B5

Jeremiah Theisen

2024-08-29

```
ncbirths = read.csv("https://github.com/TienChih/tbil-stats/raw/main/data/ncbirths.csv")
names(ncbirths)
```

```
##  [1] "fage"           "mage"           "mature"         "weeks"
##  [5] "premie"         "visits"         "marital"        "gained"
##  [9] "weight"         "lowbirthweight" "gender"         "habit"
## [13] "whitemom"
```

1.5.1

```
L1=c(13, 13, 12, 10, 10, 14, 12, 11, 10, 14, 15, 10, 11, 13, 14, 14, 14, 15, 14, 11)
L2=c(19, 12, 13, 9, 15, 5, 7, 12, 9, 14, 8, 20, 19, 15, 13, 8, 14, 14, 7, 17)
mean(L1)
```

```
## [1] 12.5
```

a.

```
mean(L1)
```

```
## [1] 12.5
```

```
median(L1)
```

```
## [1] 13
```

```
mean(L2)
```

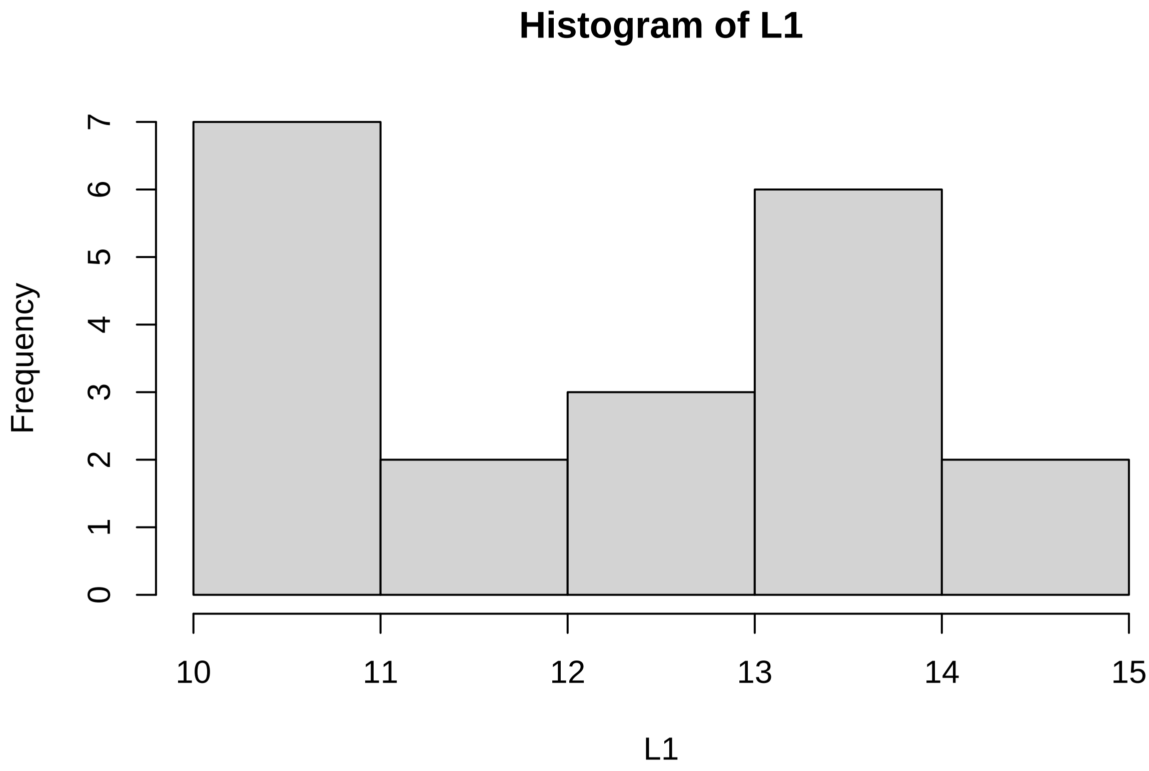
```
## [1] 12.5
```

```
median(L2)
```

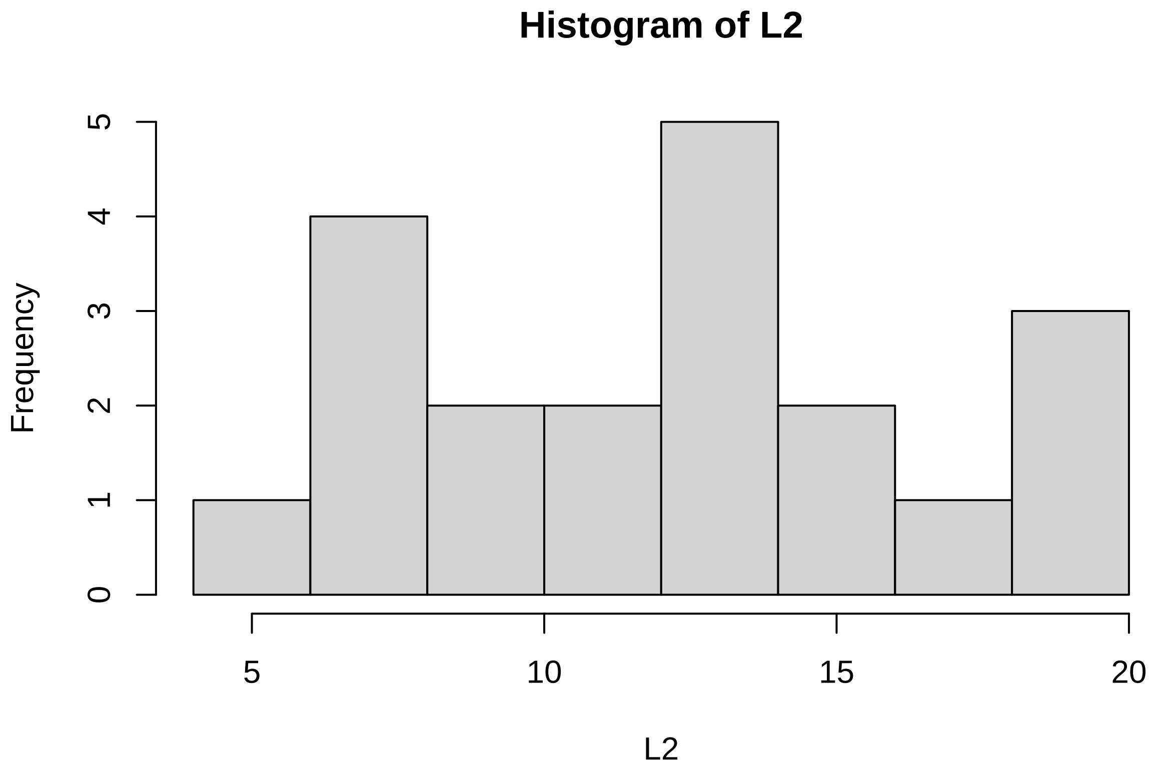
```
## [1] 13
```

b.

```r
hist(L1)
```



```r
hist(L2)
```



c.

1.5.2

a. Maximum = 9, Minimum = 2, Range = 7

b. Max = 500, Min = 0, Range = 500

c. The range here applies only to the data with the most extremely high or low value, every other value could be ignored.

1.5.3

a.

```
S=c(1,2,3,4,5)

mean(S)
```

```
## [1] 3
```

b. c.

```
S1=c((1-3),(2-3),(3-3),(4-3),(5-3))
print(S1)
```

```
## [1] -2 -1  0  1  2
```

```
S2=c((1-3)+(2-3)+(3-3)+(4-3)+(5-3))
print(S2)
```

```
## [1] 0
```

d. The result of the math is 0, but we know that the spread is not 0.

e.

```
S3=c(((1-3)^2)+((2-3)^2)+((3-3)^2)+((4-3)^2)+((5-3)^2))
print(S3)
```

```
## [1] 10
```

This fix is supposed to make sure that the spread is not 0, as any number squared is positive.

f. It is more accurate than the one in c, but still inaccurate, as 10 is too large

g.

```
T=c(1,1,2,2,3,3,4,4,5,5)
```

T should be as spread out as S, as it is also in the range 1-5, just with two of each number instead of one.

