[9]:
```
 ## Importing packages

# This R environment comes with all of CRAN and many other helpful packag
# You can see which packages are installed by checking out the kaggle/rst
# https://github.com/kaggle/docker-rstats

library(tidyverse) # metapackage with lots of helpful functions

## Running code

# In a notebook, you can run a single code cell by clicking in the cell a
# the blue arrow to the left, or by clicking in the cell and pressing Shi
# you can run code by highlighting the code you want to run and then clic
# at the bottom of this window.

## Reading in files

# You can access files from datasets you've added to this kernel in the "
# You can see the files added to this kernel by running the code below.

list.files(path = "../input")

## Saving data

# If you save any files or images, these will be put in the "output" dire
# can see the output directory by committing and running your kernel (usi
# Commit & Run button) and then checking out the compiled version of your
```

[10]:

```
#
# https://stats.stackexchange.com/questions/223446/variance-of-the-mean-c
# To demonstrate the concept in An Introduction to Statistical Learning,
# p183-184 :
# Since the mean of many highly correlated quantities has higher variance
# does the mean of many quantities that are not as highly correlated,
# the test error estimate resulting from LOOCV tends to have higher varia
# than does the test error estimate resulting from k-fold CV.
#
version=20190204
```

[11]:

```
# n = 2 : 2 dimensional, x1 and x2
# rho : correlation
# n.sim : number of (x1, x2) pairs
n <- 2
rho1 <- 0.45
rho2 <- 0.99
n.sim <- 5e3

#
# Create a data structure for making correlated variables.
#
Sigma1 <- outer(1:n, 1:n, function(i,j) rho1^abs(i-j))
Sigma2 <- outer(1:n, 1:n, function(i,j) rho2^abs(i-j))

S1 <- svd(Sigma1)
S2 <- svd(Sigma2)

Q1 <- S1$v %*% diag(sqrt(S1$d))
Q2 <- S2$v %*% diag(sqrt(S2$d))
#
# Generate two sets of sample means, one uncorrelated (x) and the other c
#
Z0 <- matrix(rnorm(n*n.sim), nrow=n)
Z1<- Q1 %*% Z0
Z2<- Q2 %*% Z0

meanZ0 <- colMeans(Z0)
meanZ1 <- colMeans(Z1)
meanZ2 <- colMeans(Z2)

var0 <- var(meanZ0)
var1 <- var(meanZ1)
var2 <- var(meanZ2)

Z0[1:2,1:8]
Z1[1:2,1:8]
Z2[1:2,1:8]

par(mfrow=c(2,3))

#
# Show scatterplots of the samples.
```

