
## 9   2  8 0.5
## 10  4  8 1.0
```

Analytical Solution

Using the table created above, we can calculate the expected value and variance analytically.

$$E(\hat{p}) = 0.3(0) + 0.6(0.5) + 0.1(1) = 0.40$$
$$E(\hat{p^2}) = 0.3(0^2) + 0.6(0.5^2) + 0.1(1^2) = 0.25$$
$$Var(\hat{p}) = 0.25 - 0.40^2 = 0.09$$

7.67. Families Dataset.

```r
families <- read_csv("http://www.amherst.edu/~nhorton/rice/chapter07/families.csv")
```

67a

- Take a SRS of 500 families. Estimate the following population parameters, calculuate estimated standard errors, and form 95% CIs of some demographic information (see table).
- Do the preceding parameters above with 5 different random samples of same sample size (n=500) and compare results.

67ahidden

```
displayres
```

% latex table generated in R 3.2.2 by xtable 1.8-0 package % Mon Feb 8 21:19:40 2016

	Mean	StandardError	LowerBound	UpperBound
Prop of Fem-Headed Families	0.22	0.02	0.18	0.26
Avg Num of Children	0.91	0.05	0.81	1.01
Prop of Low HS Headed Families	0.21	0.02	0.12	0.25
Avg Familiy Income	41415.97	1556.15	38359.00	44473.00

Mean from Samples

```
displayres1
```

	FemaleHead	ChildrenPerFamily	NoHSDiploma	FamilyIncome
Sample 1	0.19	0.92	0.23	40228.04
Sample 2	0.19	0.99	0.21	42575.25
Sample 3	0.18	0.90	0.20	43434.03
Sample 4	0.17	0.91	0.21	41316.05
Sample 5	0.19	0.85	0.25	39914.46

Standard Error from Samples

```
displayres2
```

	FemaleHead	ChildrenPerFamily	NoHSDiploma	FamilyIncome
Sample 1	0.02	0.05	0.02	1387.09
Sample 2	0.02	0.05	0.02	1404.06
Sample 3	0.02	0.05	0.02	1505.76
Sample 4	0.02	0.05	0.02	1310.62
Sample 5	0.02	0.05	0.02	1454.46

Lower Bound from Samples

```
displayres3
```

	FemaleHead	ChildrenPerFamily	NoHSDiploma	FamilyIncome
Sample 1	0.17	0.82	0.19	37503.00
Sample 2	0.15	0.88	0.18	39817.00
Sample 3	0.15	0.80	0.16	40476.00
Sample 4	0.13	0.82	0.17	38741.00
Sample 5	0.16	0.75	0.21	37057.00

Upper Bound from Samples

```
displayres4
```

	FemaleHead	ChildrenPerFamily	NoHSDiploma	FamilyIncome
Sample 1	0.24	1.02	0.27	42953.00
Sample 2	0.23	1.09	0.25	45334.00
Sample 3	0.22	1.00	0.24	46392.00
Sample 4	0.20	1.01	0.25	43891.00
Sample 5	0.23	0.94	0.29	42772.00

67(bi)

- QUESTION: Take 100 samples of size 400 and see the following calculations listed in the table.

```
set.seed(1)
sampSize<-400
numsim <- 100
samp100 <- matrix(data=NA , nrow = numsim, ncol = 4)
confintmeanYU = 0; confintmeanYL = 0 ; meanY = 0 ;sdY = 0

   for(i in 1:numsim)
   {
    samp2<-sample(families,sampSize,replace=F)
    meanY[i] <- mean(samp2$INCOME)
    sdY[i] <- sd(samp2$INCOME)/sqrt(sampSize)
    confintmeanYL[i]<- meanY[i] - 1.96*sdY[i]
    confintmeanYU[i]<- meanY[i] + 1.96*sdY[i]
    samp100[i,1] <- (meanY[i])
    samp100[i,2] <- (sdY[i])
    samp100[i,3] <- confintmeanYL[i]
    samp100[i,4] <- confintmeanYU[i]
   }

samp100size400 <- data.frame(samp100)
```

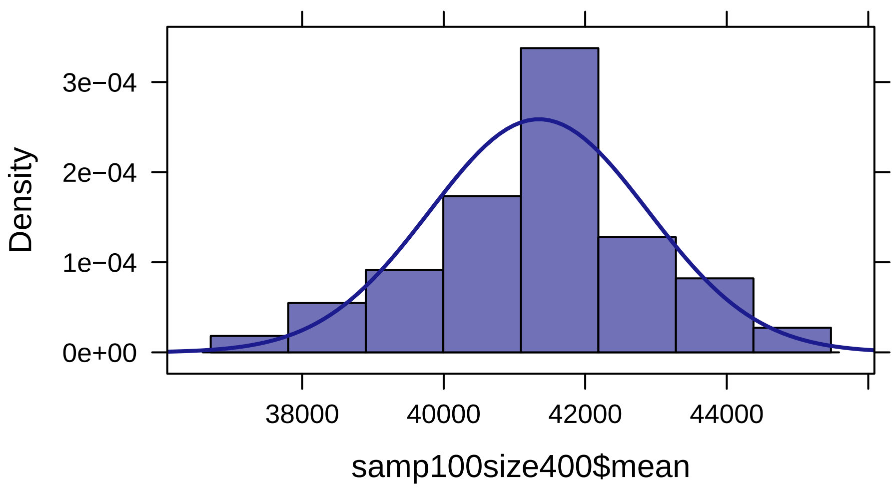
```r
colnames(samp100size400) <- c("mean", "sd", "lower_bound", "upper_bound")
head(samp100size400)
```

```
##       mean       sd lower_bound upper_bound
## 1 40465.47 1542.743    37441.69    43489.24
## 2 40533.11 1477.029    37638.13    43428.08
## 3 41854.47 1487.133    38939.69    44769.26
## 4 38570.22 1401.551    35823.18    41317.26
## 5 43129.64 1713.300    39771.57    46487.71
## 6 43310.78 1759.055    39863.03    46758.53
```

b(ii & iii)

- QUESTION: Make a histogram of the average and standard deviations of the 100 estimates.

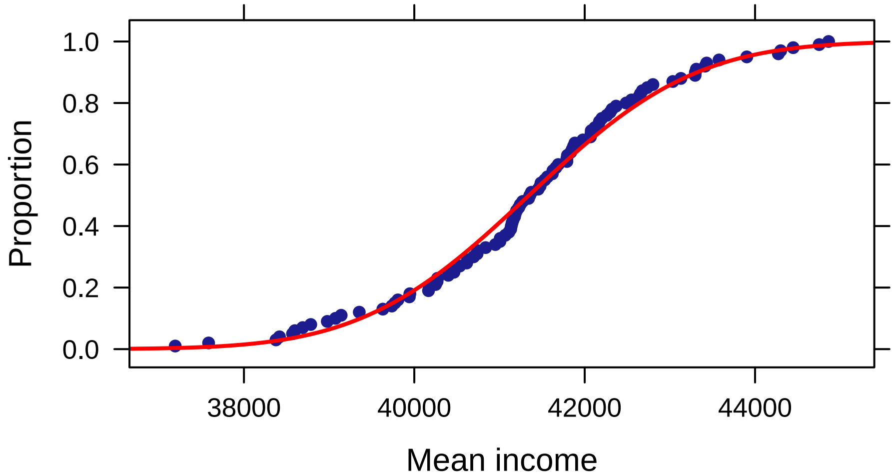
```r
meanInc <- mean(samp100size400$mean)
sdInc <- sd(samp100size400$mean)
histogram(samp100size400$mean, density=TRUE)
```



b(iv)

- QUESTION: Plot the empirical cumulative distribution function (see Section 10.2). Also, superimpose the normal cdf.

