MATH 556: MATHEMATICAL STATISTICS I

BASIC EXCHANGEABILITY CONSTRUCTIONS

An infinite sequence of random variable $X_1, X_2, \ldots, X_n, \ldots$ is *exchangeable* (or *infinitely exchangeable*) if, for any $n \geq 1$ and sets $A_1, A_2, \ldots, A_n \subseteq \mathbb{R}$ we have that

$$P_{X_1,\ldots,X_n}\left[\bigcap_{j=1}^{n}(X_j \in A_j)\right] = P_{X_{\tau(1)},\ldots,X_{\tau(n)}}\left[\bigcap_{j=1}^{n}(X_{\tau(j)} \in A_j)\right]$$

for all permutations $(\tau(1), \ldots, \tau(n))$ of the labels $(1, \ldots, n)$. In terms of cdfs, we can express this as that for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_{\tau(1)},\ldots,X_{\tau(n)}}(x_1, \ldots, x_n).$$

We have the following characterization: the infinite sequence $X_1, X_2, \ldots, X_n, \ldots$ is exchangeable if and only if the representation

$$P_{X_1,\ldots,X_n}\left[\bigcap_{j=1}^{n}(X_j \in A_j)\right] = \int \left\{\prod_{j=1}^{n} P_{X_j|T}[X_j \in A_j | T = t]\right\} dF_T(t)$$

holds for some other random variable T with distribution F_T. That is, the sequence is exchangeable if and only if elements in the sequence are conditionally independent given T, for some T with distribution F_T. In fact, T is a random variable formed as some function of $(X_1, \ldots, X_n)$ in the limiting case as $n \longrightarrow \infty$.

The representation also indicates that we can construct exchangeable random variables by following the construction

$$T \sim f_T(t)$$

$$X_1, \ldots, X_n \sim f_{X|T}(x|t) \qquad \text{independent}$$

EXAMPLE: Suppose $T \sim Uniform(0, 1)$, and $X_1, \ldots, X_n | T = t \sim Bernoulli(t)$ independently. Then

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \int_0^1 \prod_{j=1}^{n} f_{X_j|T}(x_j|t) f_T(t)\, dt = \int_0^1 t^s (1-t)^{n-s}\, dt = \frac{\Gamma(s+1)\Gamma(n-s+1)}{\Gamma(n+2)}$$

where $s = \sum_{j=1}^{n} x_j$, for $s = 0, 1, \ldots, n$, where the support of the joint pmf is the set $\{0, 1\}^n$ of binary vectors of length n. The integral is analytically tractable as the integrand is proportional to a $Beta(s+1, n-s+1)$ pdf. Note that in this construction, the quantity s is associated with a corresponding random variable

$$S = \sum_{j=1}^{n} X_i$$

which we can consider a *summary statistic*, and notice that the event $S = s$ corresponds to

$$\binom{n}{s}$$

individual sequences of x values which all have the same joint probability: this demonstrates exchangeability. Thus

$$f_S(s) = \binom{n}{s}\frac{\Gamma(s+1)\Gamma(n-s+1)}{\Gamma(n+2)} = \frac{n!}{s!(n-s)!}\frac{s!(n-s)!}{(n+1)!} = \frac{1}{n+1} \qquad s = 0, 1, \ldots, n.$$

and zero otherwise.

```
n<-10
s<-0:n
fs<-choose(n,s)*gamma(s+1)*gamma(n-s+1)/gamma(n+2)
fs
```

```
+  [1] 0.09090909 0.09090909 0.09090909 0.09090909 0.09090909 0.09090909
+  [7] 0.09090909 0.09090909 0.09090909 0.09090909 0.09090909

sum(fs)

+ [1] 1

sim.exch01<-function(nv){  #Sample the exchangeable binary variables.
    Tv<-runif(1)
    Xv<-rbinom(nv,1,Tv)
}
svals<-replicate(10000,sum(sim.exch01(n)))  #10000 replicate draws of S
table(svals)/10000

+ svals
+      0      1      2      3      4      5      6      7      8      9
+ 0.0917 0.0973 0.0945 0.0869 0.0917 0.0911 0.0940 0.0879 0.0909 0.0891
+     10
+ 0.0849
```