h.

```
mean(T)
```

```
## [1] 3
```

```
T3=c(((1-3)^2)+((1-3)^2)+((2-3)^2)+((2-3)^2)+((3-3)^2)+((3-3)^2)+((4-3)^2)+((4-3)^2)+((5-3)^2)+((5-3)^2
print(T3)
```

```
## [1] 20
```

The value is double because the total value of T is S*2.

i.

j

```
T4 = (T3/10)
S4 = (S3/5)

print(T4)
```

```
## [1] 2
```

```
print(S4)
```

```
## [1] 2
```

The values are the same, which is correct because every number in both datasets is within the range of 3-2 to 3+2.

1.5.5

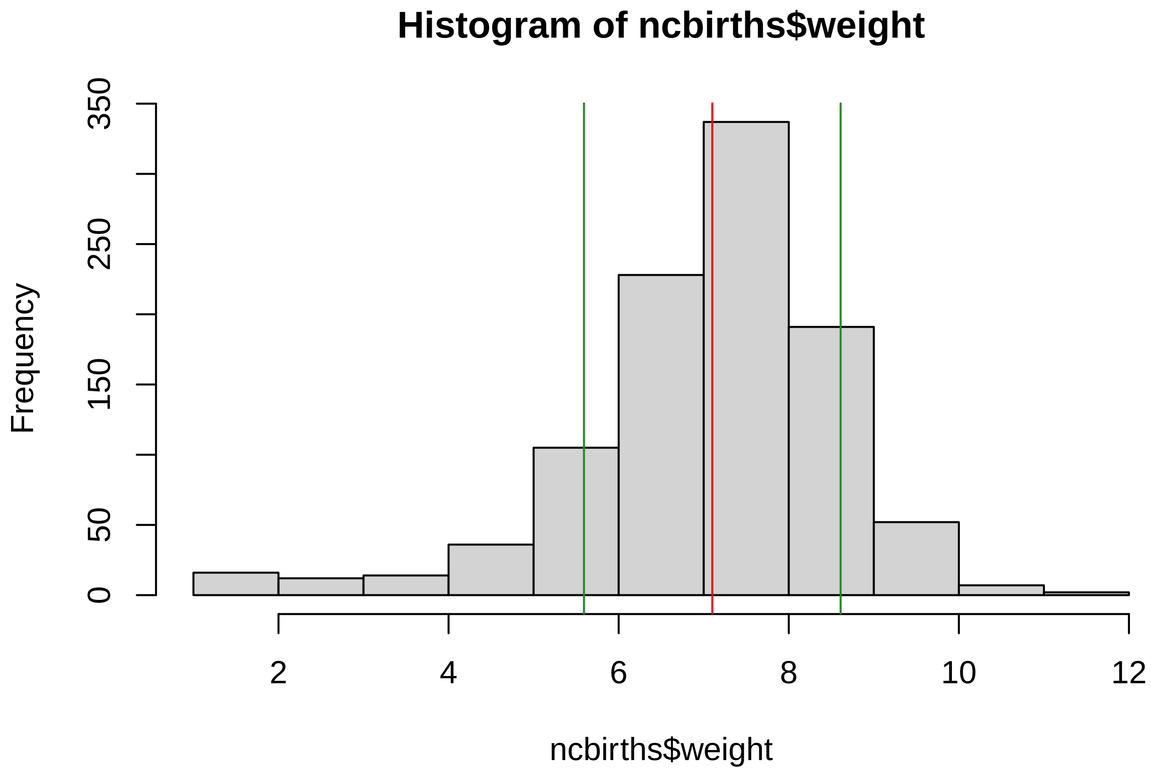
a.

```
sd(ncbirths$weight)
```

```
## [1] 1.50886
```

b.

```
m=mean(ncbirths$weight)
s=sd(ncbirths$weight)
hist(ncbirths$weight)
abline(v = m, col = "red")
abline(v = m-s, col = "forestgreen")
abline(v = m+s, col = "forestgreen")
```



1.5.6

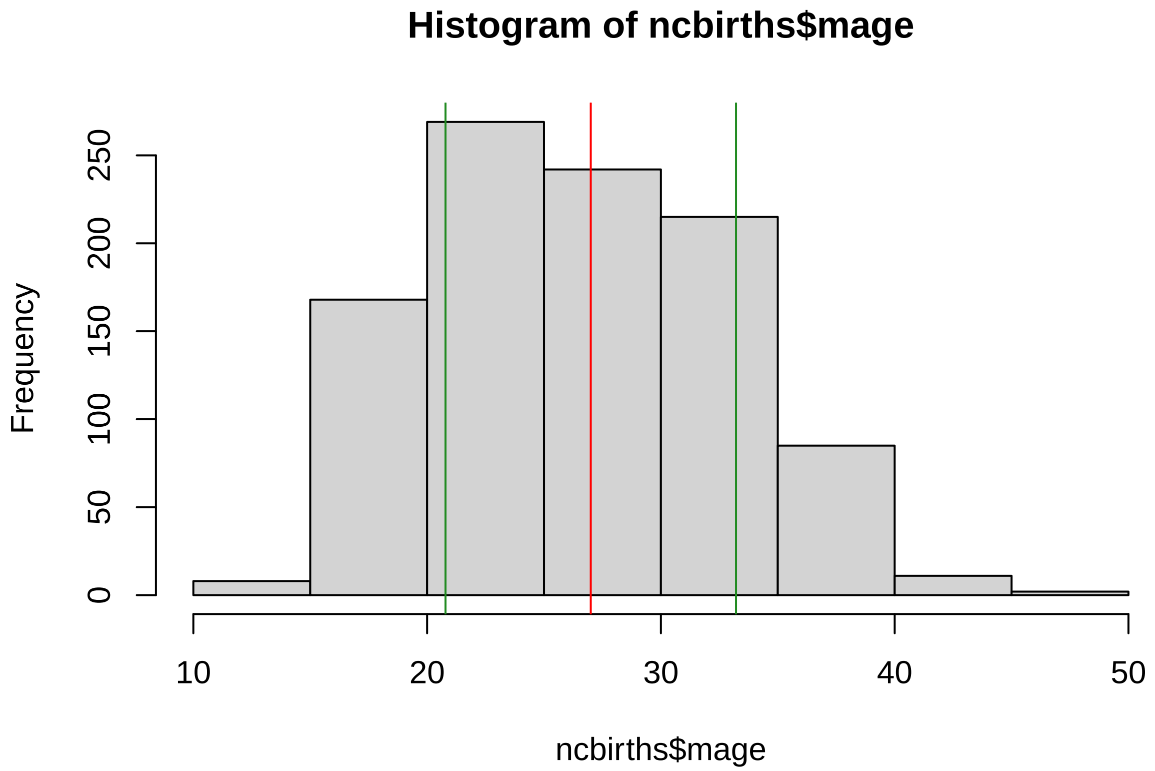
a.
b.

```
sd(ncbirths$mage)
```

```
## [1] 6.213583
```

c.

```
m=mean(ncbirths$mage)
s=sd(ncbirths$mage)
hist(ncbirths$mage)
abline(v = m, col = "red")
abline(v = m-s, col = "forestgreen")
abline(v = m+s, col = "forestgreen")
```



1.5.7

a.

```
Q = c(4, 6, 9, 12, 13, 16, 19, 23, 25, 26, 31, 32, 33, 37, 40, 42, 43, 47, 48, 49)
Q2 = median(Q)
print(Q2)
```

```
## [1] 28.5
```

b. c. d.

```
h1 = c(4, 6, 9, 12, 13, 16, 19, 23, 25, 26)
h2 = c(31, 32, 33, 37, 40, 42, 43, 47, 48, 49)
Q1 = median(h1)
Q3 = median(h2)
print(Q1)
```

```
## [1] 14.5
```

```
print(Q2)
```

```
## [1] 28.5
```

```
print(Q3)
```

```
## [1] 41
```

e. One quarter of the data is between each value

f. 26.5

1.5.8

```
a1 = c(0, 1, 2, 2, 2, 2, 3, 5, 5, 5, 5, 8, 8, 8, 9, 11, 12, 12, 14, 15, 19, 41, 45, 52, 62, 66, 85, 96)
```

