Recommended Exercise 2 in Statistical Linear Models, Spring 2021

alexaoh

22.01.2021

Problem 1 Principal Component Analysis (PCA)

```r
#str(USArrests) # Data set included in R.

# a) Find loadings/rotations of the PC's.
pc <- prcomp(USArrests, scale = T)
pc$rotation # Loadings via prcomp() function
```

```
#>                 PC1        PC2        PC3         PC4
#> Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
#> Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
#> UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
#> Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

```r
# Loadings via finding the eigenvalues of correlation
# (since we want scaled variables) matrix of the data.
eigen(cor(USArrests))$vectors
```

```
#>            [,1]       [,2]       [,3]        [,4]
#> [1,] -0.5358995  0.4181809 -0.3412327  0.64922780
#> [2,] -0.5831836  0.1879856 -0.2681484 -0.74340748
#> [3,] -0.2781909 -0.8728062 -0.3780158  0.13387773
#> [4,] -0.5434321 -0.1673186  0.8177779  0.08902432
```

```r
# b) Find sample variance of the PC's.
pc$sdev^2 # Sample variances using prcomp() function.
```

```
#> [1] 2.4802416 0.9897652 0.3565632 0.1734301
```

```r
eigen(cor(USArrests))$values # Sample variances via correlation matrix.
```

```
#> [1] 2.4802416 0.9897652 0.3565632 0.1734301
```

```r
# c) Find the scores and check that the scores for Alabama are indeed
#    the linear combinations of the data for Alabama with the loadings
#    as coefficients.
pc$x # Scores.
```

```
#>                        PC1         PC2         PC3          PC4
#> Alabama        -0.97566045  1.12200121 -0.43980366  0.154696581
#> Alaska         -1.93053788  1.06242692  2.01950027 -0.434175454
#> Arizona        -1.74544285 -0.73845954  0.05423025 -0.826264240
#> Arkansas        0.13999894  1.10854226  0.11342217 -0.180973554
#> California     -2.49861285 -1.52742672  0.59254100 -0.338559240
#> Colorado       -1.49934074 -0.97762966  1.08400162  0.001450164
```

```
#> Connecticut     1.34499236 -1.07798362 -0.63679250 -0.117278736
#> Delaware       -0.04722981 -0.32208890 -0.71141032 -0.873113315
#> Florida        -2.98275967  0.03883425 -0.57103206 -0.095317042
#> Georgia        -1.62280742  1.26608838 -0.33901818  1.065974459
#> Hawaii          0.90348448 -1.55467609  0.05027151  0.893733198
#> Idaho           1.62331903  0.20885253  0.25719021 -0.494087852
#> Illinois       -1.36505197 -0.67498834 -0.67068647 -0.120794916
#> Indiana         0.50038122 -0.15003926  0.22576277  0.420397595
#> Iowa            2.23099579 -0.10300828  0.16291036  0.017379470
#> Kansas          0.78887206 -0.26744941  0.02529648  0.204421034
#> Kentucky        0.74331256  0.94880748 -0.02808429  0.663817237
#> Louisiana      -1.54909076  0.86230011 -0.77560598  0.450157791
#> Maine           2.37274014  0.37260865 -0.06502225 -0.327138529
#> Maryland       -1.74564663  0.42335704 -0.15566968 -0.553450589
#> Massachusetts   0.48128007 -1.45967706 -0.60337172 -0.177793902
#> Michigan       -2.08725025 -0.15383500  0.38100046  0.101343128
#> Minnesota       1.67566951 -0.62590670  0.15153200  0.066640316
#> Mississippi    -0.98647919  2.36973712 -0.73336290  0.213342049
#> Missouri       -0.68978426 -0.26070794  0.37365033  0.223554811
#> Montana         1.17353751  0.53147851  0.24440796  0.122498555
#> Nebraska        1.25291625 -0.19200440  0.17380930  0.015733156
#> Nevada         -2.84550542 -0.76780502  1.15168793  0.311354436
#> New Hampshire   2.35995585 -0.01790055  0.03648498 -0.032804291
#> New Jersey     -0.17974128 -1.43493745 -0.75677041  0.240936580
#> New Mexico     -1.96012351  0.14141308  0.18184598 -0.336121113
#> New York       -1.66566662 -0.81491072 -0.63661186 -0.013348844
#> North Carolina -1.11208808  2.20561081 -0.85489245 -0.944789648
#> North Dakota    2.96215223  0.59309738  0.29824930 -0.251434626
#> Ohio            0.22369436 -0.73477837 -0.03082616  0.469152817
#> Oklahoma        0.30864928 -0.28496113 -0.01515592  0.010228476
#> Oregon         -0.05852787 -0.53596999  0.93038718 -0.235390872
#> Pennsylvania    0.87948680 -0.56536050 -0.39660218  0.355452378
#> Rhode Island    0.85509072 -1.47698328 -1.35617705 -0.607402746
#> South Carolina -1.30744986  1.91397297 -0.29751723 -0.130145378
#> South Dakota    1.96779669  0.81506822  0.38538073 -0.108470512
#> Tennessee      -0.98969377  0.85160534  0.18619262  0.646302674
#> Texas          -1.34151838 -0.40833518 -0.48712332  0.636731051
#> Utah            0.54503180 -1.45671524  0.29077592 -0.081486749
#> Vermont         2.77325613  1.38819435  0.83280797 -0.143433697
#> Virginia        0.09536670  0.19772785  0.01159482  0.209246429
#> Washington      0.21472339 -0.96037394  0.61859067 -0.218628161
#> West Virginia   2.08739306  1.41052627  0.10372163  0.130583080
#> Wisconsin       2.05881199 -0.60512507 -0.13746933  0.182253407
#> Wyoming         0.62310061  0.31778662 -0.23824049 -0.164976866
```

```r
pc$x[1,] # Scores of Alabama.
```

```
#>        PC1        PC2        PC3        PC4
#> -0.9756604  1.1220012 -0.4398037  0.1546966
```

```r
# The calculations below give the same results.
t(pc$rotation) %*% t(scale(USArrests))[,"Alabama"]
```

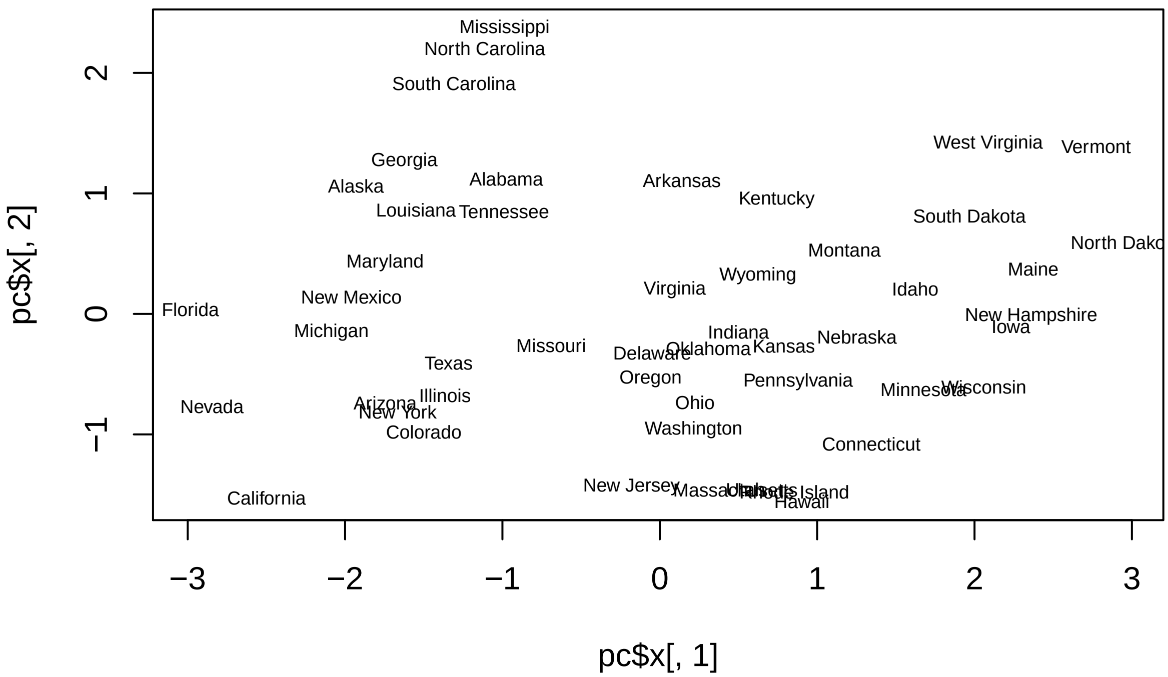
```
#>           [,1]
#> PC1 -0.9756604
```

```
#> PC2  1.1220012
#> PC3 -0.4398037
#> PC4  0.1546966
```

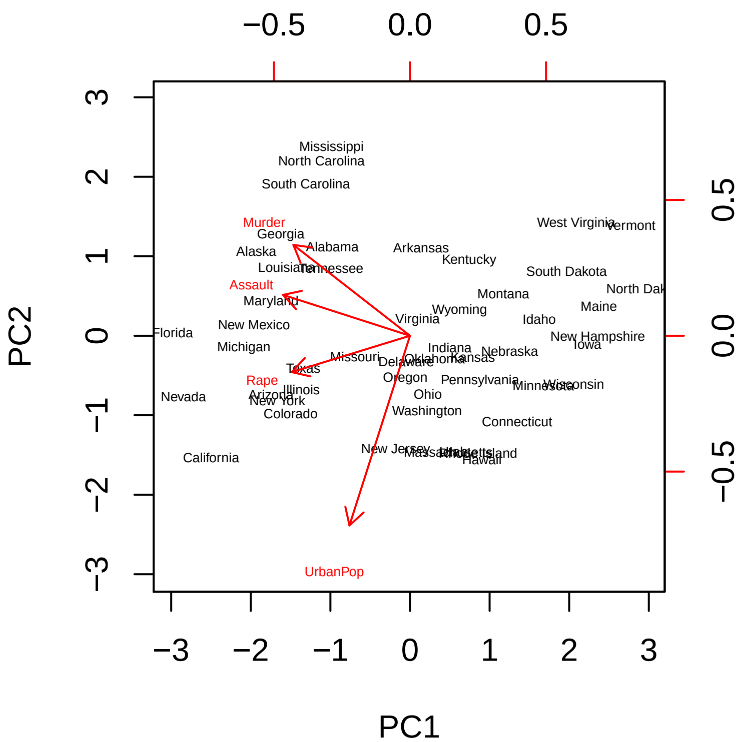
```r
t(pc$rotation) %*% t(data.matrix(scale(USArrests)))[,1]
```

