Cancer Mortality Exploration

w203 Teaching Team

Background

In this lab, imagine that your team is hired by a health government agency. They would like to understand factors that predict cancer mortality rates, with the ultimate aim of identifying communities for social interventions, and of understanding which interventions are likely to have the most impact. Your team was hired to perform an exploratory analysis to help the agency address their goals.

Data

You are given a dataset for a selection of US counties, **"cancer.csv"**. The dependent (or target) variable in this data is named "deathRate".

The labels of some of the variables are listed below; the rest of the variables should be self-explanatory.

avgAnnCount:	"2009-2013 mean incidences per county"
povertyPercent:	"Percent of population below poverty line"
popEst2015:	"Estimated population by county 2015"
PctPrivateCoverage:	"Percentage of the population with private insurance coverage"
PctPublicCoverage:	"Percentage of the population with public insurance coverage"

Objective

Perform an exploratory analysis to understand how county-level characteristics are related to cancer mortality.

```
setwd('~/Documents/MIDS/W203/hw/Lab_1/Cancer_EDA')
Cancer = read.csv('cancer.csv')
```

```
colnames(Cancer)
```

```
##  [1] "X"                     "avgAnnCount"           "medIncome"
##  [4] "popEst2015"            "povertyPercent"        "binnedInc"
##  [7] "MedianAge"             "MedianAgeMale"         "MedianAgeFemale"
## [10] "Geography"             "AvgHouseholdSize"      "PercentMarried"
## [13] "PctNoHS18_24"          "PctHS18_24"            "PctSomeCol18_24"
## [16] "PctBachDeg18_24"       "PctHS25_Over"          "PctBachDeg25_Over"
## [19] "PctEmployed16_Over"    "PctUnemployed16_Over" "PctPrivateCoverage"
## [22] "PctEmpPrivCoverage"    "PctPublicCoverage"     "PctWhite"
## [25] "PctBlack"              "PctAsian"              "PctOtherRace"
## [28] "PctMarriedHouseholds" "BirthRate"             "deathRate"
```

```
nrow(Cancer)
```

```
## [1] 3047
```

```
summary(Cancer)
```

```
##        X           avgAnnCount        medIncome        popEst2015
##  Min.   :   1.0   Min.   :    6.0   Min.   : 22640   Min.   :     827
##  1st Qu.: 762.5   1st Qu.:   76.0   1st Qu.: 38882   1st Qu.:   11684
##  Median :1524.0   Median :  171.0   Median : 45207   Median :   26643
##  Mean   :1524.0   Mean   :  606.3   Mean   : 47063   Mean   :  102637
```

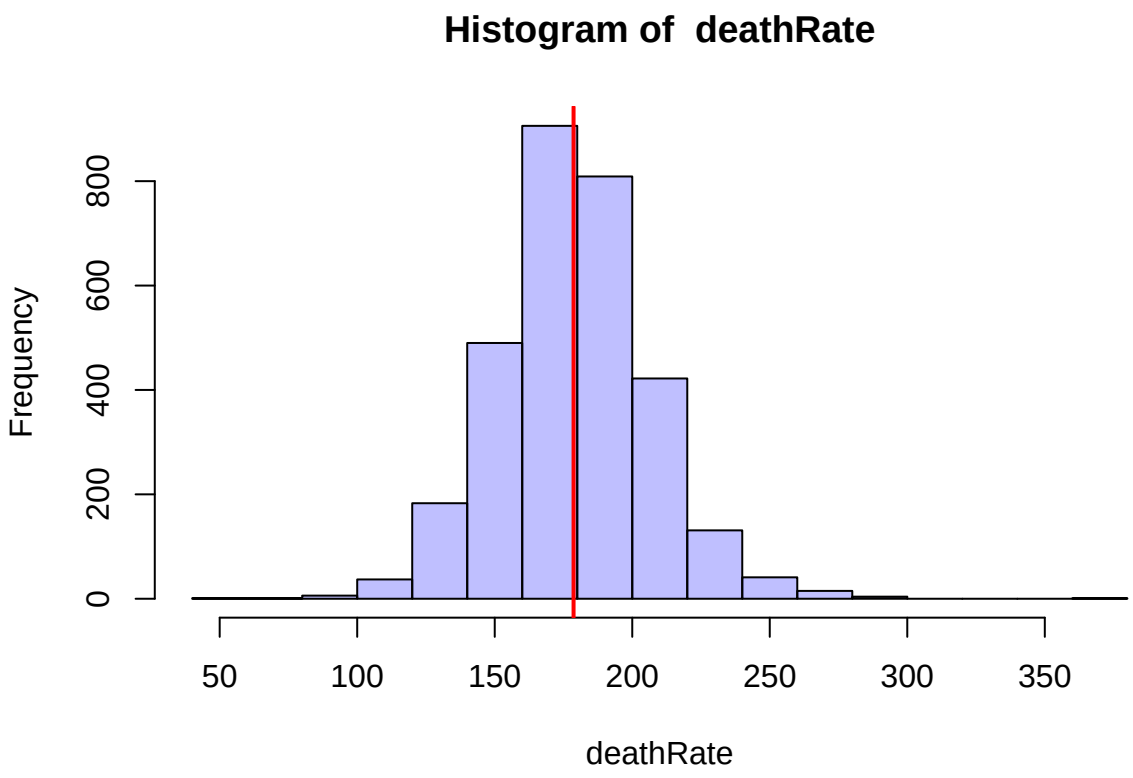
```
##  3rd Qu.:2285.5   3rd Qu.:  518.0   3rd Qu.: 52492   3rd Qu.:   68671
##  Max.   :3047.0   Max.   :38150.0   Max.   :125635   Max.   :10170292
##
##  povertyPercent               binnedInc      MedianAge
##  Min.   : 3.20   (45201, 48021.6] : 306   Min.   : 22.30
##  1st Qu.:12.15   (54545.6, 61494.5]: 306   1st Qu.: 37.70
##  Median :15.90   [22640, 34218.1] : 306   Median : 41.00
##  Mean   :16.88   (42724.4, 45201] : 305   Mean   : 45.27
##  3rd Qu.:20.40   (48021.6, 51046.4]: 305   3rd Qu.: 44.00
##  Max.   :47.40   (51046.4, 54545.6]: 305   Max.   :624.00
##                  (Other)           :1214
##  MedianAgeMale   MedianAgeFemale                              Geography
##  Min.   :22.40   Min.   :22.30   Abbeville County, South Carolina:   1
##  1st Qu.:36.35   1st Qu.:39.10   Acadia Parish, Louisiana        :   1
##  Median :39.60   Median :42.40   Accomack County, Virginia       :   1
##  Mean   :39.57   Mean   :42.15   Ada County, Idaho               :   1
##  3rd Qu.:42.50   3rd Qu.:45.30   Adair County, Iowa              :   1
##  Max.   :64.70   Max.   :65.70   Adair County, Kentucky          :   1
##                                  (Other)                         :3041
##  AvgHouseholdSize PercentMarried   PctNoHS18_24     PctHS18_24
##  Min.   :0.0221   Min.   :23.10   Min.   : 0.00   Min.   : 0.0
##  1st Qu.:2.3700   1st Qu.:47.75   1st Qu.:12.80   1st Qu.:29.2
##  Median :2.5000   Median :52.40   Median :17.10   Median :34.7
##  Mean   :2.4797   Mean   :51.77   Mean   :18.22   Mean   :35.0
##  3rd Qu.:2.6300   3rd Qu.:56.40   3rd Qu.:22.70   3rd Qu.:40.7
##  Max.   :3.9700   Max.   :72.50   Max.   :64.10   Max.   :72.5
##
##  PctSomeCol18_24 PctBachDeg18_24   PctHS25_Over   PctBachDeg25_Over
##  Min.   : 7.10   Min.   : 0.000   Min.   : 7.50   Min.   : 2.50
##  1st Qu.:34.00   1st Qu.: 3.100   1st Qu.:30.40   1st Qu.: 9.40
##  Median :40.40   Median : 5.400   Median :35.30   Median :12.30
##  Mean   :40.98   Mean   : 6.158   Mean   :34.80   Mean   :13.28
##  3rd Qu.:46.40   3rd Qu.: 8.200   3rd Qu.:39.65   3rd Qu.:16.10
##  Max.   :79.00   Max.   :51.800   Max.   :54.80   Max.   :42.20
##  NA's   :2285
##  PctEmployed16_Over PctUnemployed16_Over PctPrivateCoverage
##  Min.   :17.60      Min.   : 0.400       Min.   :22.30
##  1st Qu.:48.60      1st Qu.: 5.500       1st Qu.:57.20
##  Median :54.50      Median : 7.600       Median :65.10
##  Mean   :54.15      Mean   : 7.852       Mean   :64.35
##  3rd Qu.:60.30      3rd Qu.: 9.700       3rd Qu.:72.10
##  Max.   :80.10      Max.   :29.400       Max.   :92.30
##  NA's   :152
##  PctEmpPrivCoverage PctPublicCoverage    PctWhite         PctBlack
##  Min.   :13.5       Min.   :11.20     Min.   : 10.20   Min.   : 0.0000
##  1st Qu.:34.5       1st Qu.:30.90     1st Qu.: 77.30   1st Qu.: 0.6207
##  Median :41.1       Median :36.30     Median : 90.06   Median : 2.2476
##  Mean   :41.2       Mean   :36.25     Mean   : 83.65   Mean   : 9.1080
##  3rd Qu.:47.7       3rd Qu.:41.55     3rd Qu.: 95.45   3rd Qu.:10.5097
##  Max.   :70.7       Max.   :65.10     Max.   :100.00   Max.   :85.9478
##
##     PctAsian        PctOtherRace     PctMarriedHouseholds   BirthRate
##  Min.   : 0.0000   Min.   : 0.0000   Min.   :22.99        Min.   : 0.000
##  1st Qu.: 0.2542   1st Qu.: 0.2952   1st Qu.:47.76        1st Qu.: 4.521
```

```
##  Median : 0.5498   Median : 0.8262   Median :51.67        Median : 5.381
##  Mean   : 1.2540   Mean   : 1.9835   Mean   :51.24        Mean   : 5.640
##  3rd Qu.: 1.2210   3rd Qu.: 2.1780   3rd Qu.:55.40        3rd Qu.: 6.494
##  Max.   :42.6194   Max.   :41.9303   Max.   :78.08        Max.   :21.326
##
##    deathRate
##  Min.   : 59.7
##  1st Qu.:161.2
##  Median :178.1
##  Mean   :178.7
##  3rd Qu.:195.2
##  Max.   :362.8
##
```

```
attach(Cancer)
```

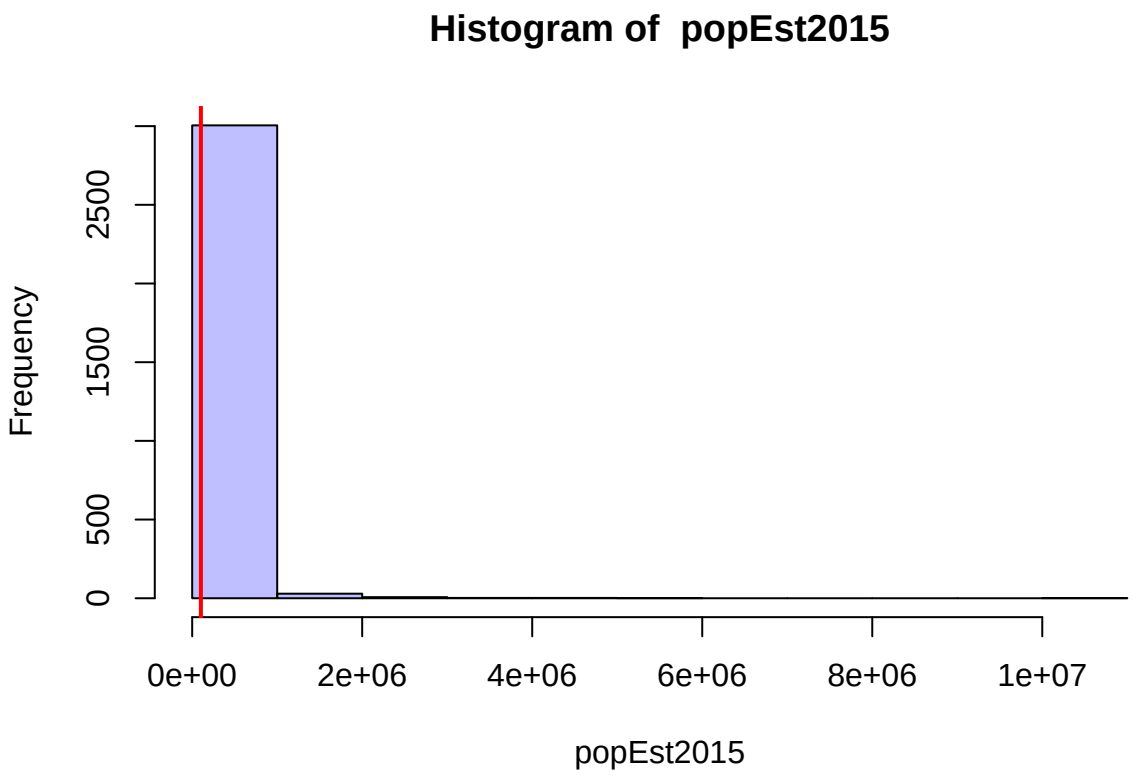
```
histWithMean <- function(vector, name) {
  hist(vector, col=rgb(0,0,1,1/4), main=paste("Histogram of ", name), xlab=name)
  abline(v = mean(vector), col="red", lwd=2)
}
```

```
histWithMean(deathRate, "deathRate")
```

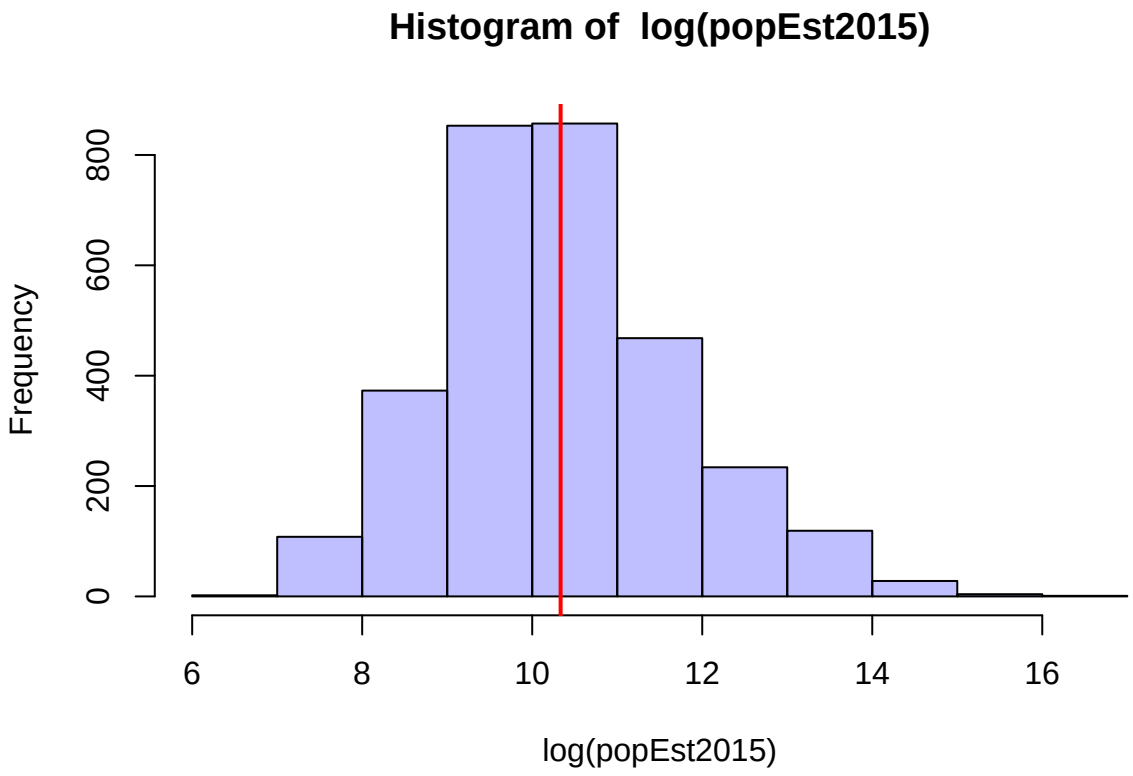


Population

```
histWithMean(popEst2015, "popEst2015") # looks like "power law distribution"
```