EXAMPLE: Suppose $T \sim Normal(0,1)$, and $X_1, \ldots, X_n | T = t \sim Normal(t, 1)$ independently. Then

$$\begin{aligned} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) &= \int_{-\infty}^{\infty} \prod_{j=1}^{n} f_{X_j|T}(x_j|t) f_T(t)\, dt \\ &= \int_{-\infty}^{\infty} \prod_{j=1}^{n} \left\{ \left(\frac{1}{2\pi}\right)^{1/2} \exp\left\{-\frac{1}{2}(x_j - t)^2\right\} \right\} \left(\frac{1}{2\pi}\right)^{1/2} \exp\left\{-\frac{1}{2}t^2\right\} dt \\ &= \left(\frac{1}{2\pi}\right)^{(n+1)/2} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[\sum_{j=1}^{n}(x_j - t)^2 + t^2\right]\right\} dt. \end{aligned}$$

Now, using the completing the square formula

$$A(t-a)^2 + B(t-b)^2 = (A+B)\left(t - \frac{Aa+Bb}{A+B}\right)^2 + \frac{AB}{A+B}(a-b)^2$$

we have

$$\sum_{j=1}^{n}(x_j - t)^2 + t^2 = \sum_{j=1}^{n}(x_j - \overline{x})^2 + (n+1)\left(t - \frac{n\overline{x}}{n+1}\right)^2 + \frac{n}{n+1}\overline{x}^2$$

so therefore, we have

$$\begin{aligned} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[\sum_{j=1}^{n}(x_j - t)^2 + t^2\right]\right\} dt &= \exp\left\{-\frac{1}{2}\left[\sum_{j=1}^{n}(x_j - \overline{x})^2 + \frac{n}{n+1}\overline{x}^2\right]\right\} \int_{-\infty}^{\infty} \exp\left\{-\frac{(n+1)}{2}\left(t - \frac{n\overline{x}}{n+1}\right)^2\right\} dt \\ &= \exp\left\{-\frac{1}{2}\left[\sum_{j=1}^{n}(x_j - \overline{x})^2 + \frac{n}{n+1}\overline{x}^2\right]\right\} \sqrt{\frac{2\pi}{n+1}} \end{aligned}$$

as the integrand is proportional to a Normal pdf. Thus for $(x_1, \ldots, x_n) \in R^n$,

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \left(\frac{1}{2\pi}\right)^{n/2} \sqrt{\frac{1}{n+1}} \exp\left\{-\frac{1}{2}\left[\sum_{j=1}^{n}(x_j - \overline{x})^2 + \frac{n}{n+1}\overline{x}^2\right]\right\}$$

which also relies only upon the summary statistics

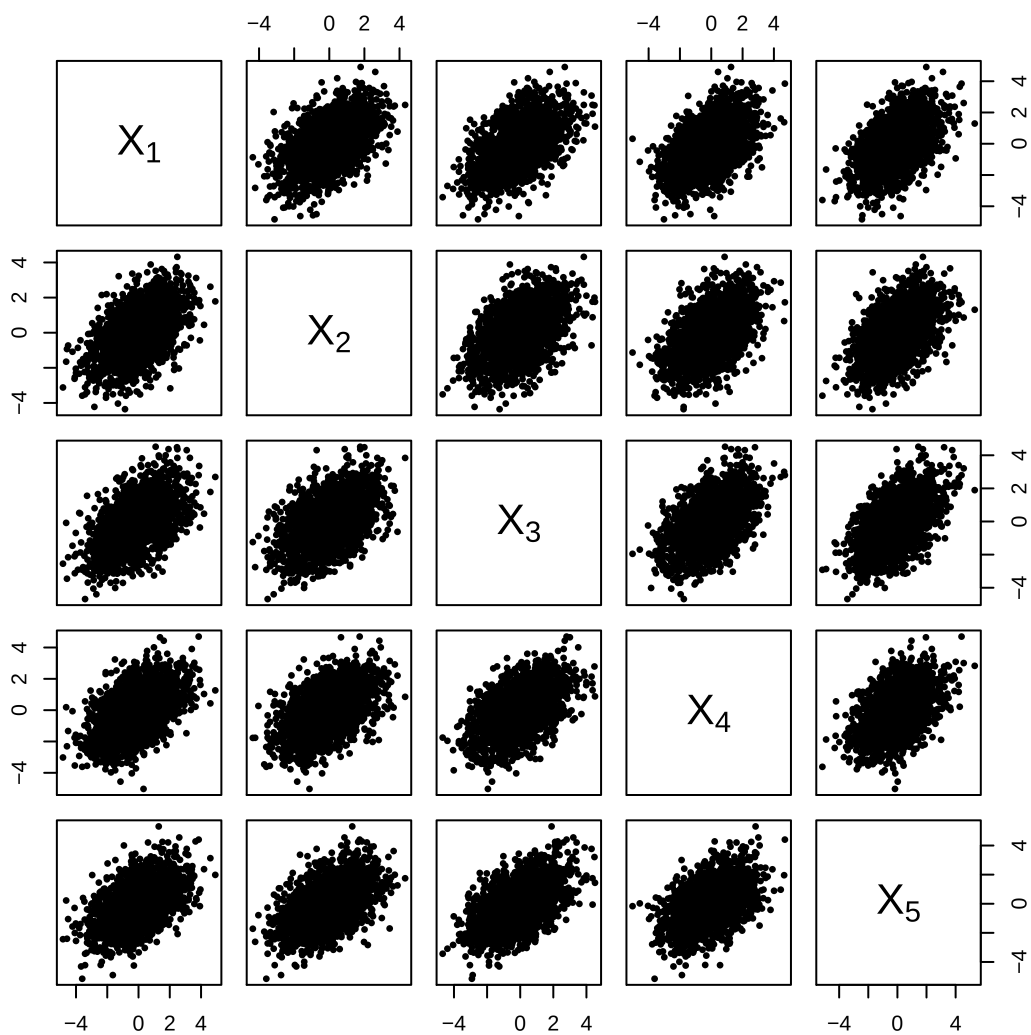
$$S_1 = \overline{x} \qquad S_2 = \sum_{j=1}^{n}(x_j - \overline{x})^2$$

