1 Correlation

Here you will generate several sets of paired data, \( Z \) and \( X \), with known theoretical correlation, and examine the actual correlation, with the aim of understanding what data sets with correlations of (roughly) \(-0.6\), \(-0.3\), \(0\), \(0.3\), \(0.6\), \(0.9\) and \(1\) look like. \( X \) and \( Y \) will be uniformly distributed in \([0, 1]\). Then we create \( Z = k \times X + (1 - k) \times Y \) for various values of \( k \) and study the correlation between \( Z \) and \( X \) as a function of \( k \). It can be shown that \( \text{Corr}(Z, X) = \frac{k}{\sqrt{k^2 + (1 - k)^2}} \).

\[
\text{Corr}(Z, X) = \frac{E(XZ) - E(X)E(Z)}{\sqrt{\text{Var}(X)\text{Var}(Z)}}
\]

Therefore

\[
E(Z) = kE(X) + (1 - k)E(Y) = k \times E(X) + (1 - k)E(X) = E(X)
\]

\[
\text{Var}(Z) = k^2\text{Var}(X) + (1 - k)^2\text{Var}(Y) = \text{Var}(X)(k^2 + (1 - k)^2)
\]

\[
E(XZ) = E(X(kX + (1 - k)Y)) = kE(X^2) + (1 - k)E(X)E(Y) = kE(X^2) + (1 - k)E(X)^2
\]

\[
= k\text{Var}(X) + E(X)^2
\]

Finally

\[
\text{Corr}(Z, X) = \frac{E(XZ) - E(X)E(Z)}{\sqrt{\text{Var}(X)\text{Var}(Z)}} = \frac{k\text{Var}(X) + E(X)^2 - E(X)E(Z)}{\sqrt{\text{Var}(X)^2(k^2 + (1 - k)^2)}}
\]

\[
= \frac{k\text{Var}(X) + E(X)^2 - E(X)^2}{\text{Var}(X)\sqrt{k^2 + (1 - k)^2}}
\]

\[
= \frac{k}{\sqrt{k^2 + (1 - k)^2}}
\]
Does this statement hold for other r.v. \( X \) and \( Y \) that are independent and identically distributed (i.i.d.)?

The R code below generates samples from \((X, Z)\) for given \( k \), i.e. \((X_1, Z_1), (X_2, Z_2), \ldots, (X_n, Z_n)\) and plots the pairs, see Figure 1 for two values of \( k \). It gives you some impression of correlated data. The correlation measures a linear relationship between two r.v.

For the special case of \( k = 1/2 \) the r.v. \( Z \) has triangular distribution, see [http://en.wikipedia.org/wiki/Triangular_distribution](http://en.wikipedia.org/wiki/Triangular_distribution). To see that you may compare the m.g.f. of \( M_Z(t) = M_{kX+(1-k)Y}(t) = M_X(kt)M_Y((1-k)t) \) with that of triangular distribution and parameters \( a = 0 \), \( b = 1 \) and \( c = 1/2 \).

```r
> size<-200
> X<-runif(size)
> Y<-runif(size)
> k.values<-c(1,0.6737,0.4286,0.2393,0,-0.459,-3)
> Z<-matrix(0,size,length(k.values))
> colnames(Z)<-c("k=1","k=0.6737","k=0.4286","k=0.2393","k=0","k=-0.459","k=-3")
> #define function for transformation
> f<-function(x,y,k){k*x+(1-k)*y}
> theoretical.correlations<- k.values/sqrt((k.values)^2+(1-k.values)^2)
> names(theoretical.correlations)<-colnames(Z)
> # the command cor(Z[,i],X) estimates the correlation between
> # X and Z[,i]
> #
> for(i in 1:length(k.values)){
>   Z[,i]<-f(X,Y,k.values[i])
>   plot(X,Z[,i],main=paste("k="\,k.values[i]," and theoretical correlation=",
>     format(theoretical.correlations[i],digits=3),sep=""),xlab="X",ylab="Z")
>   text(0.4,0.8,paste("estimated correlation=",
>     format(cor(Z[,i],X),digits=3),sep=""),col="red")
> }
```

2 Central Limit Theorem

Here we will look at an application of the Central Limit Theorem, which says (roughly speaking) that if you look at the average/sample mean (or the sum) of \( n \) random variables, the distribution of the average (sum) becomes more like a Normal distribution as \( n \) increases. The sample mean seems often very simplistic and useless, however the C.L.T. applies usually
to all parameter estimates for any type of statistical model. For example the C.L.T. also holds for parameter estimates of a linear regression model, or for parameter estimates of time-series models, etc.

