Name: Chaoyang Zhu UIN: 930001840

1.

(a) $E[\hat{\mu}] = E[\frac{1}{n}\sum_{i=1}^{n} x_i] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$

(b)

$$\begin{aligned}\sigma^2 &= Var(X) = E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - \mu^2 \\ &= E[X^2] - (E[X])^2\end{aligned} \tag{1.1}$$

$$Var(\sum_{i=1}^{n} x_i) = \sum_{i=1}^{n} Var(x_i) = \sum_{i=1}^{n} \sigma^2 = n\sigma^2 \tag{1.2}$$

$$E[\sum_{i=1}^{n} x_i] = n\mu \tag{1.3}$$

Using Eqs.(1.1), (1.2), and (1.3), the variance of the sample mean $\hat{\mu}$ can be computed as

$$\begin{aligned}Var(\hat{\mu}) &= E[(\hat{\mu}-\mu)^2] = E[\hat{\mu}^2] - \mu^2 = E[(\frac{1}{n}\sum_{i=1}^{n} x_i)^2] - \frac{1}{n^2}E[\sum_{i=1}^{n} x_i]^2 \\ &= \frac{1}{n^2}(E[(\sum_{i=1}^{n} x_i)^2] - E[\sum_{i=1}^{n} x_i]^2) = \frac{1}{n^2}Var(\sum_{i=1}^{n} x_i) = \frac{1}{n^2}n\sigma^2 \\ &= \frac{\sigma^2}{n}\end{aligned}$$

When we choose the sample, we can select the sample with larger size which will decrease the variance of $\hat{\mu}$. By this way, we can get more reliable estimates of μ.

(c)

Suppose X= [1, 2, 3, 4, 5]

$\hat{\mu}$=3

$$\hat{\sigma}_n^2 = \frac{1}{5}[(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2] = 2$$

$$\hat{\sigma}_{n-1}^2 = \frac{n}{n-1}\hat{\sigma}_n^2 = \frac{5}{4} * 2 = 2.5$$

In MATLAB,

```
>> X=[1,2,3,4,5];
>> var(X)
ans =
     2.5000
```

So, I find that $\hat{\sigma}^2_{n-1}$ is returned by the variance function in MATLAB.

2.

(a) Sample median is robust against outliers.

(b)

In the definition of χ^2 statistic, $\chi^2 = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$.

Since n_{ij} is the observed counts and e_{ij} is the expected counts of (x_1, x_2), the outliers in x_1 and x_2 will not affect the overall counts distribution of (x_1, x_2).

So, the outliers cannot affect the value of χ^2, which means the result of χ^2 test will not be affected.

Hence, χ^2 statistic is robust to outliers in x_1 and x_2.

3.

(a)

```
>> X=[69,74,68,70,72,67,66,70,76,68,72,79,74,67,66,71,74,75,75,76];
>> mean(X)
ans =
    71.4500
```

So the mean of X is 71.4500

```
>> median(X)
ans =
    71.5000
```

So the median of X is 71.5000

```
>> mode(X)
ans =
    74
```

So the mode of X is 74

(b)

```
>> Y=[153,175,155,135,172,150,115,137,200,130,140,265,185,112,140,150,165,185,210,220];
>> var(Y,1)
ans =
    1.3692e+03
```

The sample variance $\sigma^2_n = \mathrm{var}(Y, 1) = 1369.2$

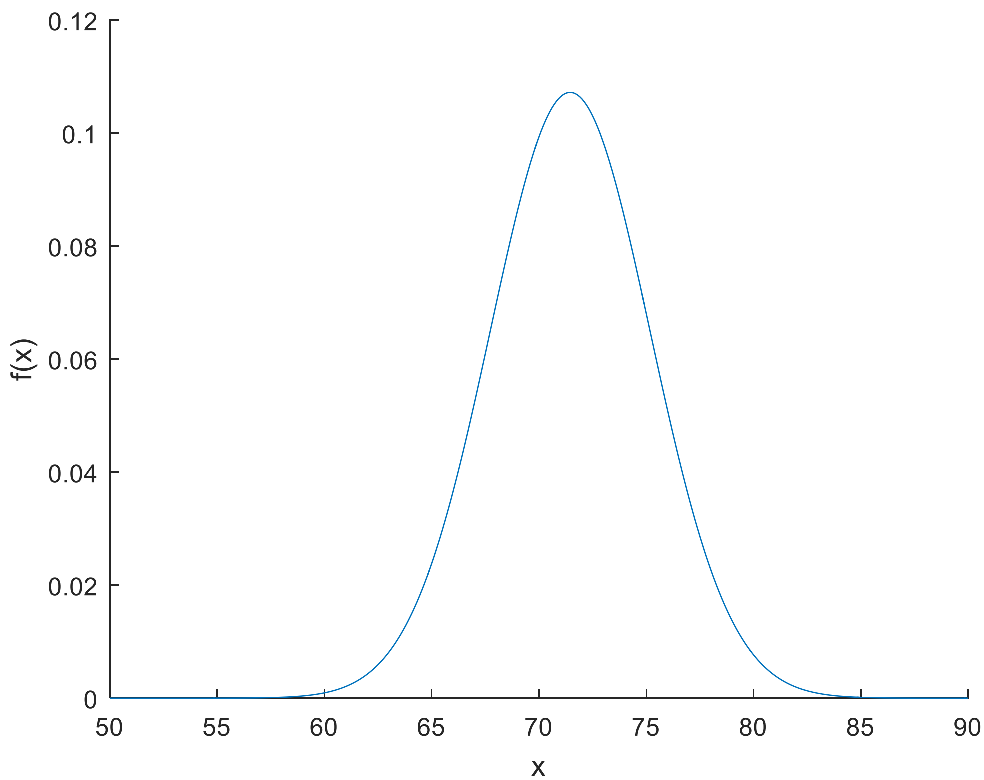
(c)