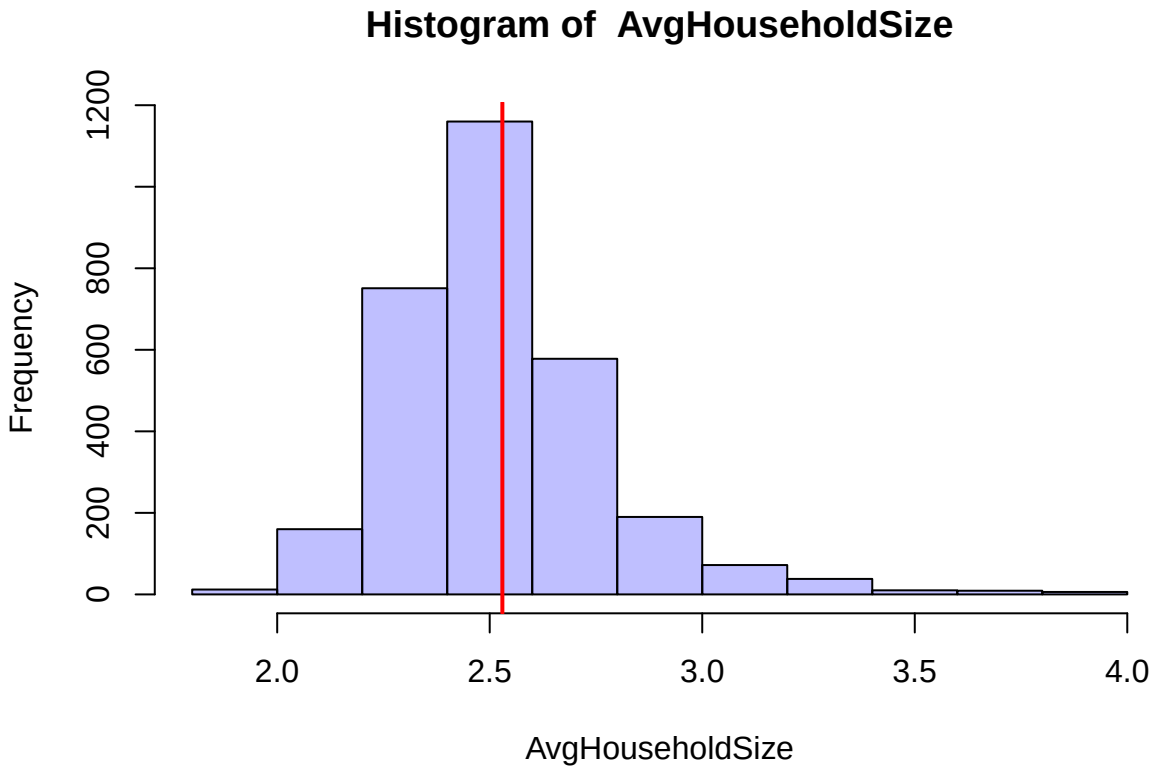



```
histWithMean(log(popEst2015), "log(popEst2015)")
```

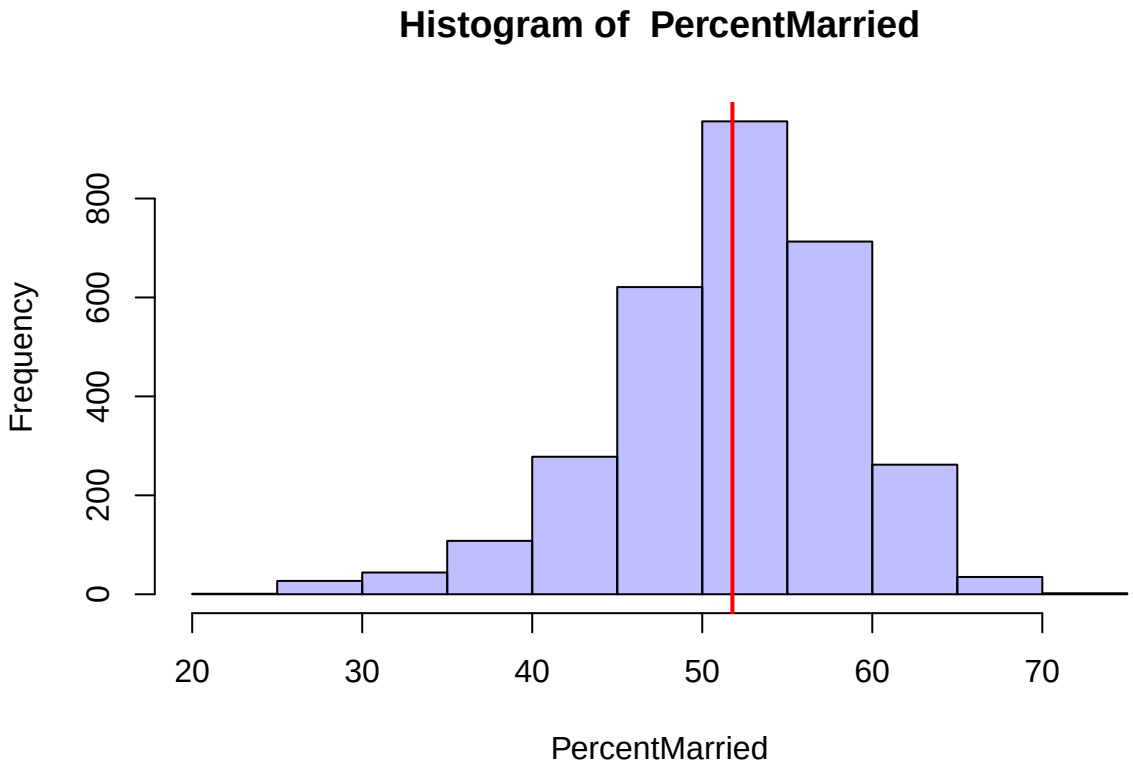



```
Cancer$logPopEst2015 = log(popEst2015)

histWithMean(AvgHouseholdSize, "AvgHouseholdSize") # impossible 0's
```

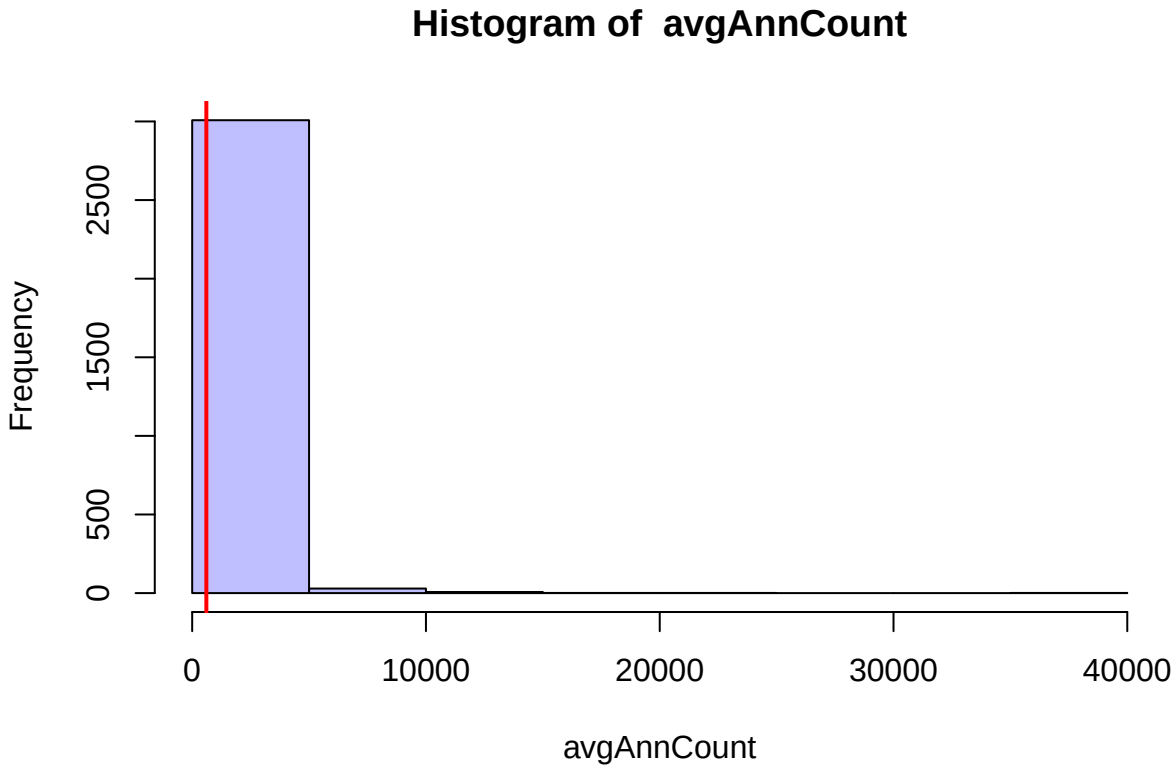

```
histWithMean(AvgHouseholdSize[AvgHouseholdSize > 1], "AvgHouseholdSize")
```



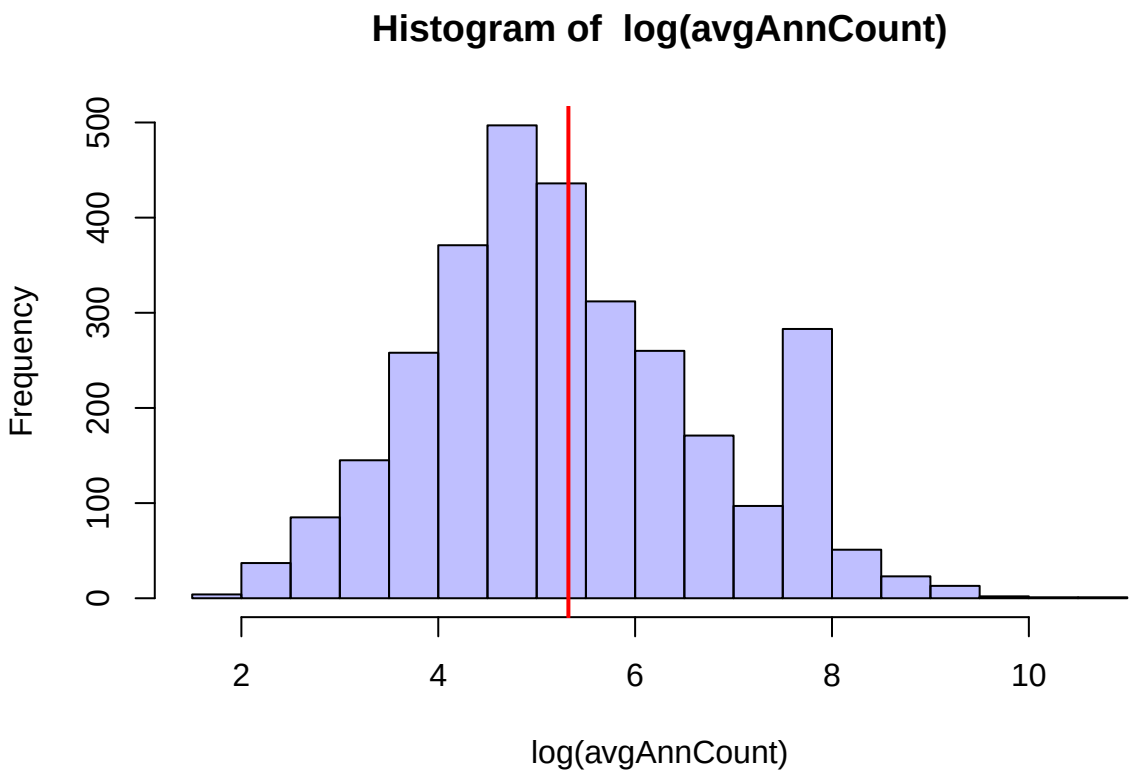

```
cleanAvgHouseholdSize <- AvgHouseholdSize > 1

histWithMean(PercentMarried, "PercentMarried")
```



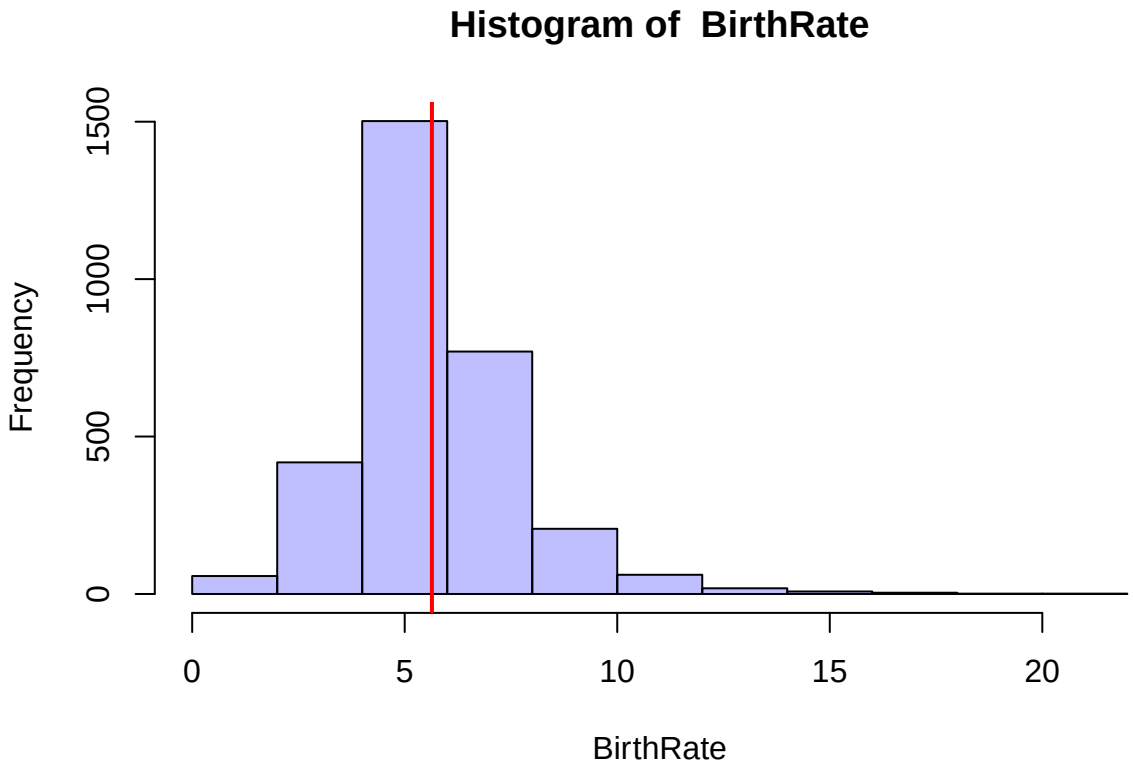

```
histWithMean(avgAnnCount, "avgAnnCount") # looks like "power law distribution"
```




```
histWithMean(log(avgAnnCount), "log(avgAnnCount)")
```