```r
xyplot(ecdf(samp100size400$mean)(knots(ecdf(samp100size400$mean)))~ knots(ecdf(samp100size400$mean)), xl
plotDist('norm' , mean=meanInc, sd= sdInc, col = "red", add=TRUE, kind='cdf')
```

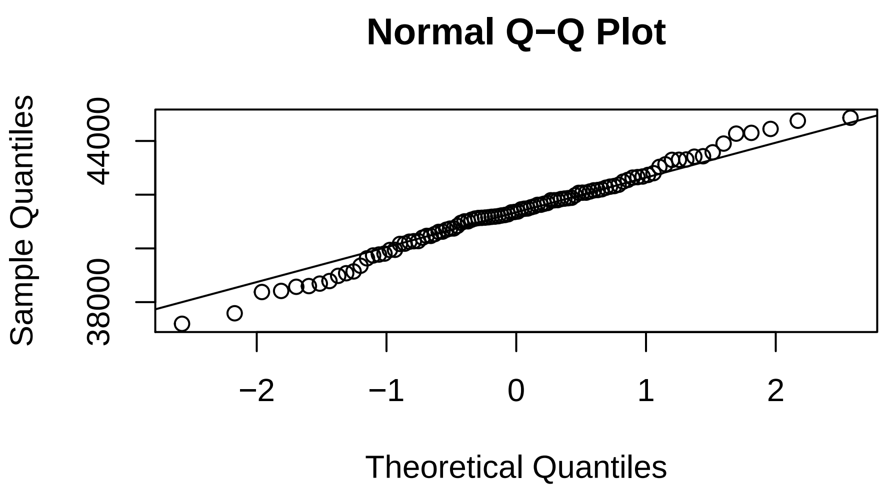


- The empirical cdf is pretty consistent of the cdf of the Normal.

b(v)

- QUESTION: Examine normality with a normal probability plot.

```
qqnorm(samp100size400$mean)
qqline(samp100size400$mean) #little deviation at the tails but it is a good fit
```



- The tails are slightly off from the line of normality, but it is Normal enough.

b(vi)

- QUESTION: For each of the 100 samples, find a 95% CI for the population avg income. How many of those intervals actually contain the population target?

```
meanFam <- mean(families$INCOME)
tally(~(lower_bound < meanFam) & (meanFam < upper_bound), data=samp100size400, format="proportion")
```

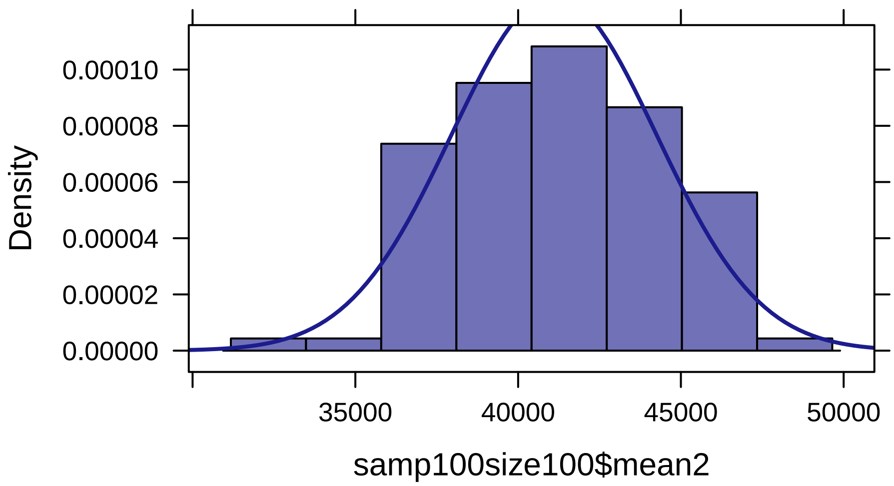
```
##                        (meanFam < upper_bound)
## (lower_bound < meanFam) TRUE FALSE
##                    TRUE  0.97  0.03
##                    FALSE 0.00  0.00
```

- Based on our simulation of a 100 samples of size 400, 97% of the samples contain the true population avg income.

b(vii)

- QUESTION: Take 100 samples of size 100. Find the averages, standard deviations, and histograms of these results and compare to the results for 100 samples of size 400. Explain how SRS relates to these comparisons.