The sample mean of X: μ=71.45

The sample variance of X: $\sigma^2 = \mathrm{var}(X, 1) = 13.8475$

In MATLAB, plot the probability density function of the normal distribution parameterized by the sample mean and sample variance of X.

```
>> x=50:0.1:90;
>> y=normpdf(x,71.45,3.7212);
>> hold on;
>> xlabel('x');
>> ylabel('f(x)');
>> plot(x,y);
```



(d)

```
>> p=normcdf(80,71.45,3.7212)
p =
       0.9892
```

So, $P(X > 80) = 1 - P(X \leq 80) = 1 - 0.9892 = 0.0108$ in the data.

(e)

```
>> mean(X)
ans =
```

71.4500

```
>> mean(Y)
ans =
   164.7000
```

So, the two dimensional mean:

$$\hat{\mu} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_i \\ \frac{1}{n}\sum_{i=1}^{n} y_i \end{pmatrix} = \begin{pmatrix} \frac{1}{20}\sum_{i=1}^{20} x_i \\ \frac{1}{20}\sum_{i=1}^{20} y_i \end{pmatrix} = \begin{pmatrix} mean(X) \\ mean(Y) \end{pmatrix} = \begin{pmatrix} 71.4500 \\ 164.7000 \end{pmatrix}$$

```
>> cov(X,Y,1)
ans =
   1.0e+03 *
     0.0138    0.1224
     0.1224    1.3692
```

So, the sample covariance matrix $\hat{\Sigma} = \begin{pmatrix} 13.8 & 122.4 \\ 122.4 & 1369.2 \end{pmatrix}$

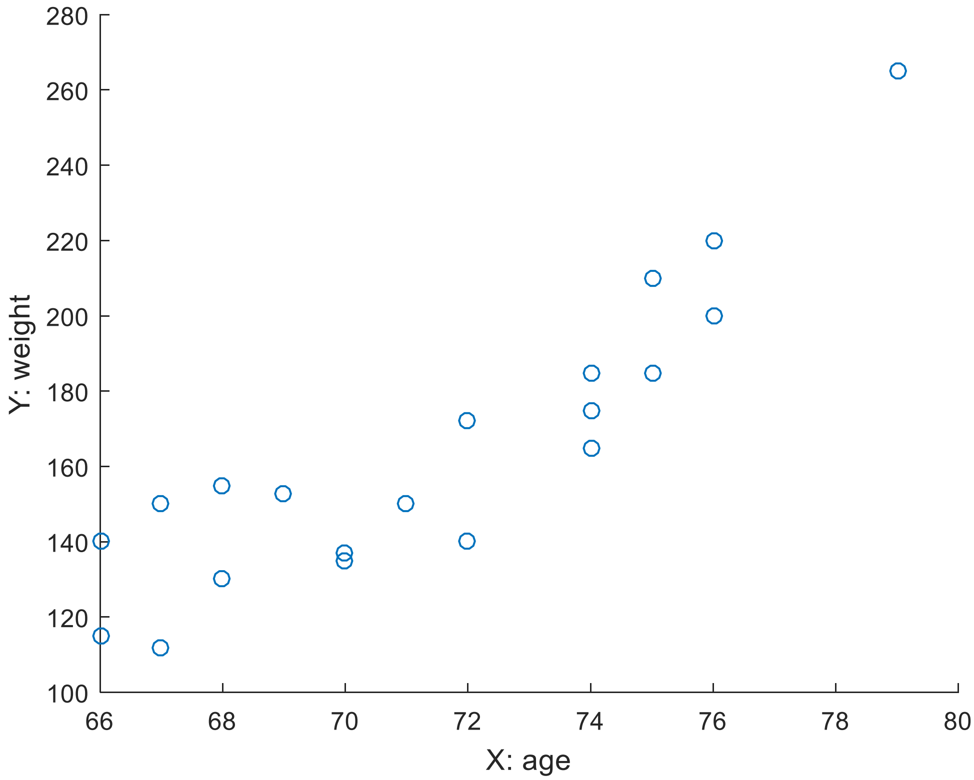
(f)

```
>> corrcoef(X,Y)
ans =
     1.0000    0.8892
     0.8892    1.0000
```

So, the correlation between age and weight is 0.8892.