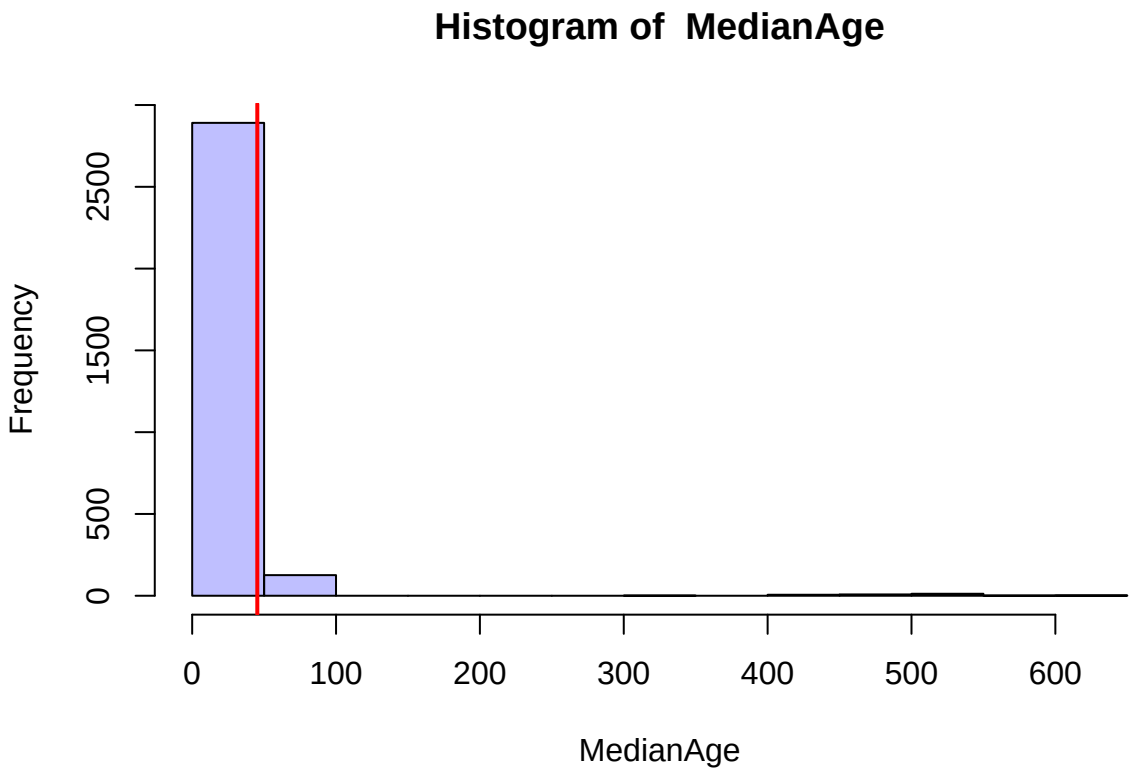



```
Cancer$logAvgAnnCount = log(avgAnnCount)

histWithMean(BirthRate, "BirthRate")
```

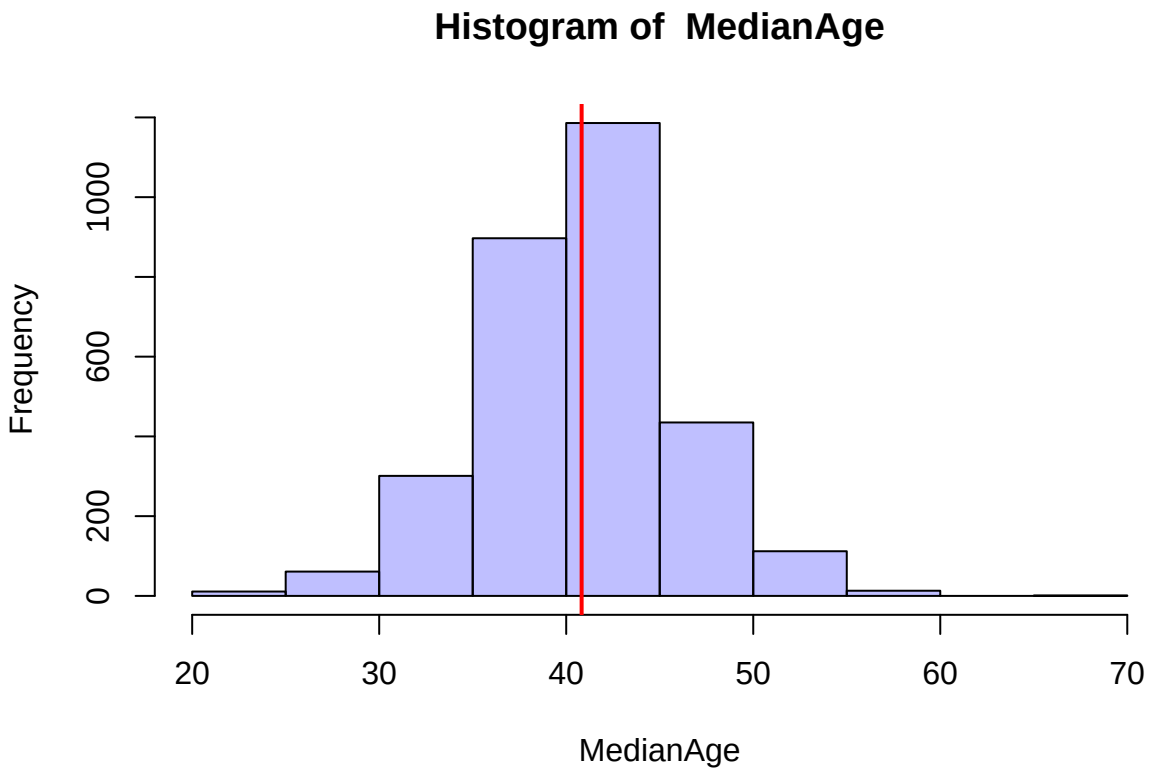


Age

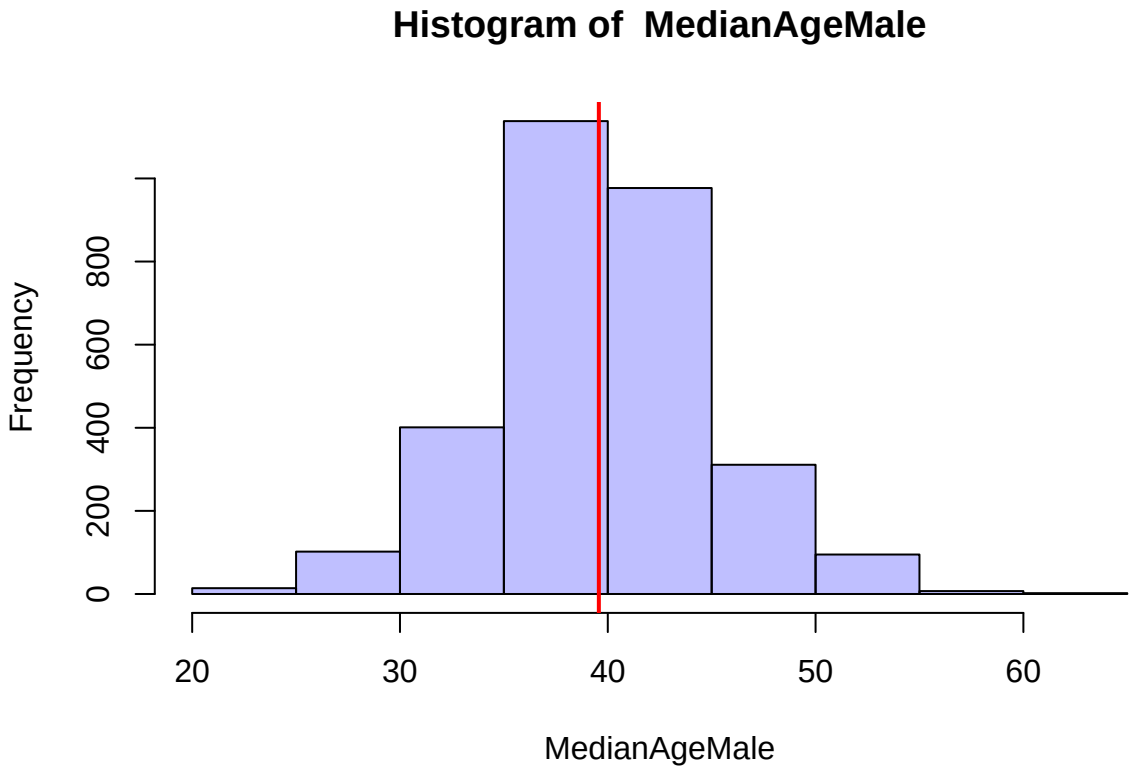
```
histWithMean(MedianAge, "MedianAge") # impossible over 200
```



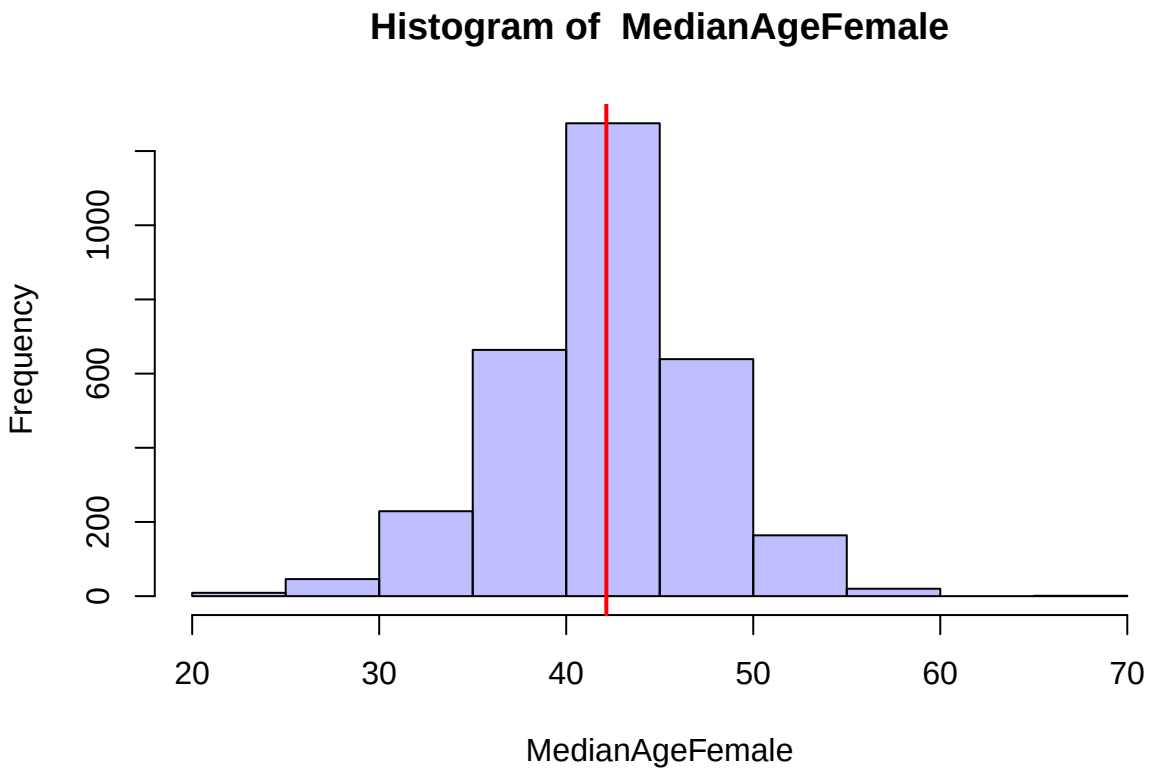

```
histWithMean(MedianAge[MedianAge < 200], "MedianAge")
```



```r
cleanMedianAge <- MedianAge < 200

histWithMean(MedianAgeMale, "MedianAgeMale")
```

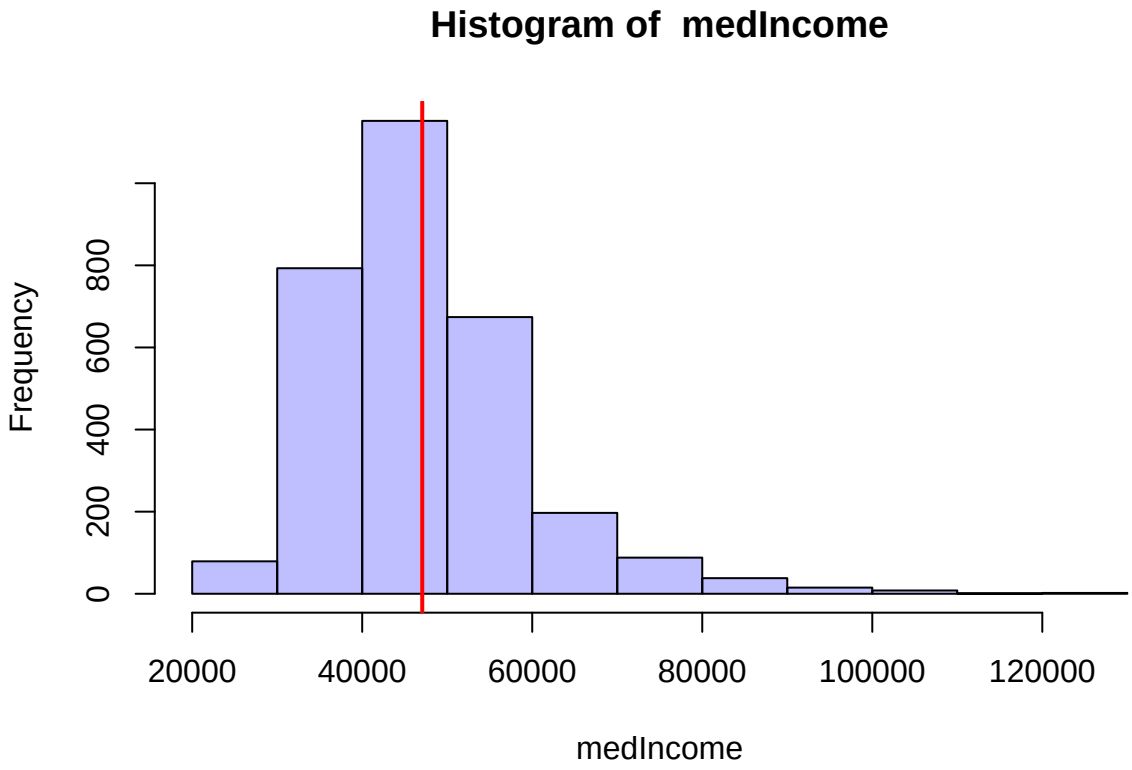



```r
histWithMean(MedianAgeFemale, "MedianAgeFemale")
```