a.

```
h1 = c(0, 1, 2, 2, 2, 2, 3, 5, 5, 5, 5, 8, 8, 8)
h2 = c(9, 11, 12, 12, 14, 15, 19, 41, 45, 52, 62, 66, 85, 96)
Q1 = median(h1)
Q2 = median(a1)
Q3 = median(h2)
min = 0
max = 96
IQR = Q3 - Q1

print(Q1)
```

```
## [1] 4
```

```
print(Q2)
```

```
## [1] 8.5
```

```
print(Q3)
```

```
## [1] 30
```

```
print(min)
```

```
## [1] 0
```

```
print(max)
```

```
## [1] 96
```

```
print(IQR)
```

```
## [1] 26
```

b.

```
o1 = Q3 + IQR * 1.5
o2 = Q1 - (IQR * 1.5)
print(o1)
```

```
## [1] 69
```

```
print(o2)
```

```
## [1] -35
```

By that math, we have outliers with 85 and 96

c.

```
a2 = c(0, 1, 2, 2, 2, 2, 3, 5, 5, 5, 5, 8, 8, 8, 9, 11, 12, 12, 14, 15, 19, 41, 45, 52, 62, 66)

a2h1 = c(0, 1, 2, 2, 2, 2, 3, 5, 5, 5, 5, 8, 8)
a2h2 = c( 8, 9, 11, 12, 12, 14, 15, 19, 41, 45, 52, 62, 66)

a2Q1 = median(a2h1)
a2Q2 = median(a2)
a2Q3 = median(a2h2)
min = 0
max = 66

print(a2Q1)
```

```
## [1] 3
```

```
print(a2Q2)
```

```
## [1] 8
```

```
print(a2Q3)
```

```
## [1] 15
```

```
print(min)
```

```
## [1] 0
```

```
print(max)
```

```
## [1] 66
```

d. https://www.desmos.com/calculator/jtxqpxqghh
e. By excluding outliers, Desmos excludes the two numbers which I also excluded, meaning that I was right about them.

1.5.9

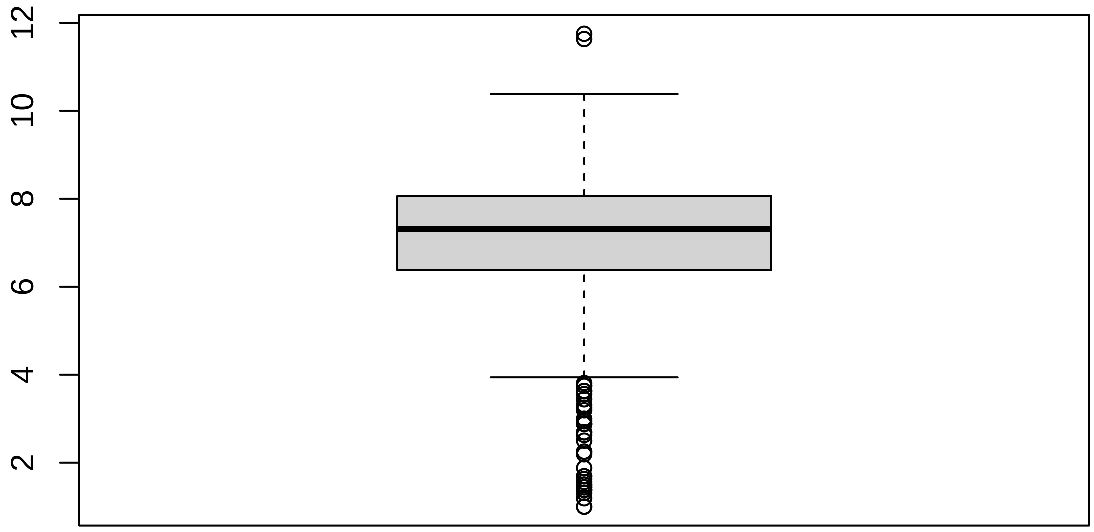
a.

```
summary(ncbirths$weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.000   6.380   7.310   7.101   8.060  11.750
```