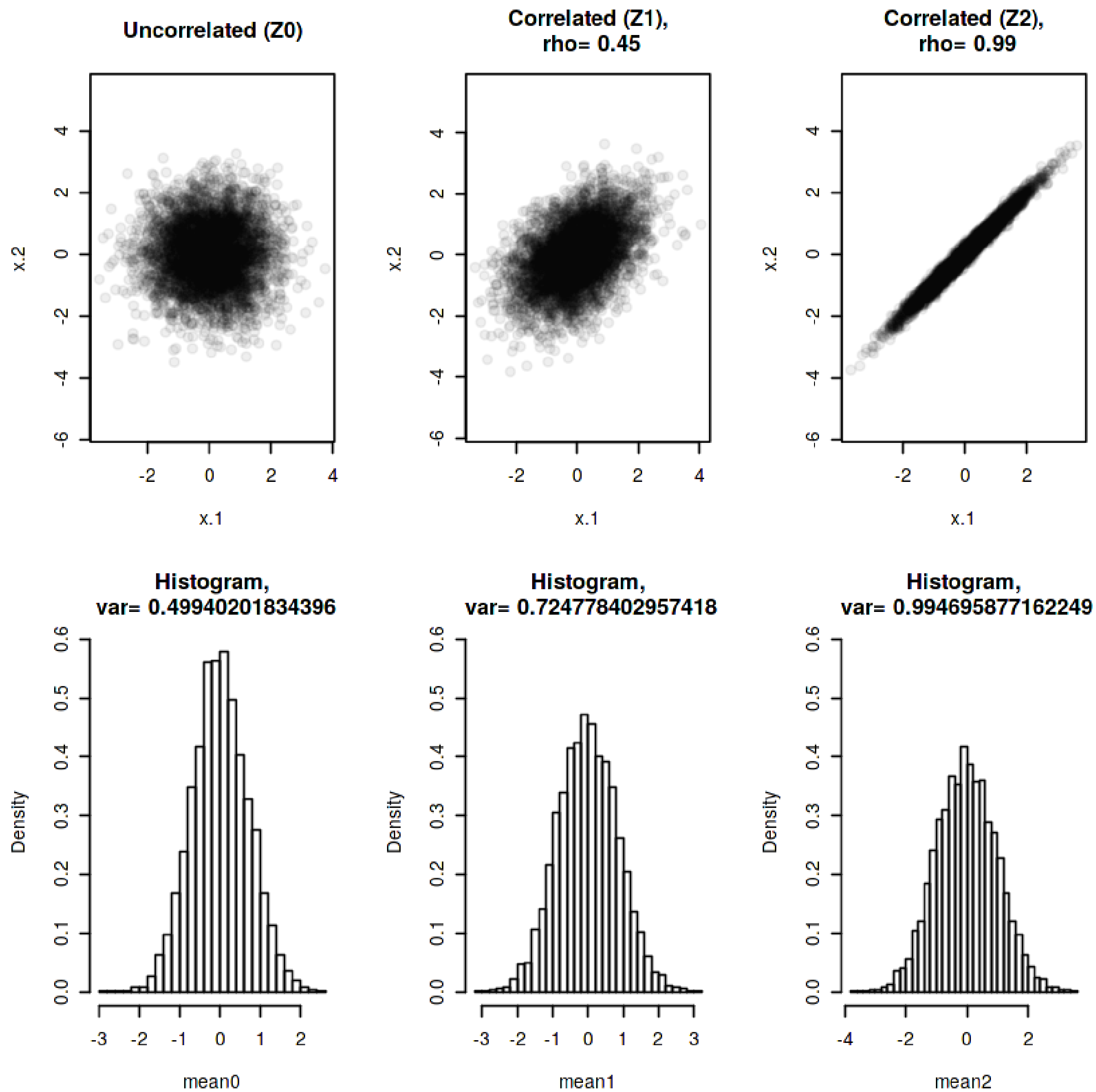
```
# t() : transpose
#
plot(t(Z0)[, 1:2],main=paste("Uncorrelated (Z0)"),
     pch=19, col="#00000010", xlab="x.1", ylab="x.2", asp=1)
plot(t(Z1)[, 1:2], main=paste("Correlated (Z1),\n rho=",rho1),
     pch=19, col="#00000010", xlab="x.1", ylab="x.2", asp=1)
plot(t(Z2)[, 1:2], main=paste("Correlated (Z2),\n rho=",rho2),
     pch=19, col="#00000010", xlab="x.1", ylab="x.2", asp=1)


#
# Display the histograms of both.
#
h.mean1 <- hist(meanZ1, breaks=30, plot=FALSE)
h.mean2 <- hist(meanZ2, breaks=30, plot=FALSE)
#h.yo <- hist(x, breaks=h.y1$breaks, plot=FALSE)
h.mean0 <- hist(meanZ0, breaks=30, plot=FALSE)

ylim <- c(0, max(h.mean0$density))
#ylim<-c(0,10)
hist(meanZ0, xlab = "mean0", main=paste("Histogram,\n var=", var0),
     freq=FALSE, breaks=h.mean0$breaks, ylim=ylim)
hist(meanZ1, xlab = "mean1", main=paste("Histogram,\n var=", var1),
     freq=FALSE, breaks=h.mean1$breaks, ylim=ylim)
hist(meanZ2, xlab = "mean2", main=paste("Histogram,\n var=", var2),
     freq=FALSE, breaks=h.mean2$breaks, ylim=ylim)

#?t
```

0.3361367	-1.0180470	-0.1944533	-1.102179	-0.1482573	0.8452272	0.1562605	0.04992798
0.7720881	0.8459182	-1.0289131	-1.061520	-0.9390530	1.3170451	-0.6587195	-1.48435782

-0.6910965	0.4232325	0.7051376	1.4951376	0.6186801	-1.41034929	0.2123844	0.7358917
0.1186763	1.3104390	-0.3739956	0.3818062	-0.3662070	-0.02902068	-0.4784864	-0.8209159

-0.3898902	0.9556832	0.2667217	1.174481	0.21428729	-0.9362406	-0.1092908	0.05515695

-0.2807005	1.0753141	0.1212114	1.024360	0.08148515	-0.7499823	-0.2024478	-0.15476295



Z0 = (x1,x2) where x1 and x2 are uncorrelated

Z1 = (x1,x2) where x1 and x2 are correlated, degree of correlation = 0.45

Z2 = (x1,x2) where x1 and x2 are correlated, degree of correlation = 0.99

Z0, Z1 and Z2 have 5000 pairs of (x1,x2)

dimension:

Z0[1:2,1:5000]

Z1[1:2,1:5000]

Z2[1:2,1:5000]

meanZ0[1:5000] = mean of Z0

meanZ1[1:5000] = mean of Z1

meanZ2[1:5000] = mean of Z2

Mean of highly correlated data pair Z2 has highest variation.

[ ]: