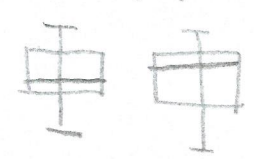


1.

1) 0,  $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X, Y)$  에서  
 $Var(X+Y) = Var(X) + Var(Y)$  가 주어진 조건이므로  $2Cov(X, Y) = 0$  입니다.  
 $Cov(X, Y) = 0$  이 되므로 이는  $E(XY) - E(X)E(Y) = 0$  과 같습니다.  
 $\therefore E(XY) = E(X)E(Y)$  가 됩니다.

2) 0, boxplot에서 왜도는 중앙값을 나타내는 박스를 통해 알 수 있습니다.  
boxplot을 통해 정확한 왜도를 구하는 불가능 하지만  이 두  
boxplot에서 좌측 boxplot의 중앙값이 제 1사분위 값에 가깝기 때문에 좌측으로  
왜도가 쏠려 있다는 것을 알 수 있습니다. 마찬가지로 우측의 그래프는 중앙값과 제 3  
사분위 수 사이에 그 구간에 데이터가 집중되어 있음을 알 수 있습니다.  
Stem-and-leaf display에서는 모든 데이터의 값을 알 수 있기 때문에 왜도의 값을  
직접 구할 수 있습니다. 그리고 잎의 갯수를 통해 데이터가 얼마나 비대칭인지 왜도의  
대략적인 값을 파악할 수 있습니다.

3)  $X_i$  iid인 random variables  $X_i$  가  $n$ 번 성공할때 까지  $Bern(p)$ 의 시행횟수이므로  
이는 음이항 분포(Negative Binomial Distribution)입니다. iid 이므로  $E(X_i) = E(X_1 + X_2 + \dots + X_n)$   
 $= E(X_1) + E(X_2) + \dots + E(X_n)$  입니다.  $E(X_i) = \frac{1}{p}$  입니다. (기하 분포)  
그러므로  $E(X_1) + E(X_2) + \dots + E(X_n) = n \cdot \frac{1}{p} = \frac{n}{p}$  입니다.

4) 0, EDA는 샘플의 크기가 커야 합니다. 샘플의 양이 매우 적은 경우 통계적으로  
왜도가 커 유익성이 없으며 샘플의 크기가 충분히 커야 다양성과 모집단에 대한 오차가 적은  
추정이 가능하기 때문입니다. (t-statistic 이용)

5) 0, Histogram과 Stem-and-leaf display 모두 정량적, Numeric 데이터의  
분포를 나타낸다는 비슷한 특징을 가지고 있습니다.

2. resistant statistic은 outlier에 대해 robust 한 통계량을 의미합니다.  
 중앙값은 resistant statistic으로 예시 자료에서 중앙값은 오름차순으로 나열한 뒤 5번째  
 6번째 자료의 평균인 13.1이 중앙값이 됩니다. 100이라는 새로운 자료가 추가되어도  
 중앙값은 13.4로 resistant를 가집니다. IQR도 resistant statistic입니다.  
 예시 자료에서 중앙값은 13.1이고 Q1 값은 제 1사분위 수로 10.4, Q3은 제 3사분위 수로  
 15.6이므로  $IQR = Q_3 - Q_1 = 15.6 - 10.4 = 5.2$ 로 마찬가지로 이상치가 추가되어도 큰  
 영향을 받지 않습니다. 최빈값은 예시 자료에서 15.6으로 이상치 100이 추가되어도 동일  
 합니다.

### 3. 1) five-number summaries.

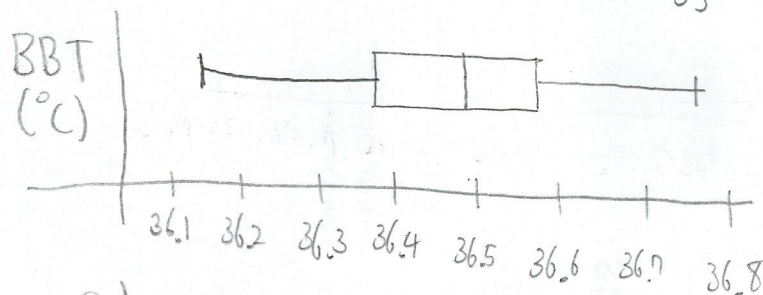
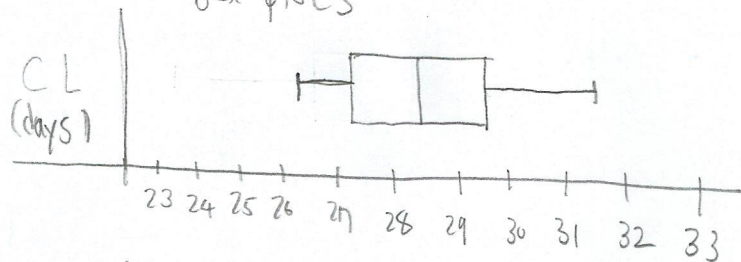
# 20 CL

M	10.5	28.4
F	5.5	27.2 29.65
I	1	26.3 31.8

# 20 BBT

M	10.5	36.49
F	5.5	36.38 36.58
I	1	36.13 36.77

box plots



### OUTLIER CUTOFFS

$$CL : F_U = 29.65 \quad F_L = 27.2$$

$$d_F = 29.65 - 27.2 = 2.45$$

$$\text{outlier cutoffs} = (27.2 - \frac{3}{2} \times 2.45, 29.65 + \frac{3}{2} \times 2.45) = (23.525, 33.32)$$

$$BBT : F_U = 36.58 \quad F_L = 36.38$$

$$d_F = 36.58 - 36.38 = 0.2$$

$$\text{outlier cutoffs} = (36.38 - \frac{3}{2} \times 0.2, 36.58 + \frac{3}{2} \times 0.2) = (36.08, 36.88)$$

⇒ 두 boxplot의 scale의 차이가 커서  
 따로 각각의 boxplot을 그렸습니다.