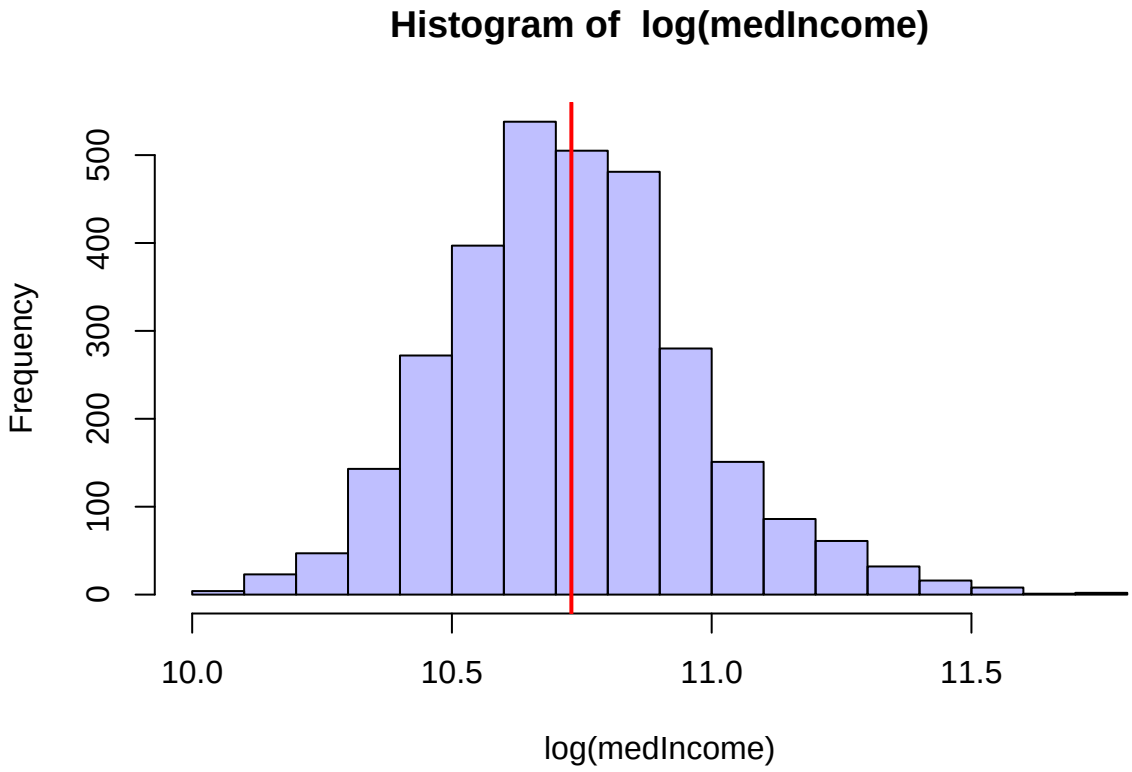


Income

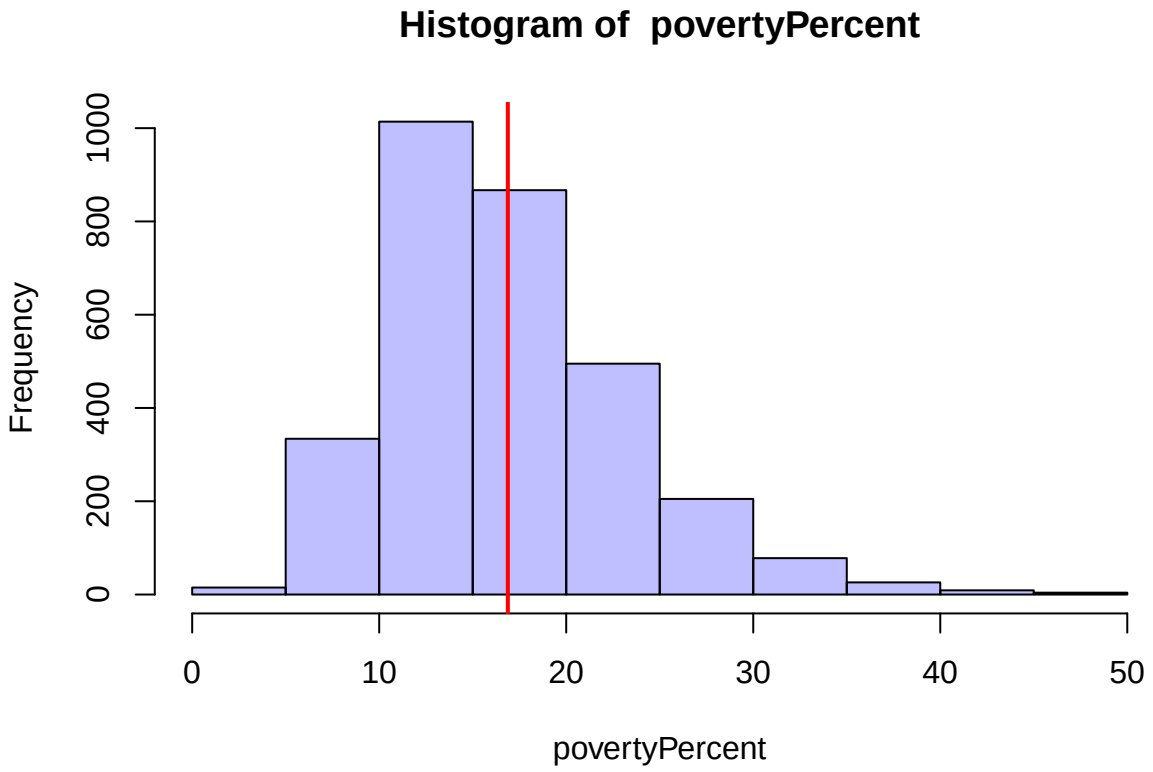
```
histWithMean(medIncome, "medIncome") # looks like "power law distribution"
```



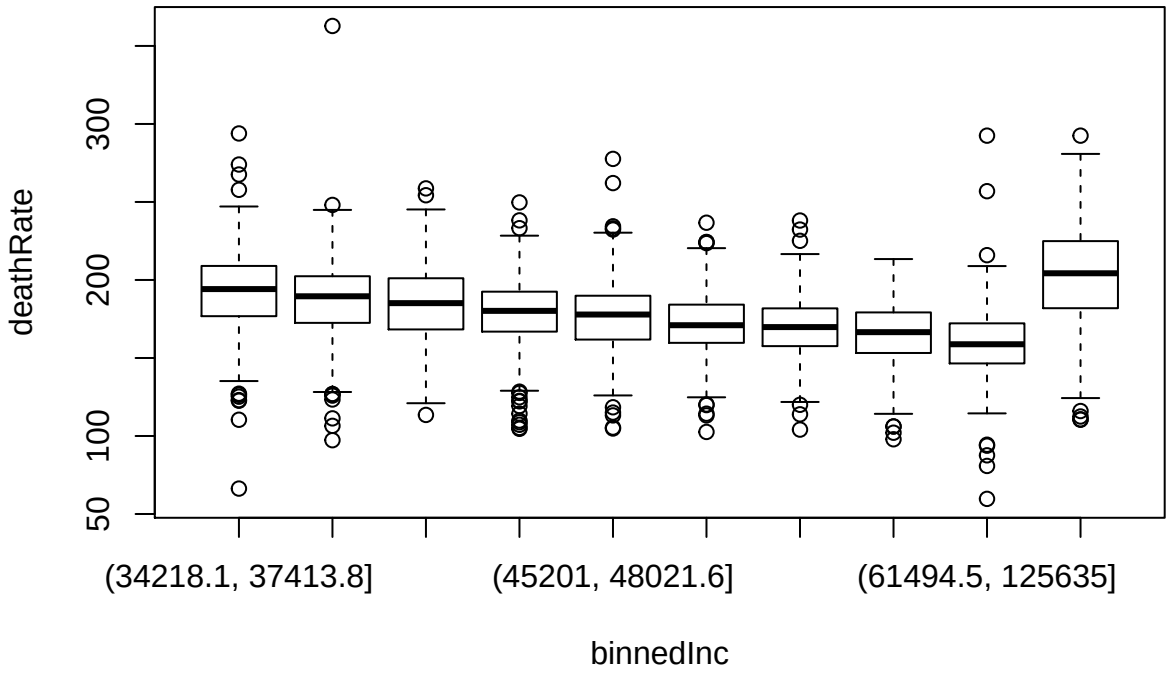

```
histWithMean(log(medIncome), "log(medIncome)")
```



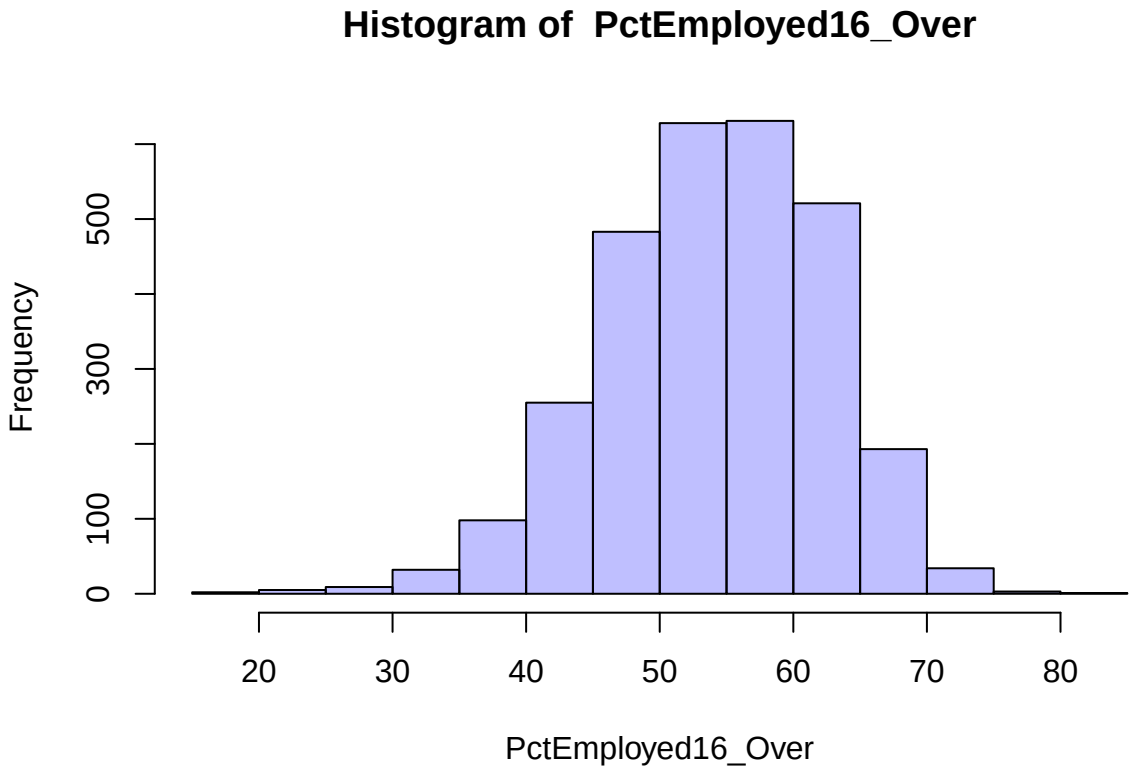
```
Cancer$logMedIncome = log(medIncome)

histWithMean(povertyPercent, "povertyPercent")
```



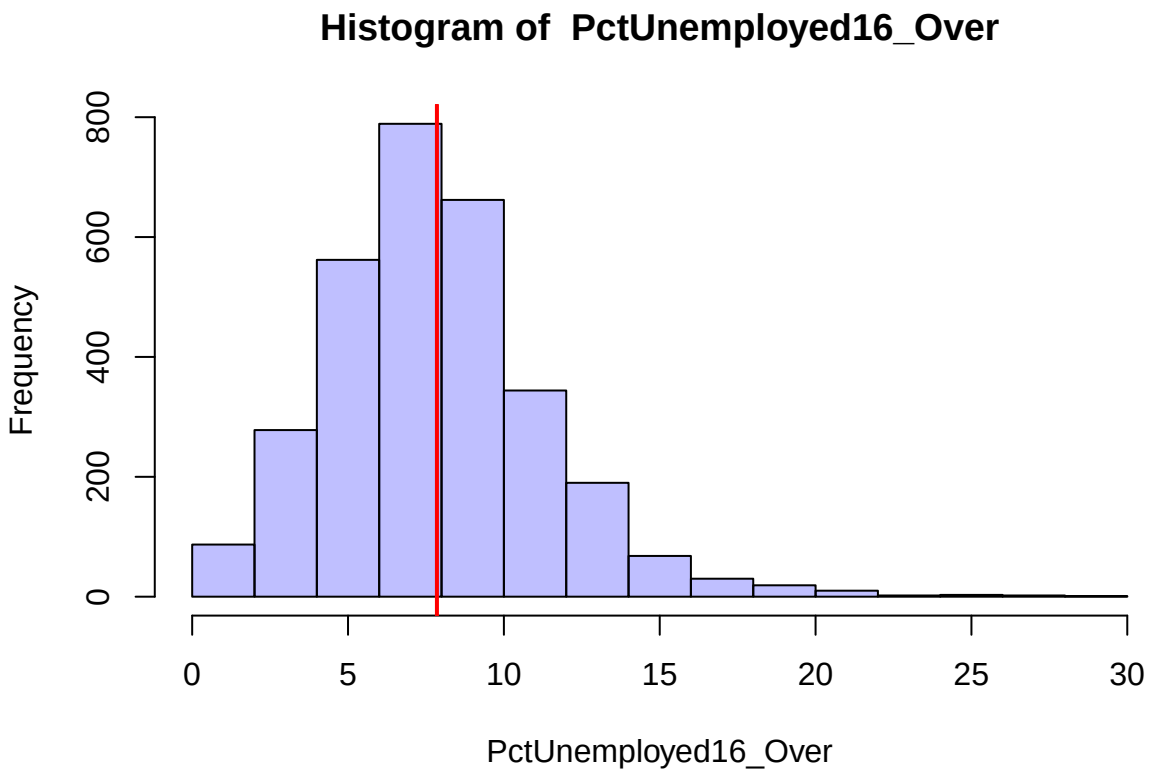

```
plot(deathRate ~ binnedInc)
```



```
histWithMean(PctEmployed16_Over, "PctEmployed16_Over") # NA's: 152
```