```
meanInc2 <- mean(samp100size100$mean2)
sdInc2 <- sd(samp100size100$mean2)
histogram(samp100size100$mean2, density=TRUE)
```



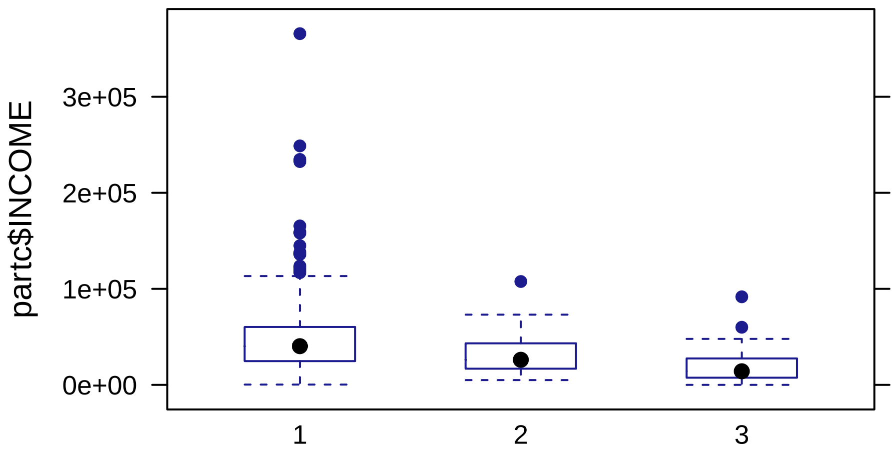
- The greater sample, the closer is the mean to the population mean and the less variation.

67c

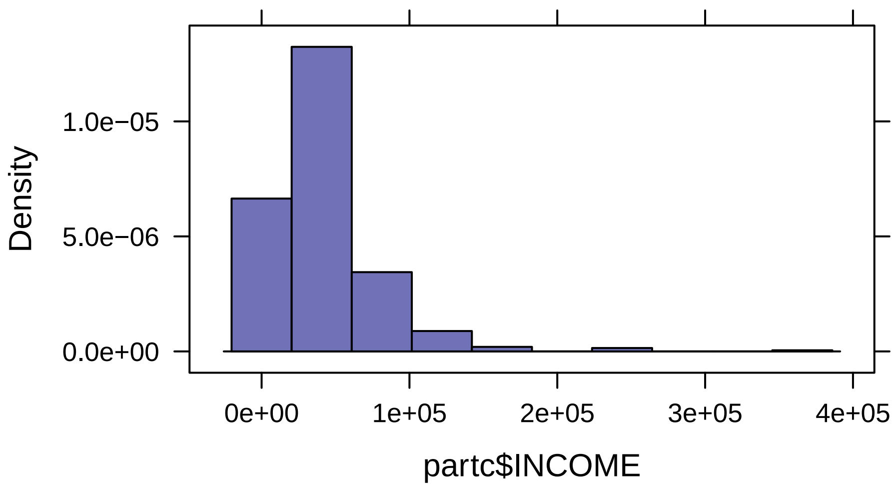
- QUESTION: For a SRS of 500, compare the incomes of the three family types via histograms and boxplots.

- Note: For TYPE, 1 = Husband-wife family ; 2 = Male-head family ; 3 = Female-head family

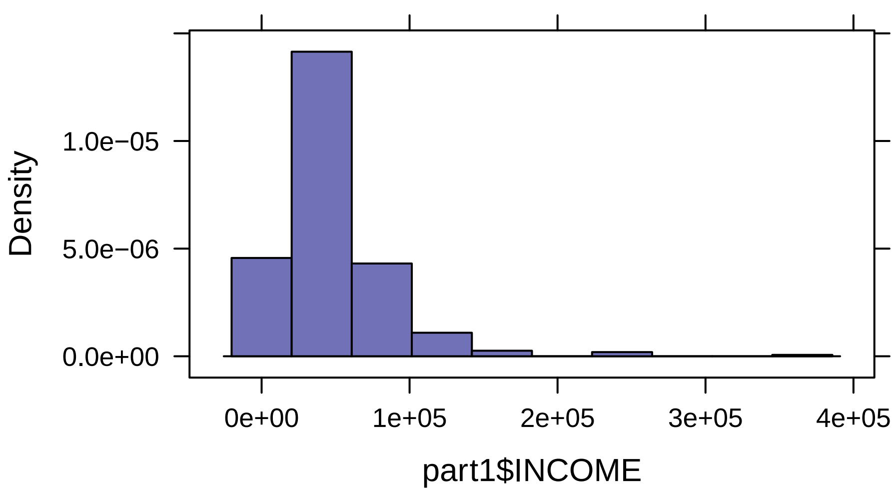
```
partc <- sample(families, 500)
bwplot(partc$INCOME ~ as.factor(TYPE), data=partc)
```



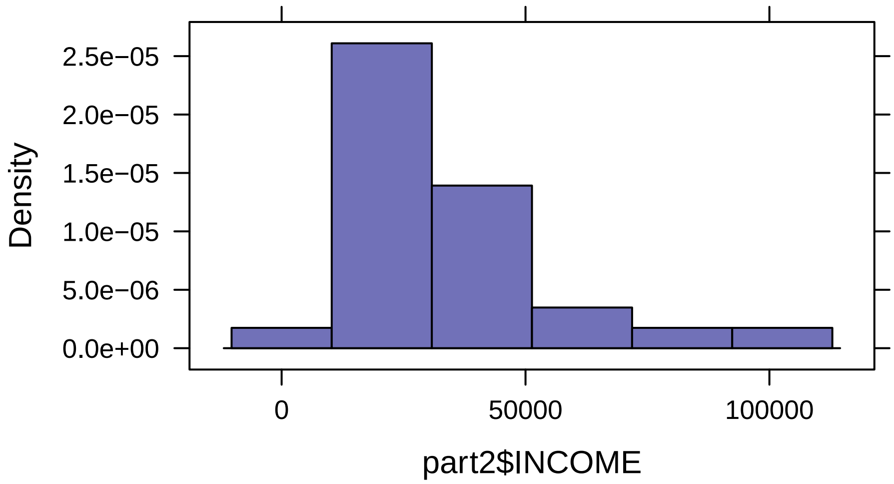

