CIS 563 HW1 Lichen Liang

1. a) with replacement: $\binom{10}{4} \cdot 0.4^4 \cdot 0.6^6 = 0.25$

without replacement: $\frac{\binom{40}{4}\binom{60}{6}}{\binom{100}{10}} = 0.264$

b) Since the data is unevenly distributed and we are using stratified sampling, then we should use without replacement. The possibility for without replacement is also higher than with replacement.

c) Sometimes data is difficult to be classified into certain strata (i.e. there is an overlap between strata) and it is time consuming since we need to know the distribution of the data then classify and sample.

2. a) Assume $x_1 = (1, 1, 1 \ldots, 1)$ and $x_2 = (0, 0, 0 \ldots, 0)$ in d dimensions, the max euclidean distance is $\sqrt{(1-0)^2 + (1-0)^2 \ldots} = \sqrt{d}$

The minimum distance is 0 in case 2 points are the same

b) As d increases, the sparsity of a point decreases and $\gamma(d, n)$ decrease to 0, meaning the ratio between the difference between max and min, and min is roughly 1.

3.a)

x	y	z
1	2	3
4	5	6

$E(x) = (1+4)/2 = 2.5$

$E(y) = (2+5)/2 = 3.5$

$E(z) = (3+6)/2 = 4.5$

$Cov(x,y) = Cov(y,x) = E(xy) - E(x)E(y)$

$= \left(\frac{1\cdot 2 + 4\cdot 5}{2}\right) - 2.5(3.5)$

$= 2.25$

$Cov(x,z) = Cov(z,x) = E(xz) - E(x)E(z) = 2.25$

$Cov(y,z) = Cov(z,y) = E(yz) - E(y)E(z) = 2.25$

$Cov(x,x) = E(x^2) - E(x)^2$

$= \left(\frac{1^2+4^2}{2}\right) - (2.5)^2 = 2.25$

$Cov(y,y) = E(y^2) - E(y)^2 = 2.25$

$Cov(z,z) = E(z^2) - E(z)^2 = 2.25$

	x	y	z
x	2.25	2.25	2.25
y	2.25	2.25	2.25
z	2.25	2.25	2.25

Covariance matrix

b) $\det \begin{bmatrix} 2.25-\lambda & 2.25 & 2.25 \\ 2.25 & 2.25-\lambda & 2.25 \\ 2.25 & 2.25 & 2.25-\lambda \end{bmatrix} = 0$

$\Rightarrow (2.25-\lambda)\left[(2.25-\lambda)^2 - (2.25)^2\right] + (2.25)\left[2.25(2.25-\lambda) - (2.25)^2\right]$

$+ 2.25\left[(2.25)^2 - 2.25(2.25-\lambda)\right] = 0$

$\Rightarrow -\lambda^3 + 6.75\lambda^2 = 0 \quad \lambda = 0, 0, 6.75$

Having zero eigenvalue means there is no variance in the data then we should discard it.

c) If eigenvalues are the same, then it would be difficult to decide on which principal component to use.