b.

```
boxplot(ncbirths$weight)
```



c. Min = 1, Q1 = 6.380, Q2 = 7.310, Q3 = 8.06, max = 11.75

1.5.10

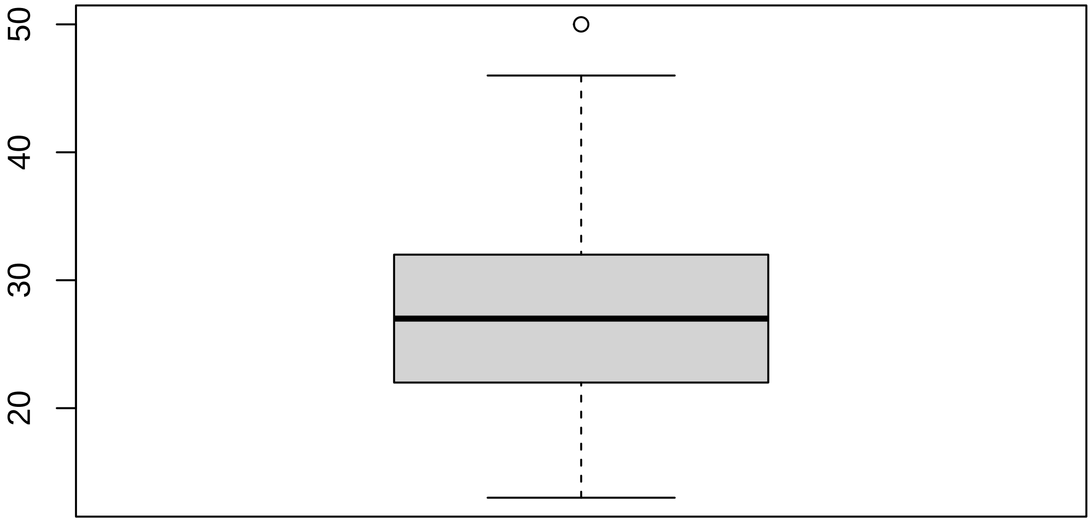
a.
b.

```
summary(ncbirths$mage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      13      22      27      27      32      50
```

c.

```
boxplot(ncbirths$mage)
```

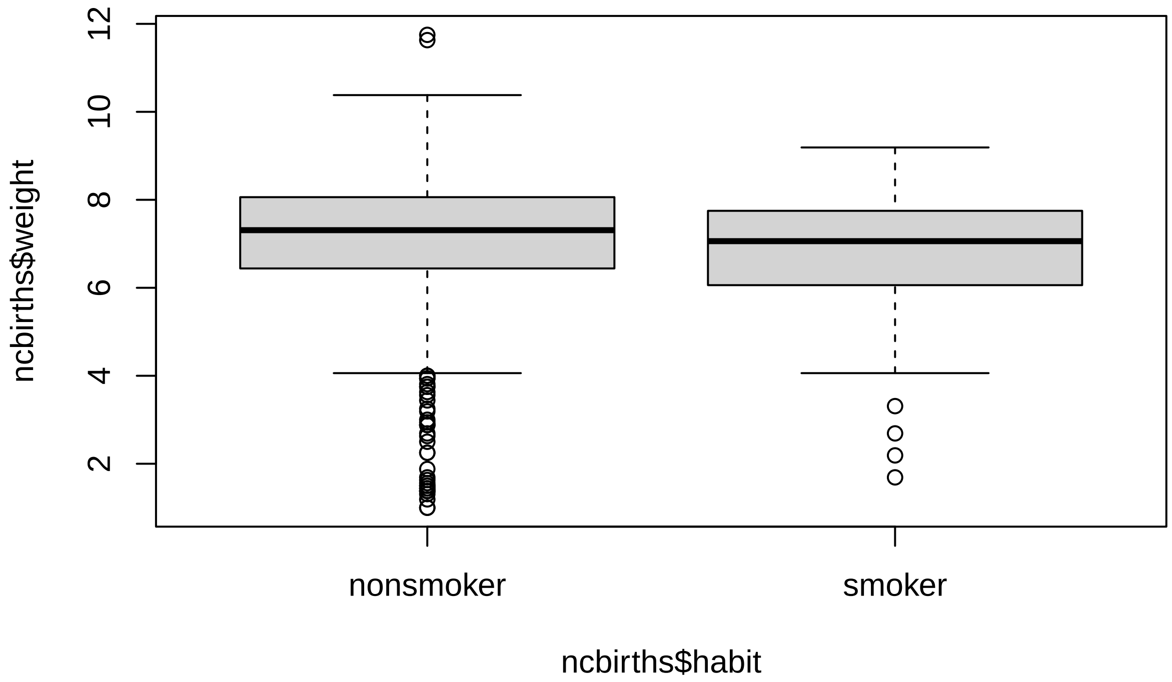


d. Min. 1st Qu. Median Mean 3rd Qu. Max. 13 22 27 27 32 50

1.5.11

a.

```
boxplot(ncbirths$weight~ncbirths$habit)
```



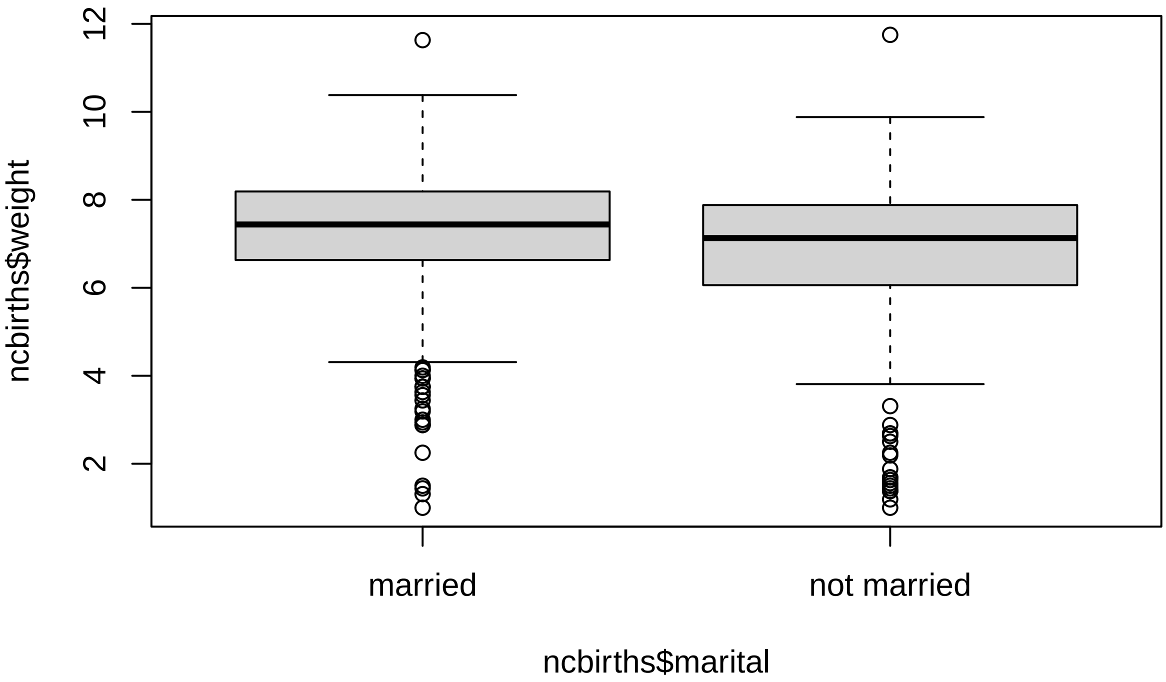
b. Babies of smokers tend to have less birth weight, but the difference is not statistically significant.

1.5.12

a.

b

```
boxplot(ncbirths$weight~ncbirths$marital)
```



c. Births from married women tend to have more weight, but the difference is not statistically significant

1.5.13

a. https://www.desmos.com/calculator/skpxiira1a

b. https://www.desmos.com/calculator/b8x9lfupzt

c. The mean changed a lot, the median remained the same. Both standard deviations also grew. The min remained the same, the max drastically changed (obviously). Q1 and Q3 changed by 0.5 each.