```
histogram(~ partc$INCOME, filterdata=partc)
```



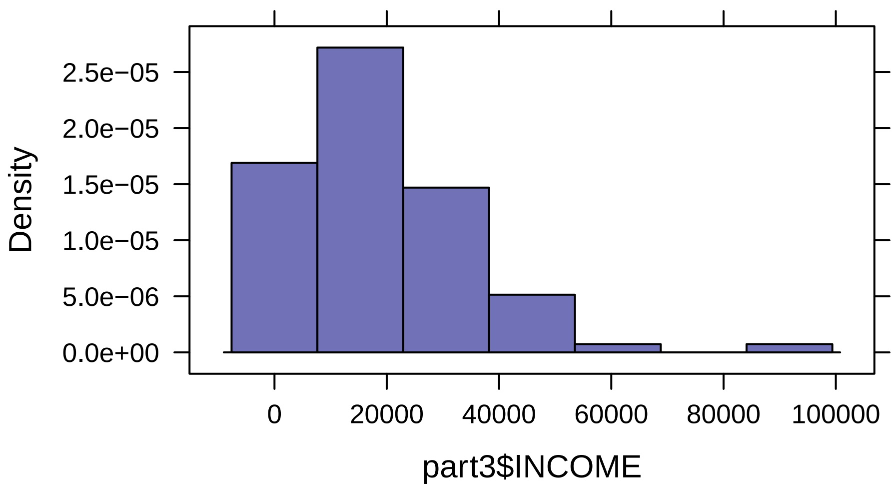