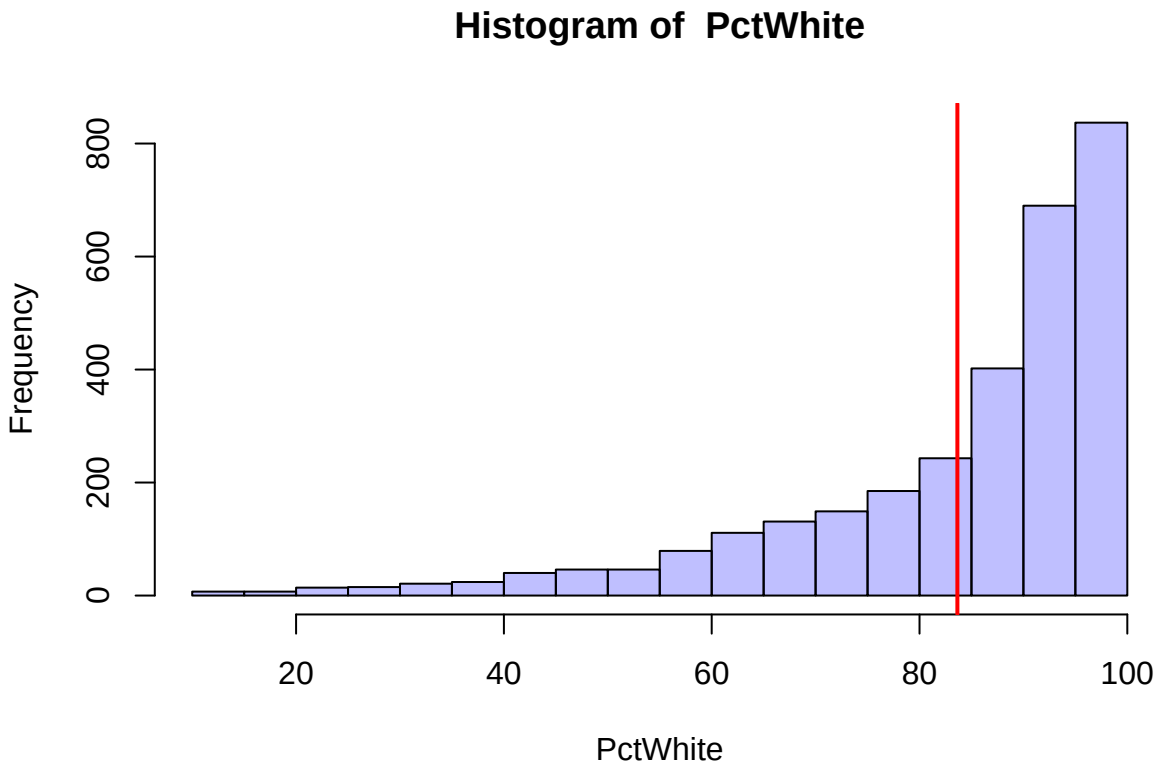



```
cleanPctEmployed16_Over <- !is.na(PctEmployed16_Over)

histWithMean(PctUnemployed16_Over, "PctUnemployed16_Over")
```

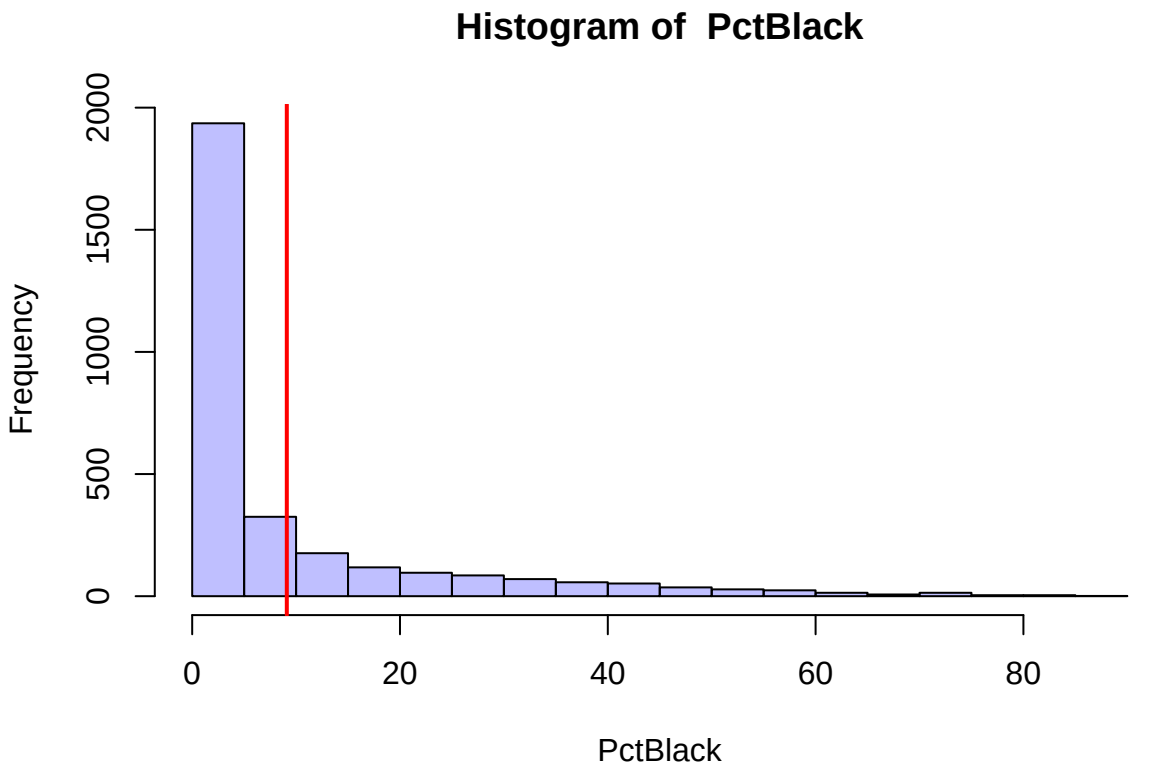


Race

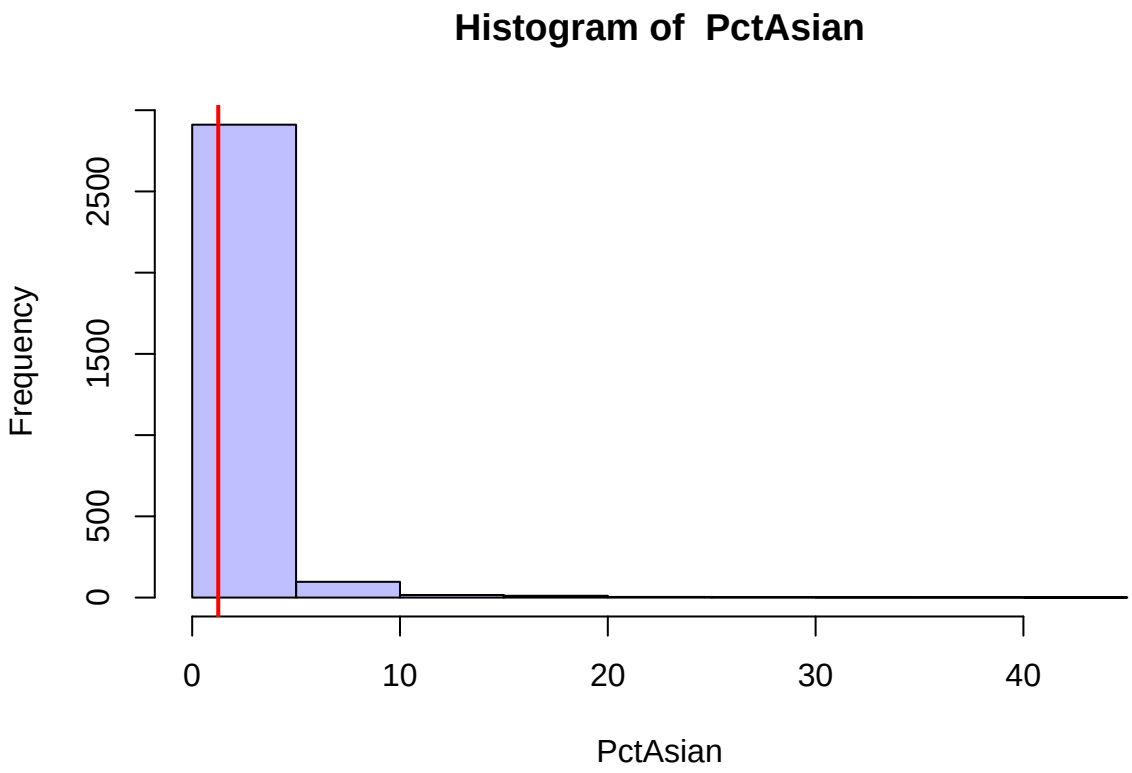
```
histWithMean(PctWhite, "PctWhite")
```



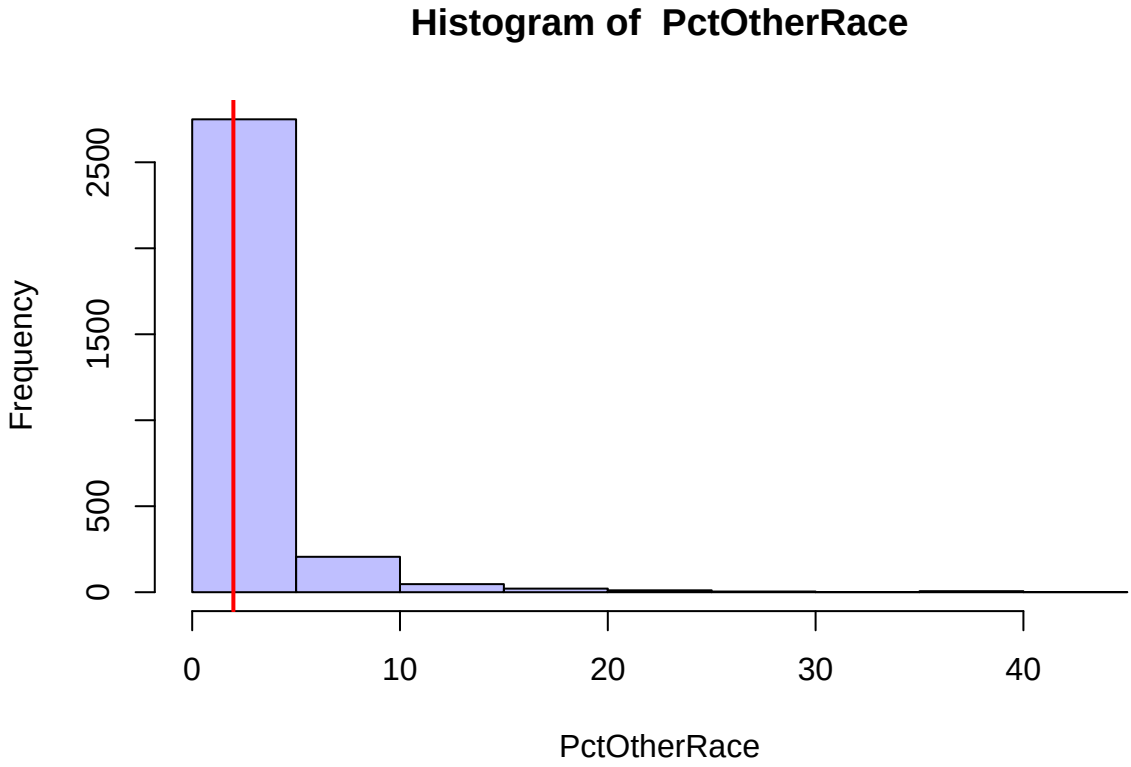

```
histWithMean(PctBlack, "PctBlack")
```



```
histWithMean(PctAsian, "PctAsian")
```

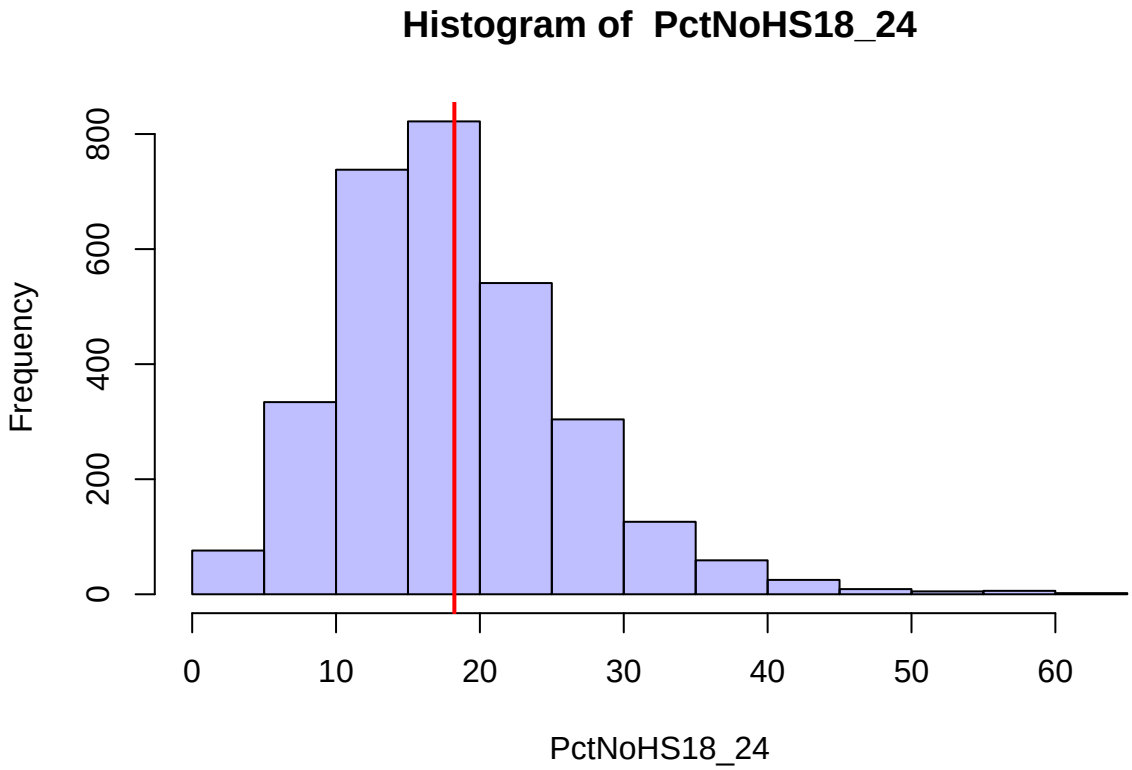



```
histWithMean(PctOtherRace, "PctOtherRace")
```