2) CL이 42.9 days BBT가 36.44°C 인 새로운 데이터가 추가된다면

먼저 CL의 boxplot은 이상치가 생깁니다. 42.9 days 는 이상치로 'o' marker로 그래프의  
 왼쪽에 표시 됩니다. 중앙값은 28.4, 제 1사분위 수는 27.2, 최솟값은 26.3으로 변하지 않으니  
 제 3사분위 수는 29.65로 증가하고 최댓값도 42.9가 됩니다.

BBT의 36.44°C는 원래 boxplot에서 확인 하면 Q1과 median 사이의 값이므로 새로운 box  
 plot은 Q1이 감소하고 중앙값도 감소합니다. Q3과 최댓값은 변하지 않습니다. 자료의 갯수  
 증가가 되어 Q1과 Q3은 계산할 때 중앙값이 제외되기 때문입니다.



4. Depth (n=20) (unit = 0.1 kg)

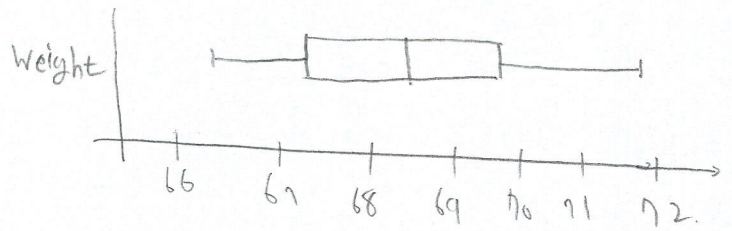
1)	1	66*	3
	5	66.	6 8 9 9
		67*	
	8	67.	5 6 6
(3)		68*	0 4 4
	9	68.	5 8 8
	6	69*	4
	5	69.	9
	4	70*	0 3
		70.	
	2	71*	2
	1	71.	8

2) weight

$$F_U = 69.65 \quad F_L = 67.2$$

$$d_F = 2.45 \quad \text{outlier cutoffs} = (63.525, 73.325)$$

$$\frac{3}{2}d_F = 3.675$$



Height

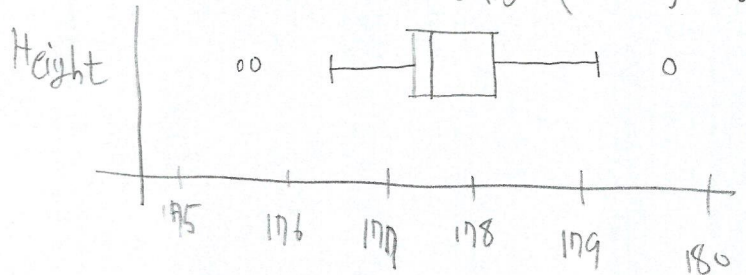
$$F_U = 178.07$$

$$F_L = 177.28$$

$$d_F = 0.79$$

$$\frac{3}{2}d_F = 1.185$$

$$\text{outlier cutoffs} = (176.095, 179.255)$$



3) weight outlier: X

height outlier: 175.52, 179.53

4) 2)번 에서 같이 그려줍니다.

5. 1) China. boxplot 의 box의 길이 ( $Q_3 - Q_1$ )가 가장 길고 outlier도 존재하여 이를 다른 나라들과 비교하였을 때 China가 가장 높은 인구 분포형을 가진 나라입니다.

2) Sweden. 1)번과 반대로 box의 길이가 가장 짧고 outlier도 존재하지 않는 Sweden이 가장 고른 인구 분포를 나타냅니다.

3) China. upper outlier cutoff 는  $Q_3 + 1.5 IQR$  입니다. 제 3 사분위수는 boxplot에서 보면 China의 제 3 사분위수가 더 큼니다.  $Q_3 - Q_1 = (\text{box의 길이})$  이므로 China의 IQR 값이 더 큼니다. 그러므로  $Q_3 + 1.5 IQR$ 은 China가 더 큰 값을 가집니다.

4) Yes. 1) China 의 2번으로 인구가 많은 도시는 China 도시중 outlier인 Shanghai 다음으로 인구가 많은 도시입니다. 즉 outlier를 제외한 데이터 중 최댓값입니다. X축의 눈금을 보면 약 42 (100,000)에 위치합니다. 2) Sweden의 가장 인구가 많은 도시는 outlier가 없기 때문에 boxplot에서 가장 우측의 눈금인 약 19 (100,000)를 나타냅니다.  $42 > 19$  이기 때문에 답은 Yes 입니다.