```
part1 <- filter(partc, TYPE ==1)
part2 <- filter(partc, TYPE ==2)
part3 <- filter(partc, TYPE ==3)
histogram(part1$INCOME)
```



```r
histogram(part2$INCOME)
```




```r
histogram(part3$INCOME)
```



- Husband-wife families (TYPE==1) have the highest average income. In the middle it is male-head families (TYPE==2). Lowest is the female-head families (TYPE==3).

67d

- QUESTION: Take a SRS of sample size 400 from each of the four regions.
- Note: For region, 1 is North, 2 is East, 3 is South, 4 is West.
- We filtered the 'families' dataset for each region. Then, we took size 400 samples from each region. We then combined all 4 datasets to have a 'totalsamp' dataset.
- Highest income is in the North. Lowest in East and South. Middle in the West.

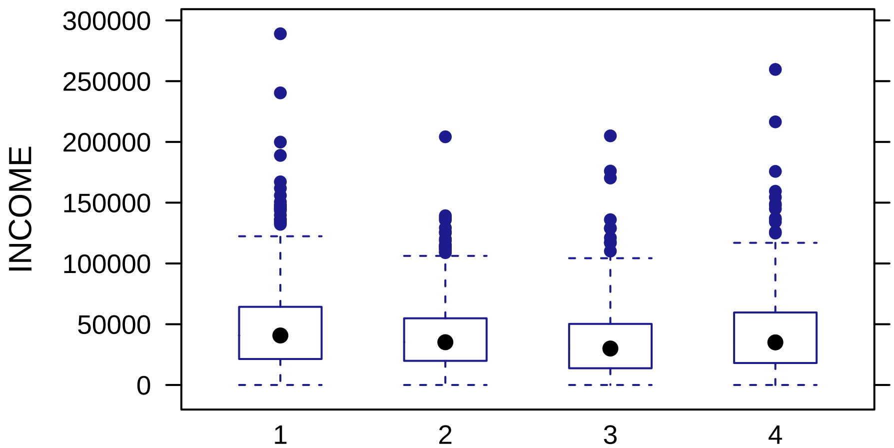
```r
set.seed(1)
NORTH <- filter(families, REGION ==1)
NORTHsamp <- sample(NORTH, 400)

EAST <- filter(families, REGION ==2)
EASTsamp <- sample(EAST, 400)

SOUTH <- filter(families, REGION ==3)
SOUTHsamp <- sample(SOUTH, 400)

WEST <- filter(families, REGION==4)
WESTsamp <- sample(WEST, 400)

totalsamp <- rbind(NORTHsamp, EASTsamp, SOUTHsamp, WESTsamp)
bwplot(INCOME ~ as.factor(REGION), data=totalsamp) #400 from each region
```



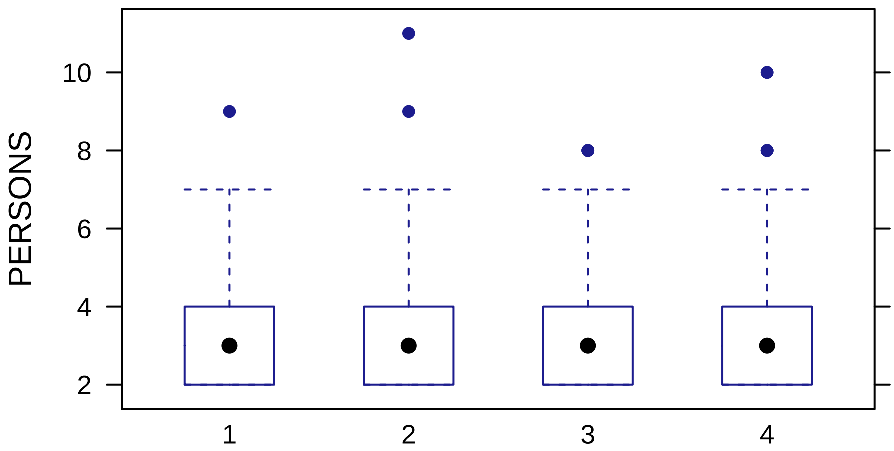