and so we observe exchangeability.

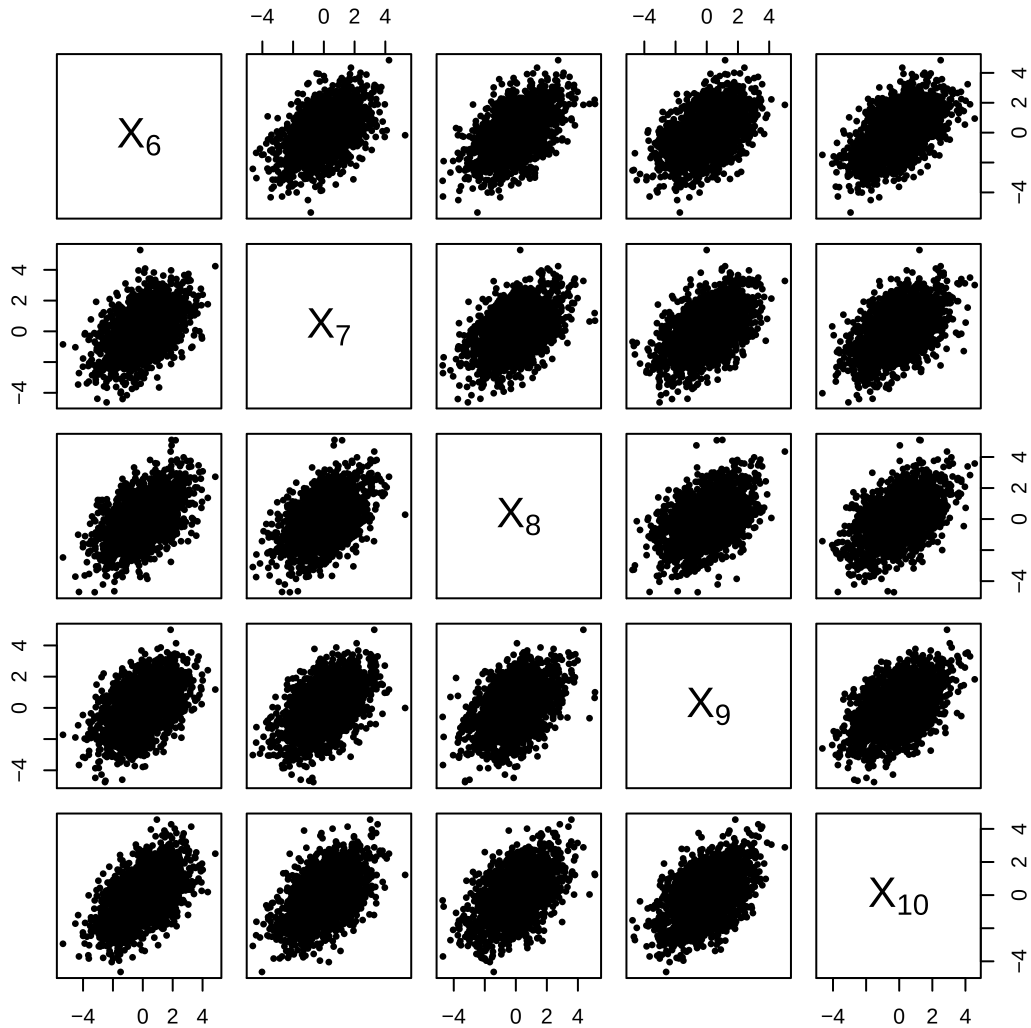
```
n<-10
sim.exch02<-function(nv){  #Sample the exchangeable variables.
    Tv<-rnorm(1)
    Xv<-rnorm(nv,Tv,1)
}
Xmat<-t(replicate(2000,sim.exch02(n)))  #10000 replicate draws of S
dim(Xmat)

+ [1] 2000   10

par(pty='s')
pairs(Xmat[,1:5],pch=19,cex=0.5,
      labels=c(expression(X[1]),expression(X[2]),expression(X[3]),
               expression(X[4]),expression(X[5])))
```




```
pairs(Xmat[,6:10],pch=19,cex=0.5,
      labels=c(expression(X[6]),expression(X[7]),expression(X[8]),
               expression(X[9]),expression(X[10])))
```