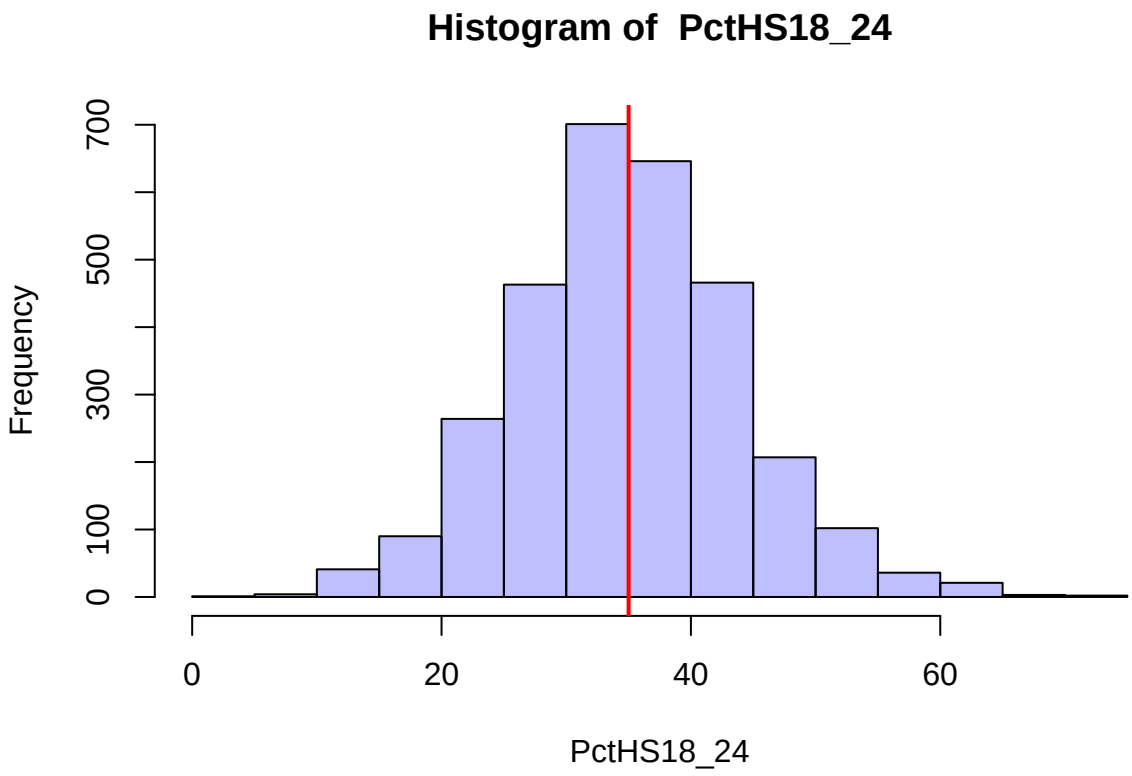


Education

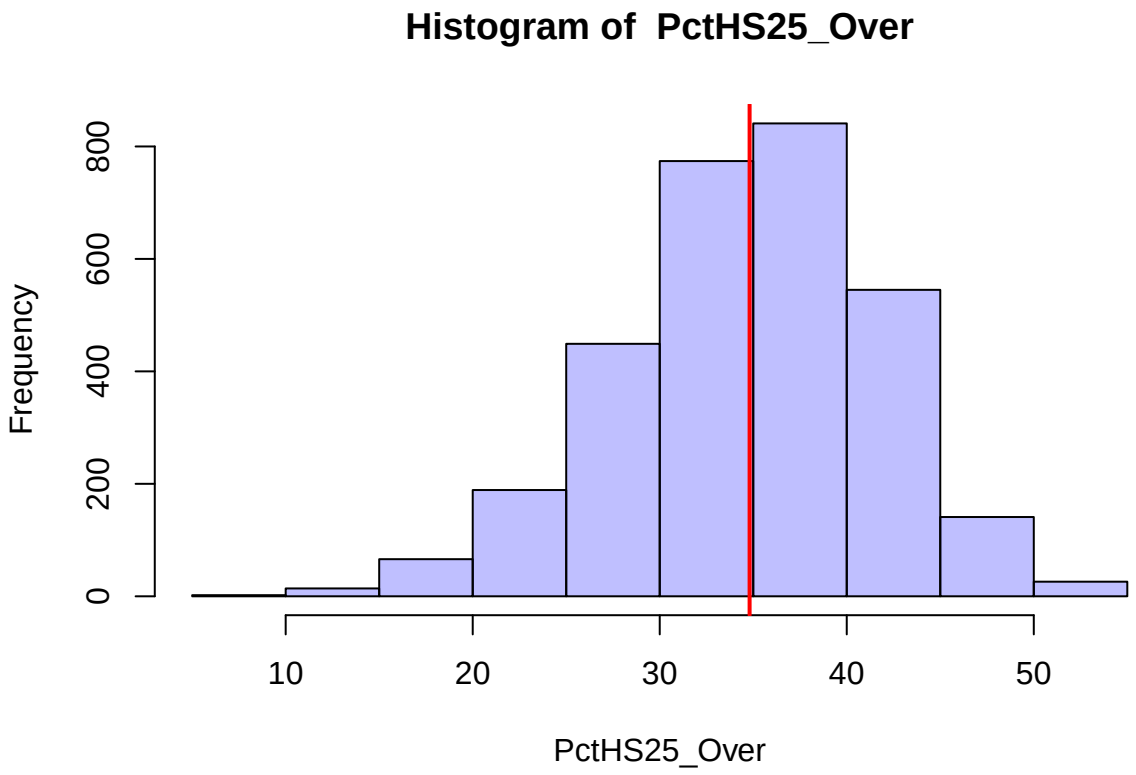
```
histWithMean(PctNoHS18_24, "PctNoHS18_24")
```



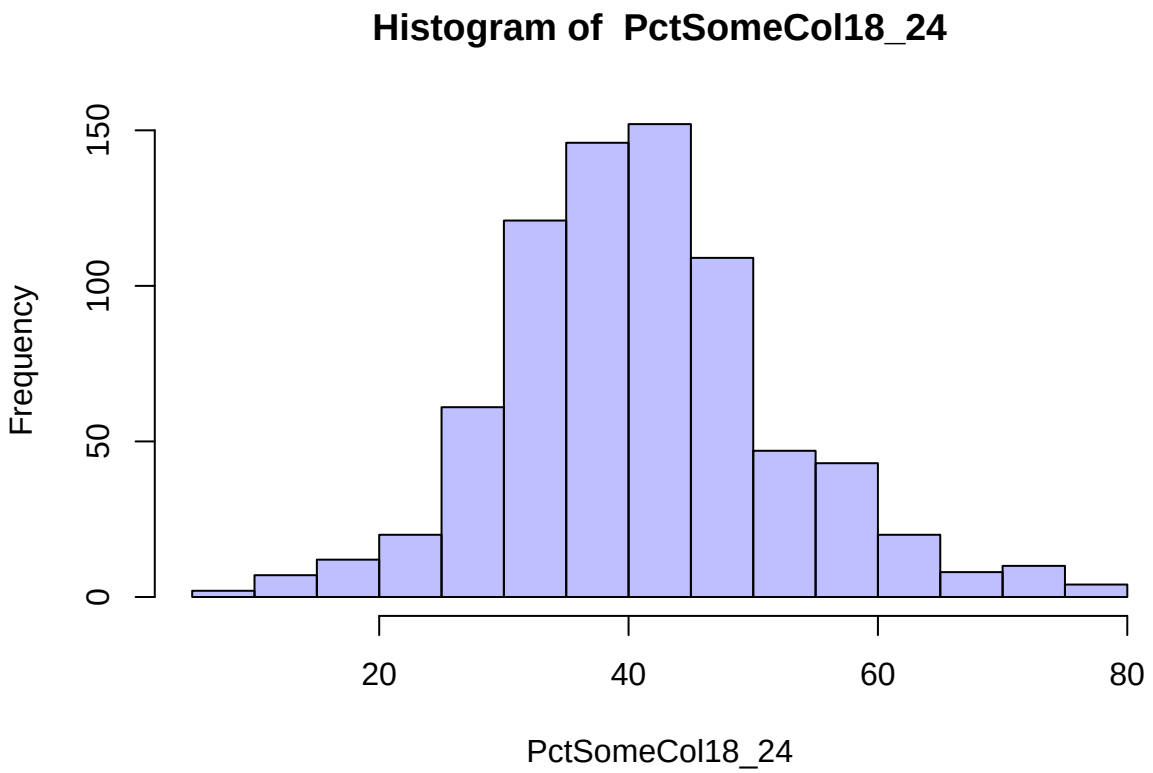

```
histWithMean(PctHS18_24, "PctHS18_24")
```



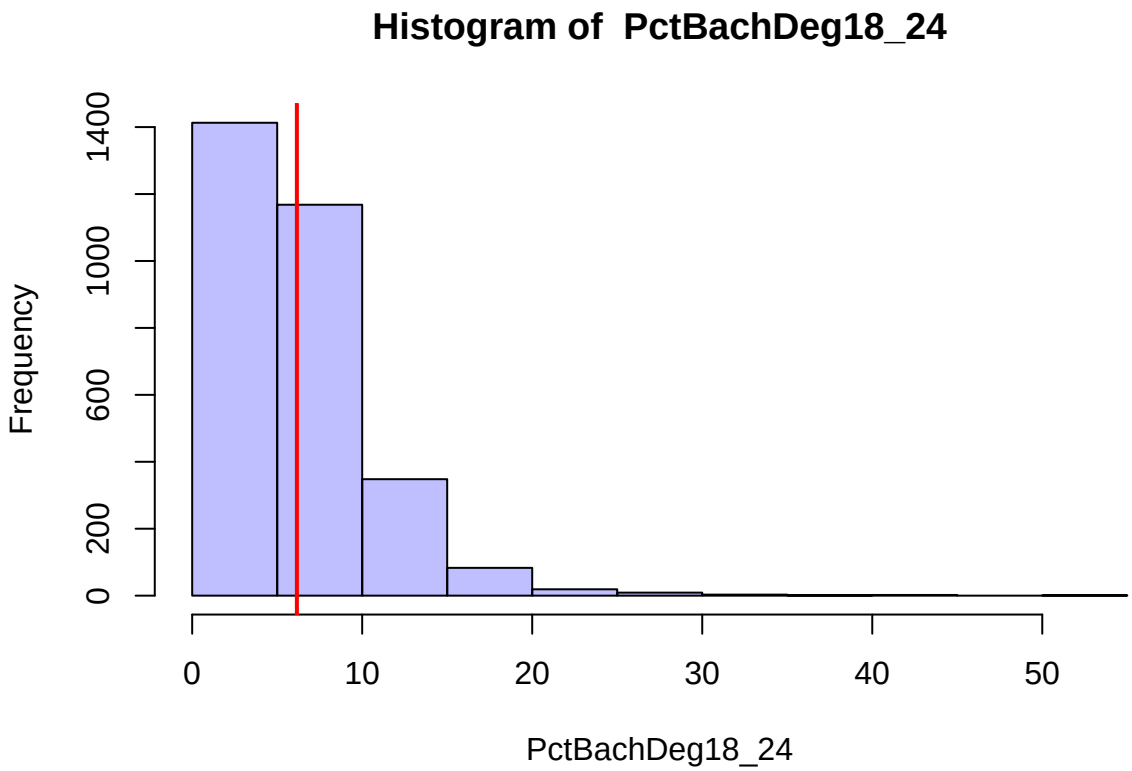
```r
histWithMean(PctHS25_Over, "PctHS25_Over")
```



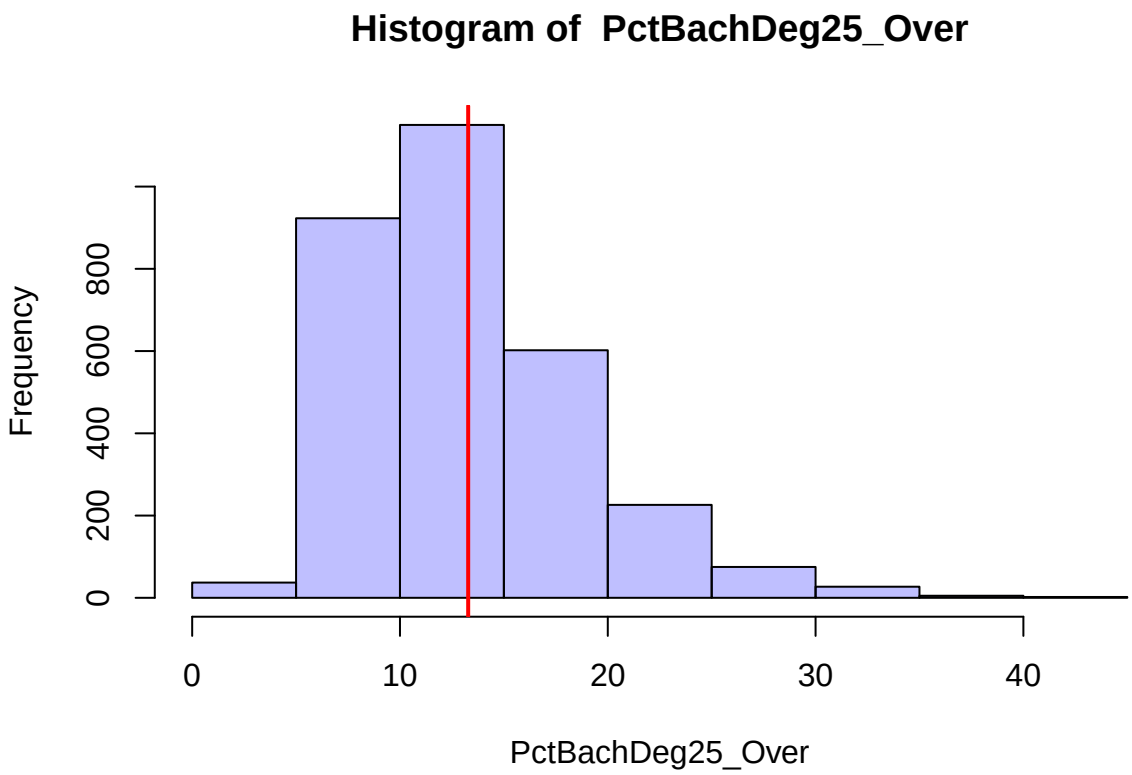

```r
histWithMean(PctSomeCol18_24, "PctSomeCol18_24") # NA's: 2285
```



```
cleanPctSomeCol18_24 <- !is.na(PctSomeCol18_24)
histWithMean(PctBachDeg18_24, "PctBachDeg18_24")
```

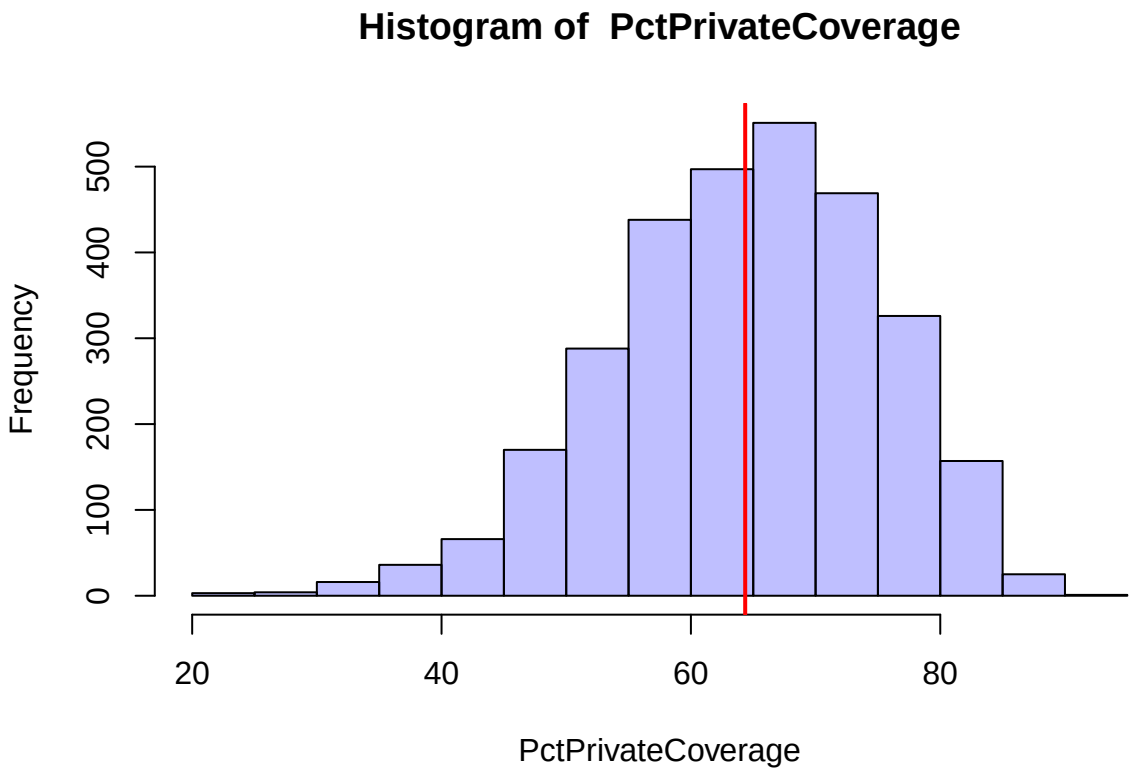



```
histWithMean(PctBachDeg25_Over, "PctBachDeg25_Over")
```