```
#>           [,1]
#> PC1 -0.9756604
#> PC2  1.1220012
#> PC3 -0.4398037
#> PC4  0.1546966
```

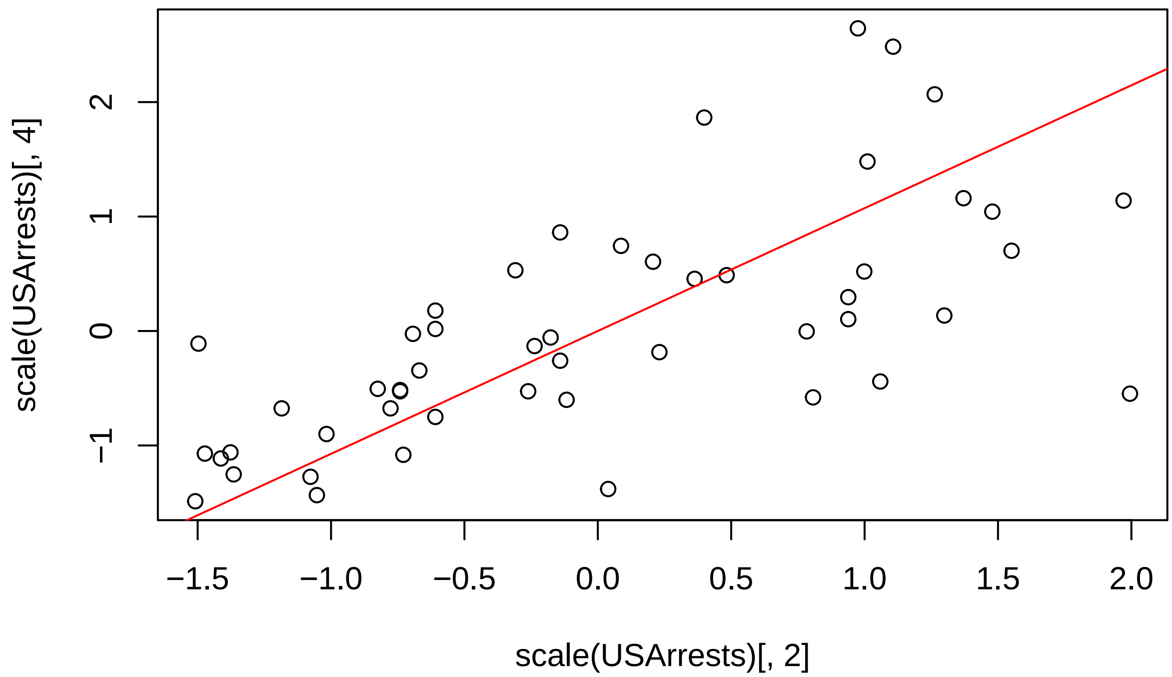
```r
# d) Biplot of the two first PC's.
plot(pc$x[,1],pc$x[,2],type="n")
text(pc$x[,1],pc$x[,2],rownames(USArrests),cex=0.6)
```



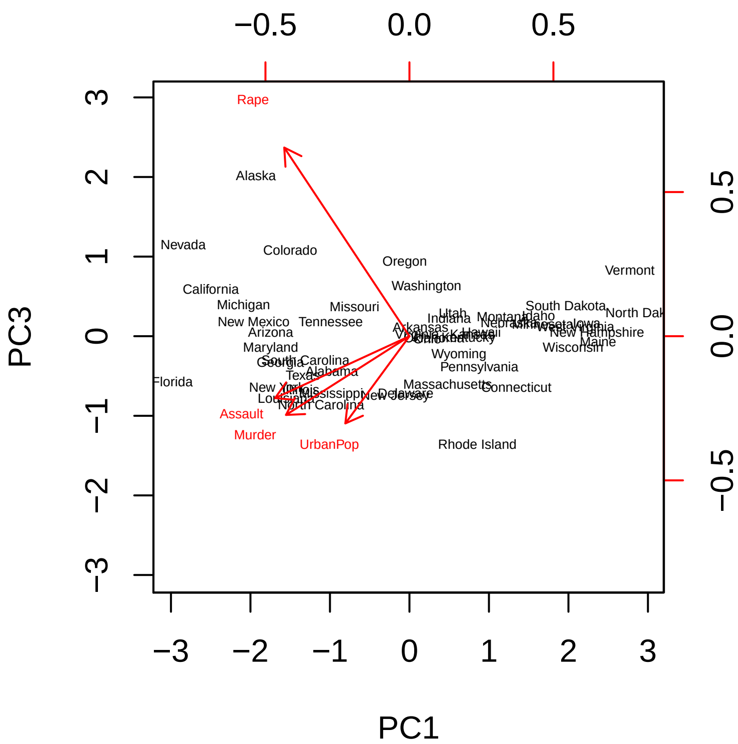

```r
biplot(pc, scale = 0,cex=0.4)
```




```r
plot(scale(USArrests)[, 2], scale(USArrests)[, 4])
abline(0, pc$rotation[2, 1] / pc$rotation[4, 1], col = "red")
```



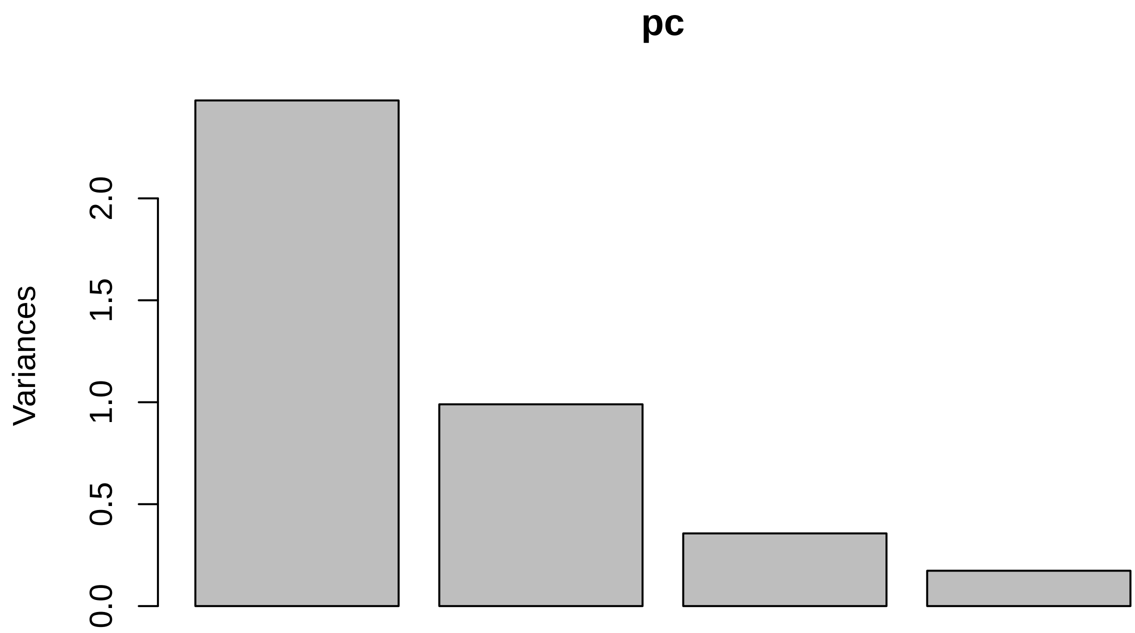

```r
biplot(pc, choices = c(1,3), scale = 0,cex=0.4)
```




```r
# e)
summary(pc)
```

```
#> Importance of components:
#>                           PC1    PC2     PC3     PC4
#> Standard deviation     1.5749 0.9949 0.59713 0.41645
#> Proportion of Variance 0.6201 0.2474 0.08914 0.04336
#> Cumulative Proportion  0.6201 0.8675 0.95664 1.00000
```

```r
plot(pc) # Screeplot.
```




```r
# It is apparent that we need 2 PC's to capture more than 80% of the variability.
```

---

f) All the tasks from above repeated, but without scaling.

```r
# a) Find loadings/rotations of the PC's.
pc.noscale <- prcomp(USArrests, scale = F)
pc.noscale$rotation # Loadings via prcomp() function
```

```
#>                 PC1         PC2         PC3         PC4
#> Murder   0.04170432 -0.04482166  0.07989066 -0.99492173
#> Assault  0.99522128 -0.05876003 -0.06756974  0.03893830
#> UrbanPop 0.04633575  0.97685748 -0.20054629 -0.05816914
#> Rape     0.07515550  0.20071807  0.97408059  0.07232502
```

```r
# Loadings via finding the eigenvalues of correlation
# (since we want scaled variables) matrix of the data.
eigen(cov(USArrests))$vectors
```

```
#>             [,1]        [,2]        [,3]        [,4]
#> [1,] -0.04170432  0.04482166  0.07989066  0.99492173
#> [2,] -0.99522128  0.05876003 -0.06756974 -0.03893830
#> [3,] -0.04633575 -0.97685748 -0.20054629  0.05816914
#> [4,] -0.07515550 -0.20071807  0.97408059 -0.07232502
```

```r
# b) Find sample variance of the PC's.
pc.noscale$sdev^2 # Sample variances using prcomp() function.
```

```
#> [1] 7011.114851  201.992366   42.112651    6.164246
```

```r
eigen(cov(USArrests))$values # Sample variances via correlation matrix.
```

```
#> [1] 7011.114851  201.992366   42.112651    6.164246
```

```r
# c) Find the scores and check that the scores for Alabama are indeed
#    the linear combinations of the data for Alabama with the loadings
#    as coefficients.
pc.noscale$x # Scores.
```

```
#>                       PC1         PC2          PC3        PC4
#> Alabama         64.802164 -11.4480074  -2.49493284 -2.4079009
#> Alaska          92.827450 -17.9829427  20.12657487  4.0940470
#> Arizona        124.068216   8.8304030  -1.68744836  4.3536852
#> Arkansas        18.340035 -16.7039114   0.21018936  0.5209936
#> California     107.422953  22.5200698   6.74587299  2.8118259
#> Colorado        34.975986  13.7195840  12.27936280  1.7214637
#> Connecticut    -60.887282  12.9325302  -8.42065719  0.6999023
#> Delaware        66.731025   1.3537978 -11.28095735  3.7279812
#> Florida        165.244370   6.2746901  -2.99793315 -1.2476807
#> Georgia         40.535177  -7.2902396   3.60952946 -7.3436728
#> Hawaii        -123.536106  24.2912079   3.72444284 -3.4728494
#> Idaho          -51.797002  -9.4691910  -1.52006356  3.3478283
#> Illinois        78.992097  12.8970605  -5.88326477 -0.3676407
#> Indiana        -57.550961   2.8462647   3.73816049 -1.6494302
#> Iowa          -115.586790  -3.3421305  -0.65402935  0.8694960
#> Kansas         -55.789694   3.1572339   0.38436416 -0.6527917
#> Kentucky       -62.383181 -10.6732715   2.23708903 -3.8762164
#> Louisiana       78.277631  -4.2949175  -3.82786965 -4.4835590
#> Maine          -89.261044 -11.4878272  -4.69240562  2.1161995
#> Maryland       129.330136  -5.0070315  -2.34717282  1.9283242
#> Massachusetts  -21.266283  19.4501790  -7.50714835  1.0348189
#> Michigan        85.451527   5.9045567   6.46434210 -0.4990479
#> Minnesota      -98.954816   5.2096006   0.00657376  0.7318957
#> Mississippi     86.856358 -27.4284196  -5.00343624 -3.8797577
#> Missouri         7.986289   5.2756414   5.50057972 -0.6794055
#> Montana        -62.483635  -9.5105021   1.83835536 -0.2459426
#> Nebraska       -69.096544  -0.2111959   0.46802086  0.6565664
#> Nevada          83.613578  15.1021839  15.88869482 -0.3341962
#> New Hampshire -114.777355  -4.7345584  -2.28238693  0.9359106
#> New Jersey     -10.815725  23.1373389  -6.31015739 -1.6124273
#> New Mexico     114.868163  -0.3364531   2.26126996  1.3812478
#> New York        84.294231  15.9239655  -4.72125960 -0.8920194
#> North Carolina 164.325514 -31.0966153 -11.69616350  2.1111927
#> North Dakota  -127.495597 -16.1350394  -1.31182982  2.3009639
#> Ohio           -50.086822  12.2793244   1.65733077 -2.0291157
#> Oklahoma       -19.693723   3.3701310  -0.45314329  0.1803457
#> Oregon         -11.150240   3.8660682   8.12998050  2.9140109
#> Pennsylvania   -64.689142   8.9115466  -3.20646858 -1.8749353
#> Rhode Island     3.063973  18.3739704 -17.47001970  2.3082597
#> South Carolina 107.281069 -23.5361159  -2.03279501 -1.2517463
#> South Dakota   -86.106720 -16.5978586   1.31437998  1.2522874
#> Tennessee       17.506264  -6.5065756   6.10012753 -3.9228558
```

```
#> Texas            31.291122  12.9849566  -0.39340922 -4.2420040
#> Utah            -49.913397  17.6484577   1.78816852  1.8677052
#> Vermont        -124.714469 -27.3135591   4.80277765  2.0049857
#> Virginia        -14.817448  -1.7526150   1.04538813 -1.1738408
#> Washington      -25.075839   9.9679669   4.78112764  2.6910819
#> West Virginia   -91.544647 -22.9528778  -0.40198344 -0.7368781
#> Wisconsin      -118.176328   5.5075792  -2.71132077 -0.2049724
#> Wyoming         -10.434539  -5.9244529  -3.79444682  0.5178674
```

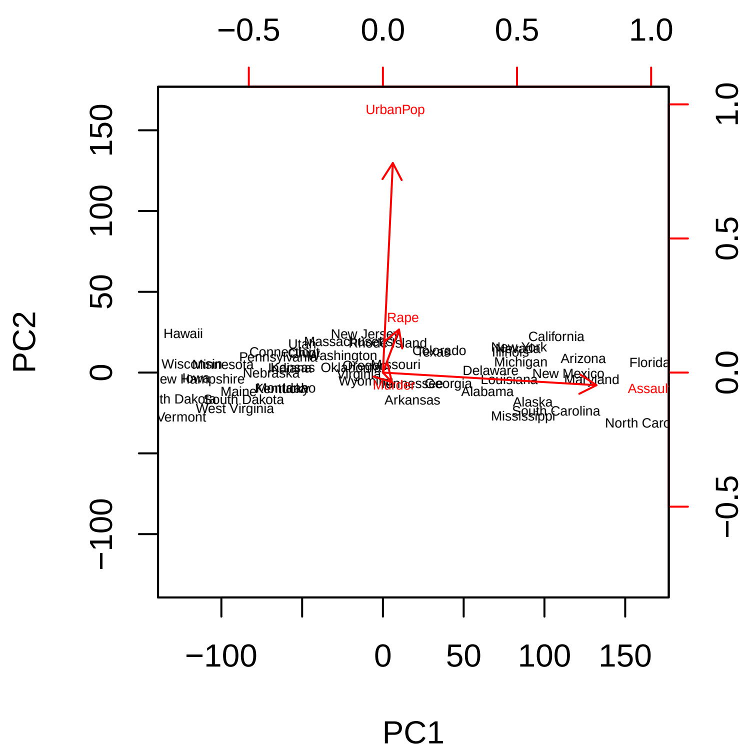
```
pc.noscale$x[1,] # Scores of Alabama.
```