Create a Matrix A of size 100 times 200, where each value in the matrix A represents a random number from a mixture distribution, and where each column represents a sample of size 100. We have here 200 samples of samples, to get an idea of the distribution of the sample mean. Why do we do this? If we obtain 1 sample only, we obtain one value of the sample mean. One value cannot be used to get an idea/approximation of the true distribution. Therefore we generate 200 samples and from each sample we obtain one sample mean. In the end, we have 200 values of the sample mean and these can be used to estimate the density using a histogram, or \texttt{density()}. We will start with 200 samples of size \( k = 1 \) and will finish with 200 samples of size \( k = 100 \).

A mixture distribution is a linear combination of densities. Let \( f_i(x) = f_i(x, \mu_i, \sigma_i^2), i = 1, 2 \) be Gaussian density and let \( p \) be a mixing probability. Then a Gaussian mixture density with two components is

\[
f(x) = pf_1(x) + (1 - p)f_2(x)
\]

```r
> size<-200
> n<-100
> #x represents values of the domain, outside of this range the density is eff. zero
> x<-seq(-10,20,by=0.01)
> #we define density of mixture as density of normal
> dmixnorm<-function(x,mu1,mu2,sigma1,sigma2,p){
```

Figure 1: Left: \( k = 1 \), Right: \( k = -0.459 \)
We fix the parameters of the mixture distribution, namely $\mu_1$, $\sigma_1$, $\mu_2$, $\sigma_2^2$ and $p$. Then we plot the true density function and the

```r
# fixing parameters of mixture
> mu1<-2 # mean of first normal
> mu2<-10 # mean of 2nd normal
> sigma1<-1 # var of 1st normal
> sigma2<-2 # var of 2nd normal
> p<-0.3 # mixture proportion
> # plot true density
> plot(x, dmixnorm(x, mu1=mu1, mu2=mu2, sigma1=sigma1, sigma2=sigma2, p=p), type="l",
+     xlab="x", ylab="f(x)\), xlim=c(-5,20), ylim=c(0,0.14), main="pdf of mixture")
> # plot estimated density
> plot(density(rmixnorm(size, mu1=mu1, mu2=mu2, sigma1=sigma1, sigma2=sigma2, p=p)),
+     main="estimated density", xlim=c(-5,20), ylim=c(0,0.14))
> # A contains 200 samples of size 100 each
> dat<-rmixnorm(n*size, mu1=mu1, mu2=mu2, sigma1=sigma1, sigma2=sigma2, p=p)
> A<-matrix(dat, n, size) # fill matrix with dat

> # we show now the histogram, the true density and the estimated density
> # sample size for each of the 200 sample start at 1 and finish at 100
> for(i in 1:n){
+   row.means<-apply(matrix(A[1:i,], i, size), 2, mean)
+   flush.console() # that command is necessary to update a single plot
+   # otherwise (in RStudio) each time a new plot will be generated
```
Figure 2: Left: True mixture density, Right: Estimated mixture density based on sample size 200

```r
h <- hist(row.means, breaks=20, main=paste("Histogram of row.means for the first ", i, " rows", sep=""), xlab="x", xlim=c(0,15))
maximum.height.histogram <- max(h$counts)

# plot estimated density into histogram
res <- density(row.means)
lines(res$y, maximum.height.histogram * res$x, col="red")

# note the first moment of mixture is sum_i p_i E(X_i) if X_i is ith r.v.
# variance of mixture is sum_i p_i^2 Var(X_i),
# variance of sample mean is var/n

# plot normal density into histogram
lines(x, maximum.height.histogram * dnorm(x, mean=p*mu1+(1-p)*mu2, 
sd=sqrt((p^2*sigma1^2+(1-p)^2*sigma2^2)/i)), col="blue")

legend("topleft", c("histogram", "scaled estimated density", "normal distribution"), 
col=c("black", "red", "blue"), lty=c(1,1,1))

# wait 1 second before proceeding with next i, i.e. i+1
Sys.sleep(1)
```
Figure 3: Left: $k = 10$, Right: $k = 80$