```r
favstats(INCOME ~ as.factor(REGION), data=totalsamp)$mean #population
```

```
## [1] 47948.99 41351.03 36027.92 43233.35
```

(dii)

QUESTION: Does it appear there are differences in family size across the 4 regions?

```r
bwplot(PERSONS ~ as.factor(REGION), data=totalsamp) #400 from each region
```




```
favstats(PERSONS ~ as.factor(REGION), data=totalsamp)$mean #population
```

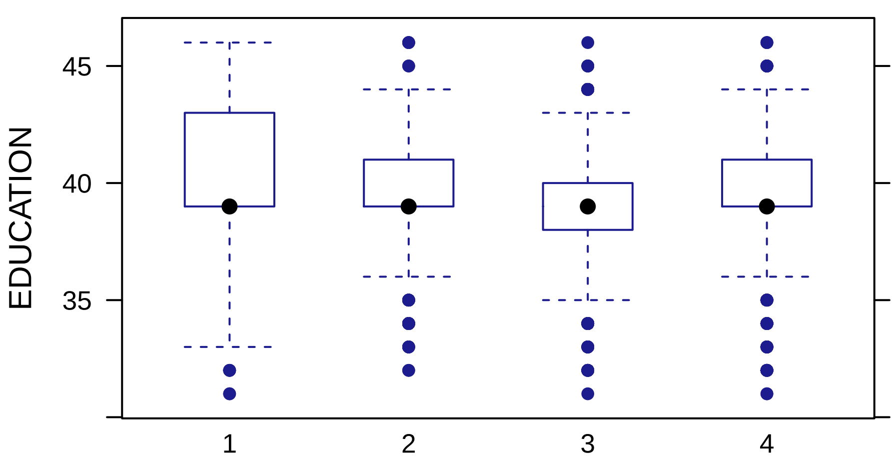
```
## [1] 3.1500 3.1125 3.1950 3.2575
```

- All families are about the same size across regions.

(diii)

- QUESTION: Are there differences in eucation level among the 4 regions?

```
bwplot(EDUCATION ~ as.factor(REGION), data=totalsamp) #population
```




```
favstats(EDUCATION ~ as.factor(REGION), data=totalsamp)$mean
```

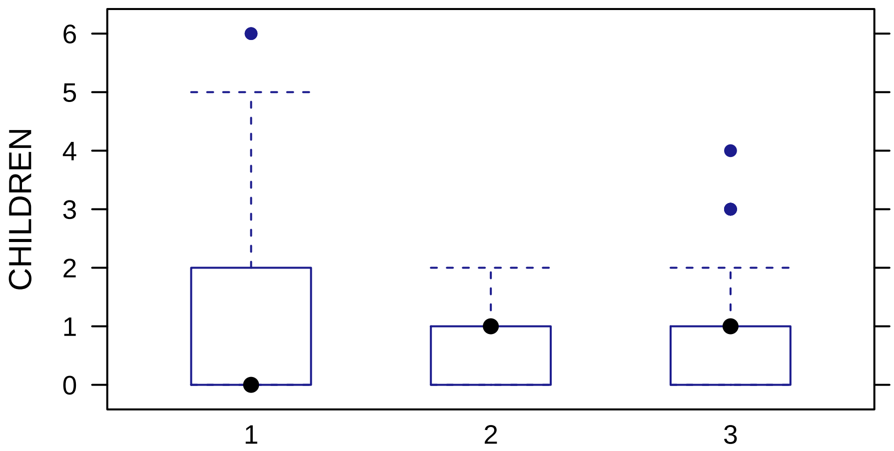
```
## [1] 39.6450 39.4525 39.0000 39.3450
```

- Neglible differences in means of education levels.

(e)

- QUESTION: For a simple random sample of 400, compare the # of children of the three family types.

```
set.seed(1)
ourques <- sample(families, 400)
bwplot(CHILDREN ~ as.factor(TYPE), data=ourques)
```



(f)

```
set.seed(5)
sampsize <- 400
samplef <- sample(families, sampsize)
mean(samplef$INCOME)
```

```
## [1] 42397.22
```

```
sd(samplef$INCOME)
```

```
## [1] 29142.44
```

```
favstats(samplef$INCOME)
```

```
##    min      Q1  median      Q3    max     mean       sd   n missing
##  -9999 20157.5 37267.5 57427.5 203253 42397.22 29142.44 400       0
```

```
confint(t.test(samplef$INCOME))
```

```
##   mean of x    lower    upper level
## 1  42397.22 39532.62 45261.82  0.95
```

```r
propalloc <- tally(families$REGION, format="prop") * sampsize; propalloc
```

```
##
##         1         2         3         4
##  92.50330  94.69990 122.65415  90.14264
```

```r
stratNORTH <- sample(NORTH, 92)
stratEAST <- sample(EAST, 95)
stratSOUTH <- sample(SOUTH, 123)
stratWEST <- sample(WEST, 90)
totalstratsamp <- rbind(stratNORTH, stratEAST, stratSOUTH, stratWEST)

t.test(totalstratsamp$INCOME) #sample mean
```

```
##
##  One Sample t-test
##
## data:  totalstratsamp$INCOME
## t = 27.642, df = 399, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  39829.50 45928.79
## sample estimates:
## mean of x
##  42879.14
```

```r
t.test(families$INCOME) #population mean
```

```
##
##  One Sample t-test
##
## data:  families$INCOME
## t = 270.29, df = 43885, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  41035.76 41635.26
## sample estimates:
## mean of x
##  41335.51
```

```r
sumStratSamp <- sum(tally(~ REGION, data=families, format = "prop") *var(INCOME ~ REGION, data=totalstra
vareverything <- sumStratSamp/sampsize
sdstratified <- sqrt(vareverything); sdstratified
```

```
## [1] 1552.199
```

```r
NORTHf <- filter(samplef, REGION ==1)
EASTf <- filter(samplef, REGION ==2)
SOUTHf <- filter(samplef, REGION ==3)
WESTf <- filter(samplef, REGION ==4)
```

```r
f1 <- mean(~ INCOME, data=NORTHf)
f2 <- mean(~ INCOME, data=EASTf)
f3 <- mean(~ INCOME, data=SOUTHf)
f4 <- mean(~ INCOME, data=WESTf)

ftotal <- (f1 + f2 + f3 + f4) /4

g1 <- var(~ INCOME, data=NORTHf) /90
g2 <- var(~ INCOME, data=EASTf) / 109
g3 <- var(~ INCOME, data=SOUTHf)/116
g4 <- var(~ INCOME, data=WESTf) /85

stratvar1<- (g1+g2+g3+g4) /4
sqrt(stratvar1)
```

```
## [1] 2923.982
```

```r
g1 <- var(~ INCOME, data=NORTHf)
g2 <- var(~ INCOME, data=EASTf)
g3 <- var(~ INCOME, data=SOUTHf)
g4 <- var(~ INCOME, data=WESTf)

stratvar <- (g1+g2+g3+g4) /400 ## Stuggling w/ Stratification
sqrt(stratvar)
```

```
## [1] 2893.859
```