(g)

```
>> scatter(X,Y)
>> hold on;
>> xlabel('X: age');
>> ylabel('Y: weight');
```



4.

(a)

```
>> D=[9,22;0,2;8,19;10,18;1,2]
D =
     9    22
     0     2
     8    19
    10    18
     1     2
>> mean(D)
ans =
    5.6000    12.6000
```

So, the sample mean $\hat{\mu} = \begin{pmatrix} 5.6 \\ 12.6 \end{pmatrix}$

```
>> cov(D,1)
ans =
   17.8400    35.8400
   35.8400    76.6400
```

So, the sample covariance matrix $\hat{\Sigma} = \begin{pmatrix} 17.84 & 35.84 \\ 35.84 & 76.64 \end{pmatrix}$

(b)

```
>> eig(cov(D,1))
ans =
    0.8841
   93.5959
```

So, the eigenvalues of $\hat{\Sigma}$ are 0.8841 and 93.5959

(c) 1

(d)

Step 1: Center the data

```
>> CenteredData=D-repmat(mean(D),5,1)
CenteredData =
    3.4000    9.4000
   -5.6000  -10.6000
    2.4000    6.4000
    4.4000    5.4000
   -4.6000  -10.6000
```

Step 2: Compute covariance matrix

```
>> cov(CenteredData,1)
ans =
   17.8400   35.8400
   35.8400   76.6400
```

Step 3: Compute eigenvectors and eigenvalues

```
>> [V,D]=eig(cov(CenteredData,1))
V =
   -0.9039    0.4277
    0.4277    0.9039
D =
    0.8841         0
         0   93.5959
```

Since eigenvalue 93.5959 > 0.8841, the corresponding eigenvector $u = \begin{pmatrix} 0.4277 \\ 0.9039 \end{pmatrix}$ is the first principal component of D.

(e)

```
>> U=[0.4277 0.9039];
>> X=[9 0 8 10 1;22 2 19 18 2];
>> U*X
ans =
   23.7351    1.8078   20.5957   20.5472    2.2355
```

So, the coordinate of each data point projected on the first principal component is:

23.7351 1.8078 20.5957 20.5472 2.2355 respectively.

(f)

$$
\begin{aligned}
MSE(u) &= \frac{1}{n}\sum_{i=1}^{n}\|\varepsilon_i\|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\|x_i' - x_i\|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i' - x_i)^T(x_i' - x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i'^T - x_i^T)(x_i' - x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i'^T x_i' - x_i'^T x_i - x_i^T x_i' + \|x_i\|^2)
\end{aligned}
$$

Because $\varepsilon_i = x_i' - x_i$ is orthogonal to the approximation x_i',

which means $(x_i' - x_i)^T x_i' = 0$.

So, $x_i'^T x_i' = x_i^T x_i'$.

Hence, $MSE(u) = \frac{1}{n}\sum_{i=1}^{n}(\|x_i\|^2 - x_i'^T x_i)$

Noting that $x_i' = (u^T x_i)u$, we have

$$
\begin{aligned}
MSE(u) &= \frac{1}{n}\sum_{i=1}^{n}(\|x_i\|^2 - ((u^T x_i)u)^T x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\|x_i\|^2 - u^T(x_i^T u)x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\|x_i\|^2 - (u^T x_i)(x_i^T u)) \\
&= \frac{1}{n}\sum_{i=1}^{n}\|x_i\|^2 - \frac{1}{n}\sum_{i=1}^{n}u^T(x_i x_i^T)u \\
&= \frac{1}{n}\sum_{i=1}^{n}\|x_i\|^2 - u^T(\frac{1}{n}\sum_{i=1}^{n}x_i x_i^T)u \\
&= \frac{1}{n}\sum_{i=1}^{n}\|x_i\|^2 - u^T\Sigma u
\end{aligned}
$$

5.

Since the correlation between X and Y is zero, that is $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = 0$.

Hence, $\sigma_{xy} = 0$.

Since
$$\begin{aligned}\sigma_{xy} &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[XY - \mu_Y X - \mu_X Y + \mu_X \mu_Y] \\ &= E[XY] - \mu_Y E[X] - \mu_X E[Y] + \mu_X \mu_Y \\ &= E[XY] - \mu_X \mu_Y\end{aligned}$$

Hence, $E[XY] - \mu_X \mu_Y = 0$

That is $\frac{a + c}{5} - \frac{2}{5} * \frac{2a + b + 2c}{5} = 0$.

Hence, $5(a + c) - 2(2a + b + 2c) = 0$.

So, the relation between a, b and c is $a + c = 2b$.