```
#>        PC1        PC2        PC3        PC4
#>  64.802164 -11.448007  -2.494933  -2.407901
```

```
# The calculations below give the same results.
t(pc.noscale$rotation) %*% data.matrix(USArrests)[1,]
```

```
#>           [,1]
#> PC1 239.703489
#> PC2  46.453944
#> PC3  -5.873077
#> PC4  -5.784048
```

```
# d) Biplot of the two first PC's.
biplot(pc.noscale, scale = 0,cex=0.4)
```




```
# e)
summary(pc.noscale)
```

```
#> Importance of components:
#>                            PC1      PC2    PC3     PC4
#> Standard deviation     83.7324 14.21240 6.4894 2.48279
#> Proportion of Variance  0.9655  0.02782 0.0058 0.00085
#> Cumulative Proportion   0.9655  0.99335 0.9991 1.00000
```

```r
# It is apparent that the components with the highest variance dominate (PC1 most significantly).

cov(USArrests) # Here we see which of the variables have the highest variables.
```

```
#>              Murder   Assault   UrbanPop      Rape
#> Murder    18.970465  291.0624   4.386204  22.99141
#> Assault  291.062367 6945.1657 312.275102 519.26906
#> UrbanPop   4.386204  312.2751 209.518776  55.76808
#> Rape      22.991412  519.2691  55.768082  87.72916
```

Problem 4 Normal and chi-squared distributions in R

```r
?rnorm
?pchisq
```

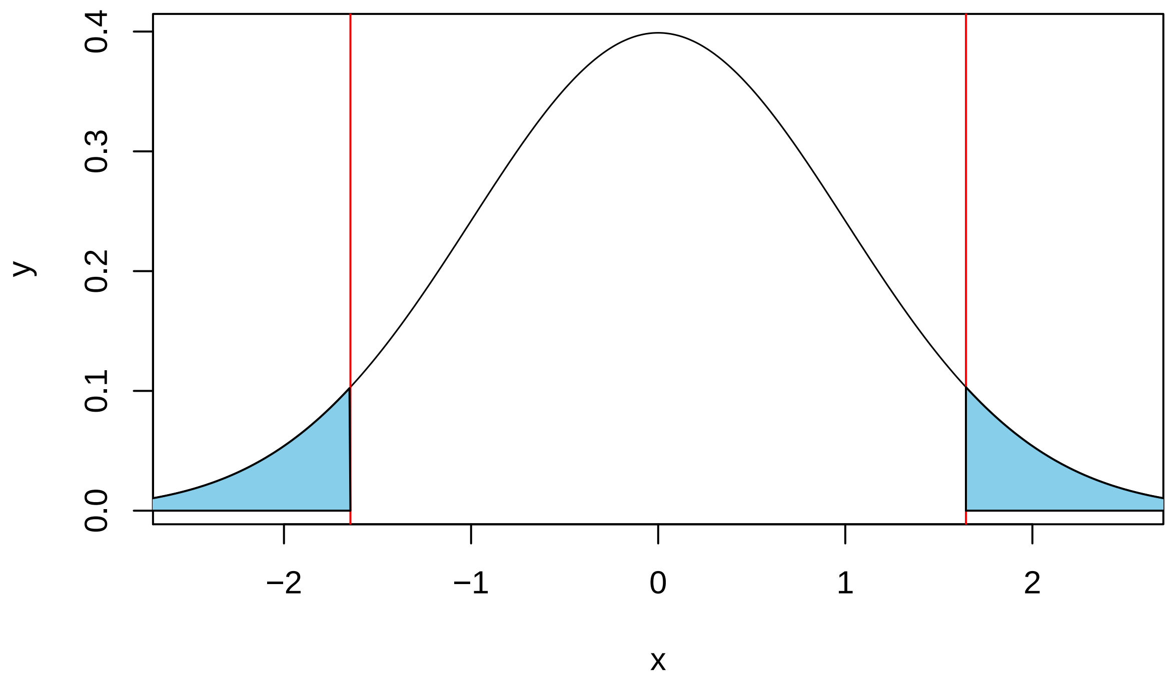
b)

Make a plot of the standard normal pdf.