We have that for $j = 1, \ldots, n$,

$$\mathbb{E}_{X_j}[X_j] = 0 \qquad \text{Var}_{X_j}[X_j] = 2$$

```
apply(Xmat,2,mean)

+  [1] -0.039762526 -0.052441837 -0.019957830 -0.031307409 -0.016080719
+  [6] -0.032508626  0.012736473  0.005388054  0.004598635 -0.050503142

apply(Xmat,2,var)

+  [1] 2.088259 1.868450 2.002085 2.024241 1.999073 2.053981 1.942118
+  [8] 1.937480 2.034434 1.890146
```

Also, for the covariances, using iterated expectation we have

$$\text{Cov}_{X_j,X_k}[X_j, X_k] \equiv \mathbb{E}_{X_j,X_k}[X_j X_k] = \mathbb{E}_T\left[\mathbb{E}_{X_j,X_k|T}[X_j X_k|T]\right] = \mathbb{E}_T\left[\mathbb{E}_{X_j|T}[X_j|T]\mathbb{E}_{X_k|T}[X_k|T]\right]$$

as X_j and X_k have expectation zero, and are conditionally independent given T. Thus, as $\mathbb{E}_{X_j|T}[X_j|T] = T$ for each j, we have

$$\text{Cov}_{X_j,X_k}[X_j, X_k] = \mathbb{E}_T[T^2] = 1$$

and hence

$$\text{Corr}_{X_j,X_k}[X_j, X_k] = \frac{\text{Cov}_{X_j,X_k}[X_j, X_k]}{\sqrt{\text{Var}_{X_j}[X_j]\text{Var}_{X_k}[X_k]}} = \frac{1}{2}.$$

```
round(cor(Xmat),3)

+        [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10]
+  [1,] 1.000 0.503 0.525 0.488 0.492 0.495 0.504 0.489 0.495 0.507
+  [2,] 0.503 1.000 0.497 0.498 0.483 0.488 0.498 0.471 0.504 0.485
+  [3,] 0.525 0.497 1.000 0.506 0.509 0.526 0.478 0.491 0.482 0.482
+  [4,] 0.488 0.498 0.506 1.000 0.464 0.479 0.489 0.489 0.485 0.498
+  [5,] 0.492 0.483 0.509 0.464 1.000 0.484 0.476 0.457 0.488 0.469
+  [6,] 0.495 0.488 0.526 0.479 0.484 1.000 0.474 0.505 0.490 0.514
+  [7,] 0.504 0.498 0.478 0.489 0.476 0.474 1.000 0.485 0.485 0.498
+  [8,] 0.489 0.471 0.491 0.489 0.457 0.505 0.485 1.000 0.475 0.498
+  [9,] 0.495 0.504 0.482 0.485 0.488 0.490 0.485 0.475 1.000 0.492
+ [10,] 0.507 0.485 0.482 0.498 0.469 0.514 0.498 0.498 0.492 1.000
```