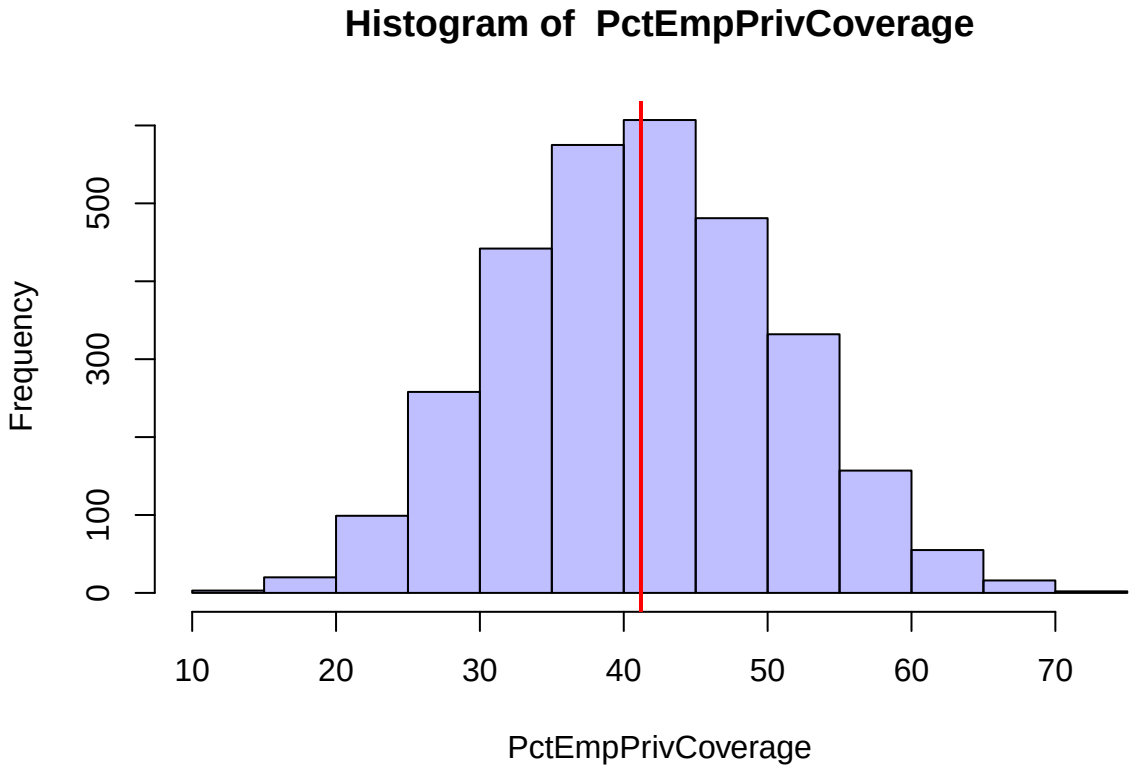


Insurance

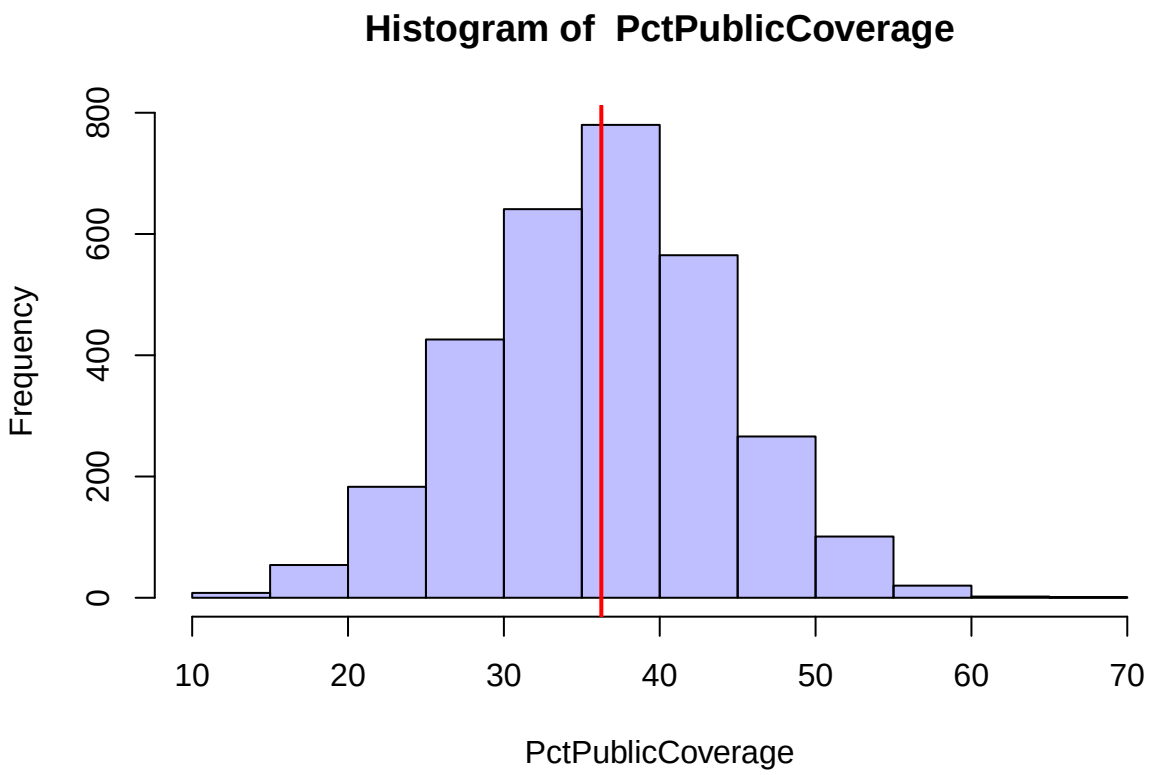
```
histWithMean(PctPrivateCoverage, "PctPrivateCoverage")
```




```
histWithMean(PctEmpPrivCoverage, "PctEmpPrivCoverage")
```



```
histWithMean(PctPublicCoverage, "PctPublicCoverage")
```




```
# some columns were added, so lets update the attached version
detach(Cancer)
attach(Cancer)
```

Get all the numeric type columns, find correlations with deathRate, and sort by descending absolute value

```
numericColumns <- sapply(Cancer, is.numeric)
NumericCancer <- Cancer[, numericColumns]
correlations <- apply(NumericCancer, 2, function(col) cor(col, deathRate))
correlations[order(abs(correlations), decreasing=TRUE)]
```

```
##            deathRate    PctBachDeg25_Over         logMedIncome
##          1.000000000         -0.485477318         -0.452277367
##       povertyPercent            medIncome         PctHS25_Over
##          0.429388980         -0.428614927          0.404589076
##    PctPublicCoverage   PctPrivateCoverage PctUnemployed16_Over
##          0.404571656         -0.386065507          0.378412442
## PctMarriedHouseholds      PctBachDeg18_24   PctEmpPrivCoverage
##         -0.293325341         -0.287817410         -0.267399428
##       PercentMarried           PctHS18_24             PctBlack
##         -0.266820464          0.261975940          0.257023560
##         PctOtherRace              PctAsian             PctWhite
##         -0.189893571         -0.186331105         -0.177399980
##          avgAnnCount           popEst2015         PctNoHS18_24
##         -0.143531620         -0.120073096          0.088462610
```

```
##       logAvgAnnCount            BirthRate        logPopEst2015
##         -0.087968621         -0.087406970         -0.070621122
##                    X     AvgHouseholdSize        MedianAgeMale
##          0.051913500         -0.036905314         -0.021929429
##      MedianAgeFemale            MedianAge      PctSomeCol18_24
##          0.012048386          0.004375077                   NA
##   PctEmployed16_Over
##                   NA
```

Check correlations for the "cleaned" variables

```
cor(AvgHouseholdSize[cleanAvgHouseholdSize], deathRate[cleanAvgHouseholdSize])
```

```
## [1] -0.03464102
```

```
cor(MedianAge[cleanMedianAge], deathRate[cleanMedianAge])
```

```
## [1] -0.004288054
```

```
cor(PctEmployed16_Over[cleanPctEmployed16_Over], deathRate[cleanPctEmployed16_Over]) # -0.4120458
```

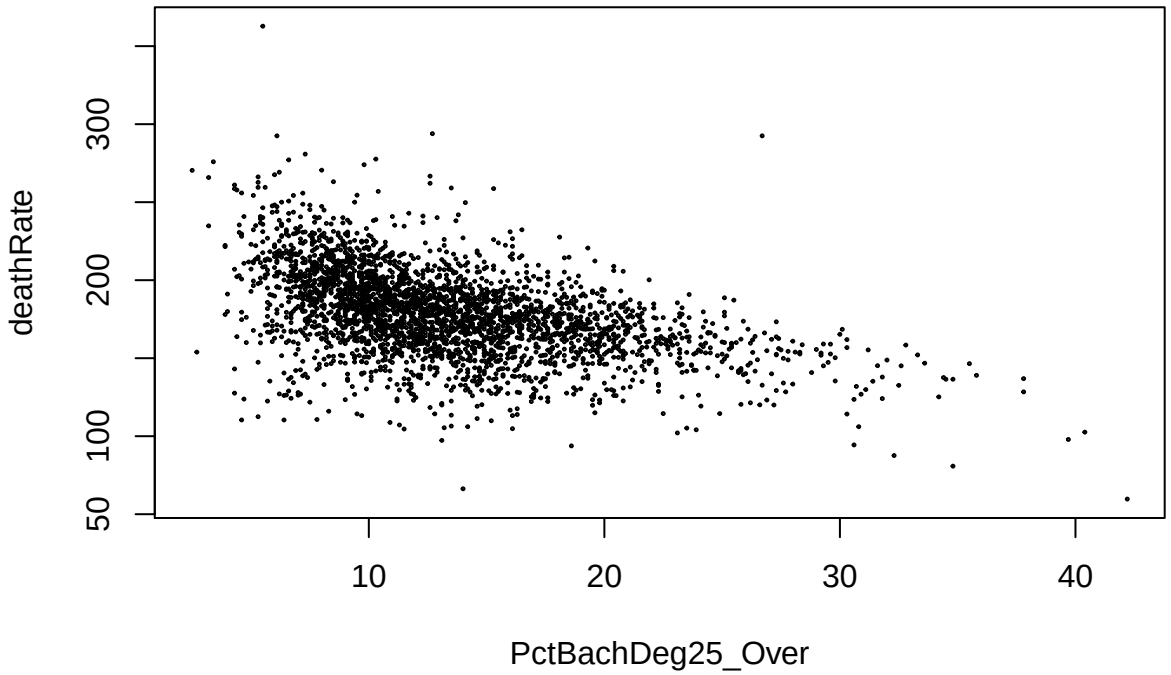
```
## [1] -0.4120458
```

```
cor(PctSomeCol18_24[cleanPctSomeCol18_24], deathRate[cleanPctSomeCol18_24])
```

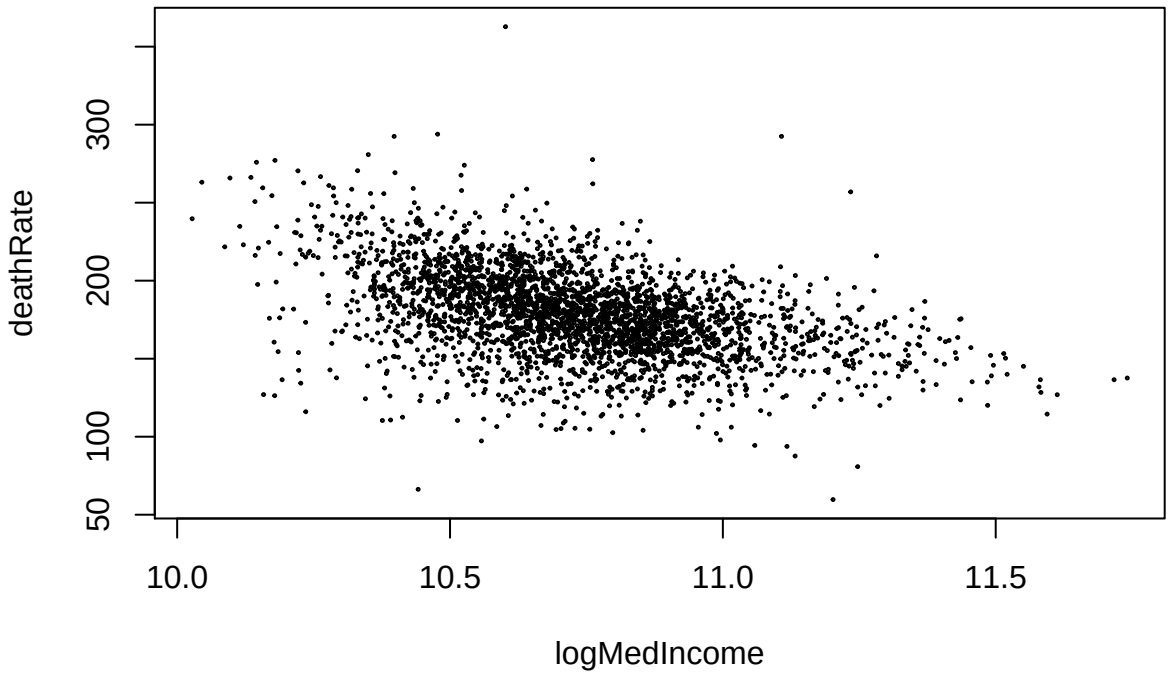
```
## [1] -0.1886877
```

Plot deathRate with all variables with at least a weak correlation

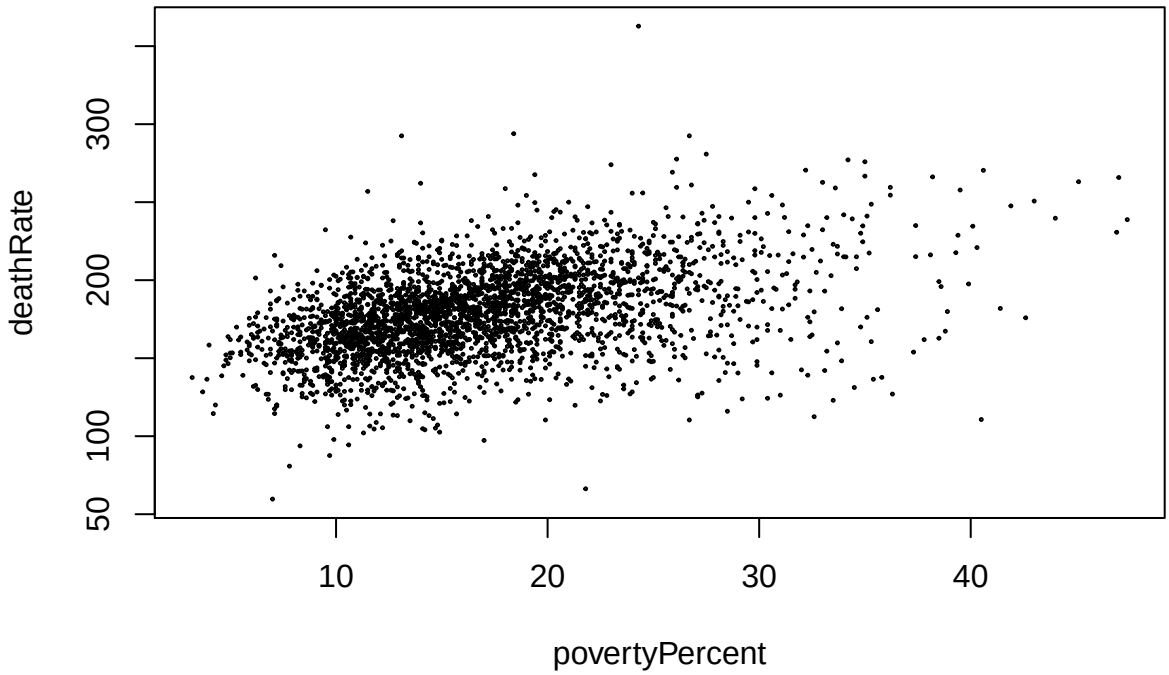
```
plot(PctBachDeg25_Over, deathRate, cex=0.2)
```