```r
left.endpoint <- -3
right.endpoint <- 3
x <- seq(left.endpoint, right.endpoint, by = 0.01)
y <- dnorm(x)

plot(x,y, type = "l", lwd = 0.8, xlim = c(left.endpoint+0.5, right.endpoint-0.5))
# Kan heller skrive:
# plot(dnorm, left.endpoint, right.endoint)
abline(v = c(qnorm(0.05), qnorm(0.95)), col = "red")

# Coloring in the tails next.
left.x <- seq(left.endpoint, qnorm(0.05), by = 0.01)
right.x <- seq(qnorm(0.95), right.endpoint, by = 0.01)
left.y <- dnorm(left.x)
right.y <- dnorm(right.x)
polygon(x = c(qnorm(0.95), right.x, right.endpoint) , y = c(0, right.y, 0), col = "skyblue")
polygon(x = c(left.endpoint, left.x, qnorm(0.05)) , y = c(0, left.y, 0), col = "skyblue")
```



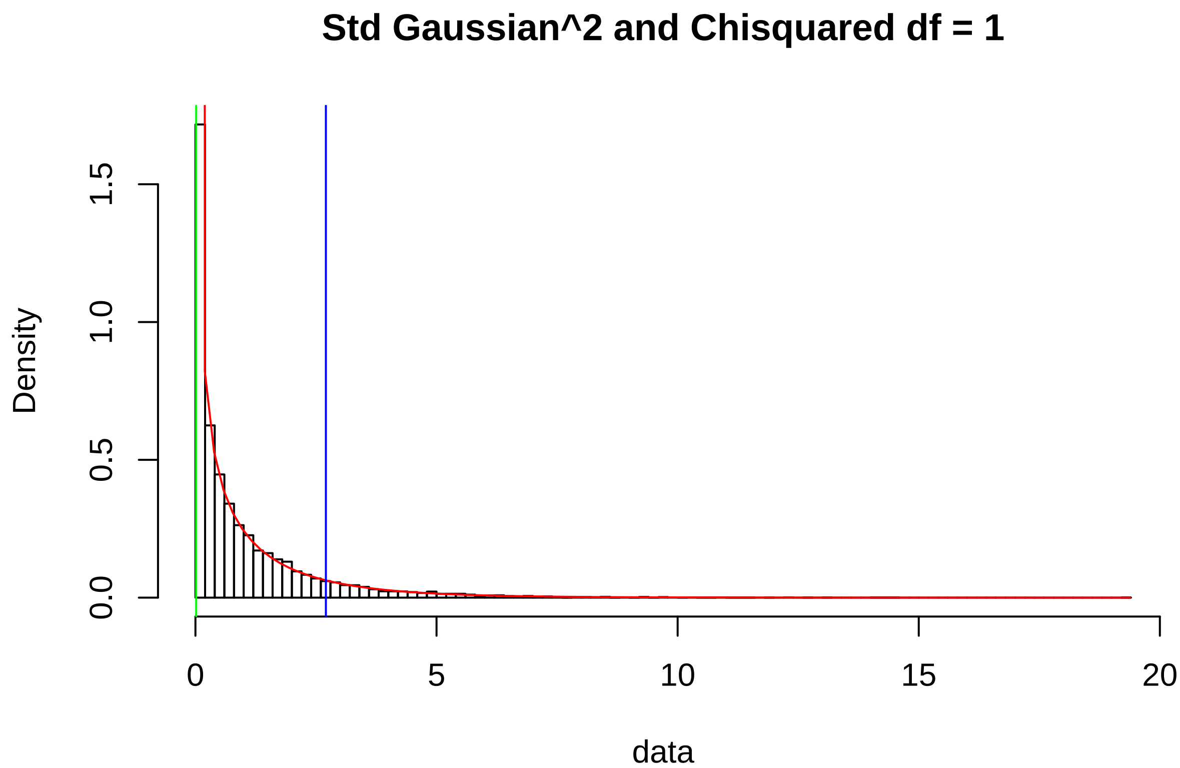
c)

```
data <- rnorm(10000)^2

hist(data, nclass = 100, freq = F, main = "Std Gaussian^2 and Chisquared df = 1")

# Add chi-squared.
plot(function(x) dchisq(x, df = 1), from = min(data), to = max(data), add = TRUE, col = "red")

# Add quantiles.
abline(v = c(qchisq(0.1, df = 1), qchisq(0.9, df = 1)), col = c("green", "blue"))
```


Std Gaussian^2 and Chisquared df = 1
Density
0.0
0.5
1.0
1.5
0
5
10
15
20
data