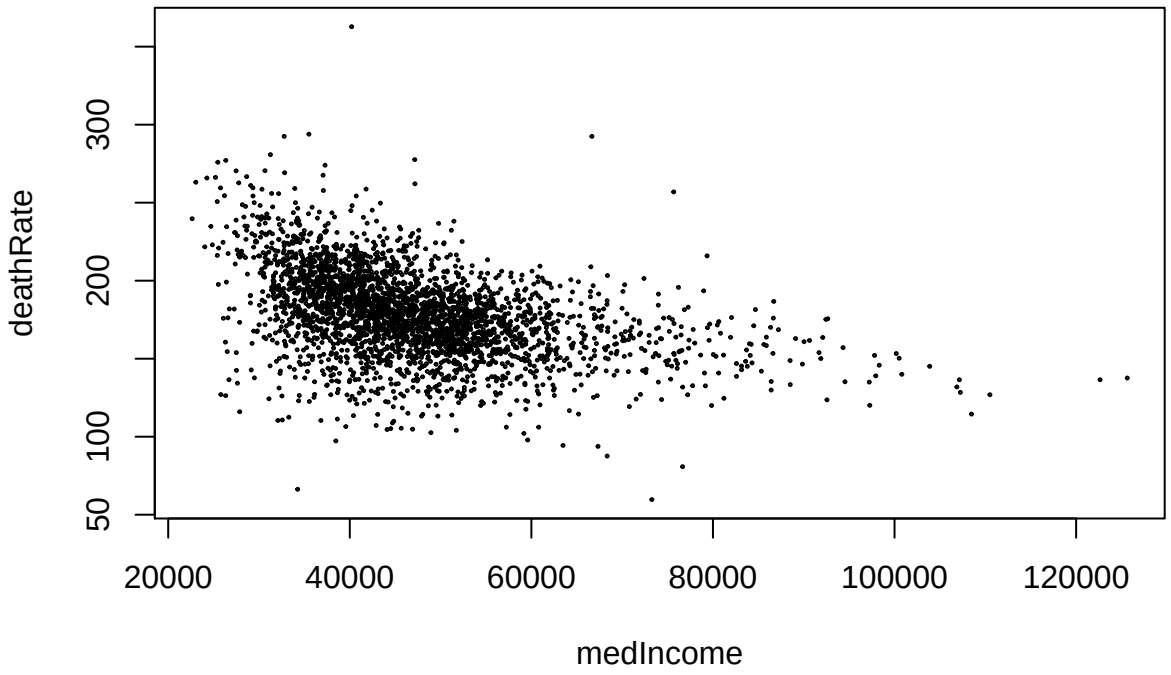

```
plot(logMedIncome, deathRate, cex=0.2)
```



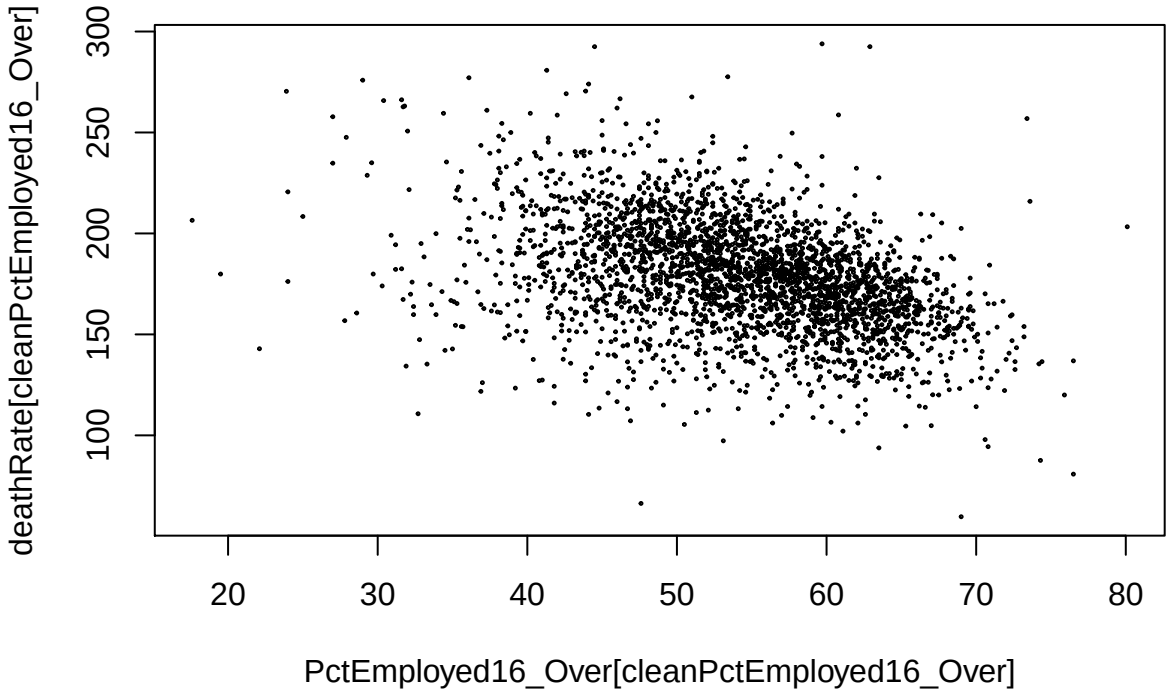

```
plot(povertyPercent, deathRate, cex=0.2)
```



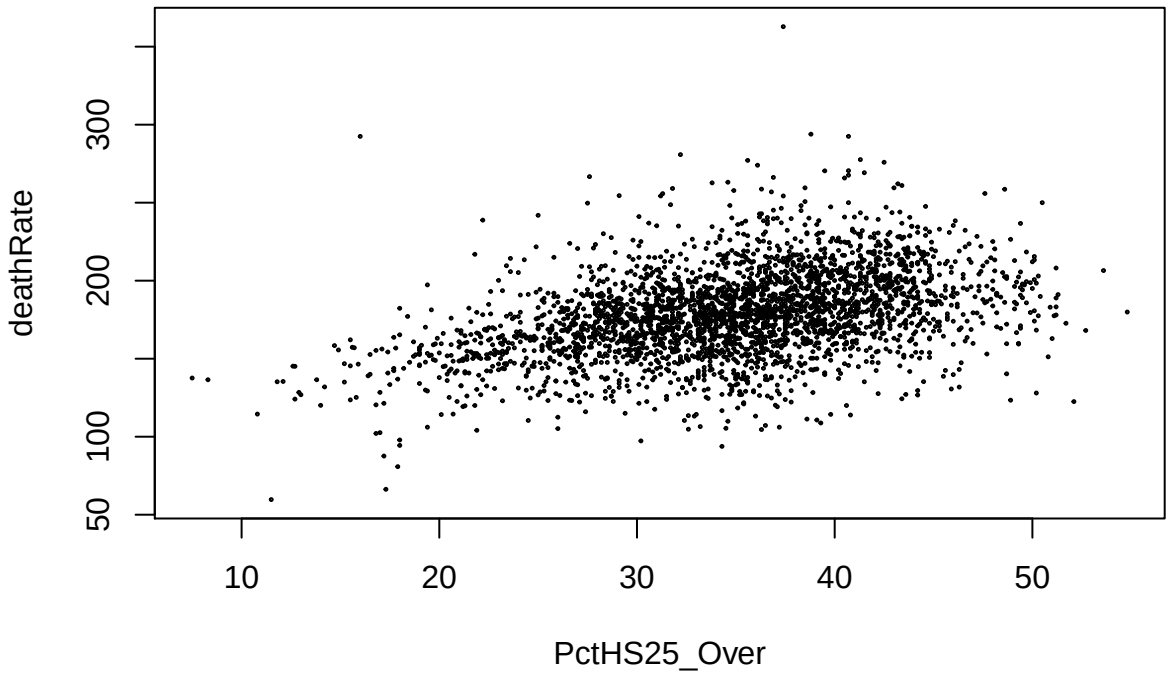

```
plot(medIncome, deathRate, cex=0.2)
```



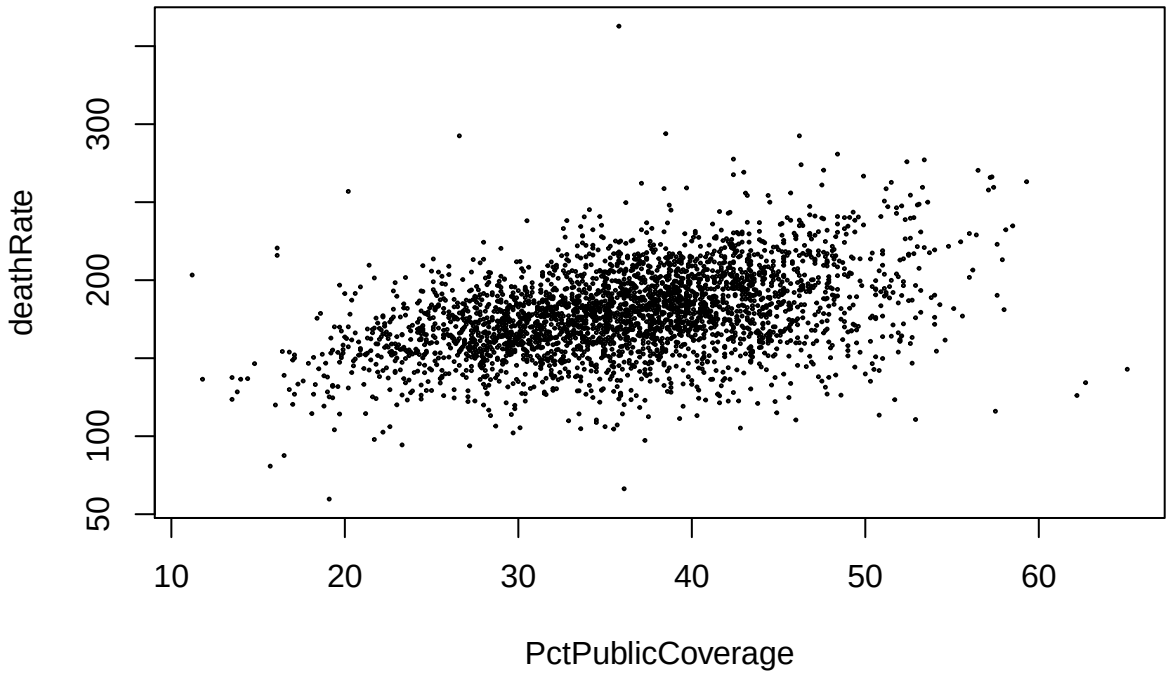

```
plot(PctEmployed16_Over[cleanPctEmployed16_Over], deathRate[cleanPctEmployed16_Over], cex=0.2)
```



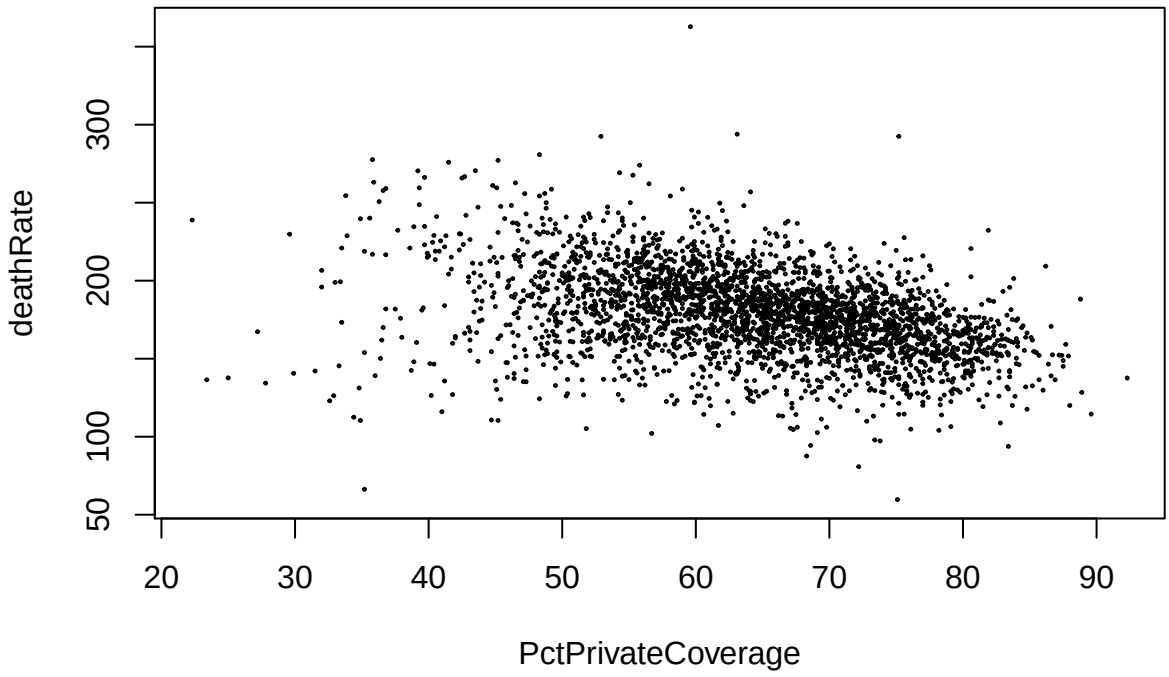

```
plot(PctHS25_Over, deathRate, cex=0.2)
```




```
plot(PctPublicCoverage, deathRate, cex=0.2)
```




```
plot(PctPrivateCoverage, deathRate, cex=0.2)
```




```
plot(PctUnemployed16_Over, deathRate, cex=0.2)
```