B4

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"
```

1.4.1

a. This is close to the average because of the small data set, but would not work on a larger data set, especially with since there are smaller intervals, such as $3,500 and $3,600.

b. This best represents the data here, because of the large exception of $15,000.

c. I think that this is the best method of calculating averages, but is not accurate here because it is skewed by one large value, and is higher than all but one value in the data set.

1.4.2

a. $4 + 8 + 9 + 11 = 32$, $32 / 4 = 8$

b. Yes. As an example, the mean of 1, 1, 1, 1, 1, 1, 100 is 15.14, which is not in the data set

1.4.3

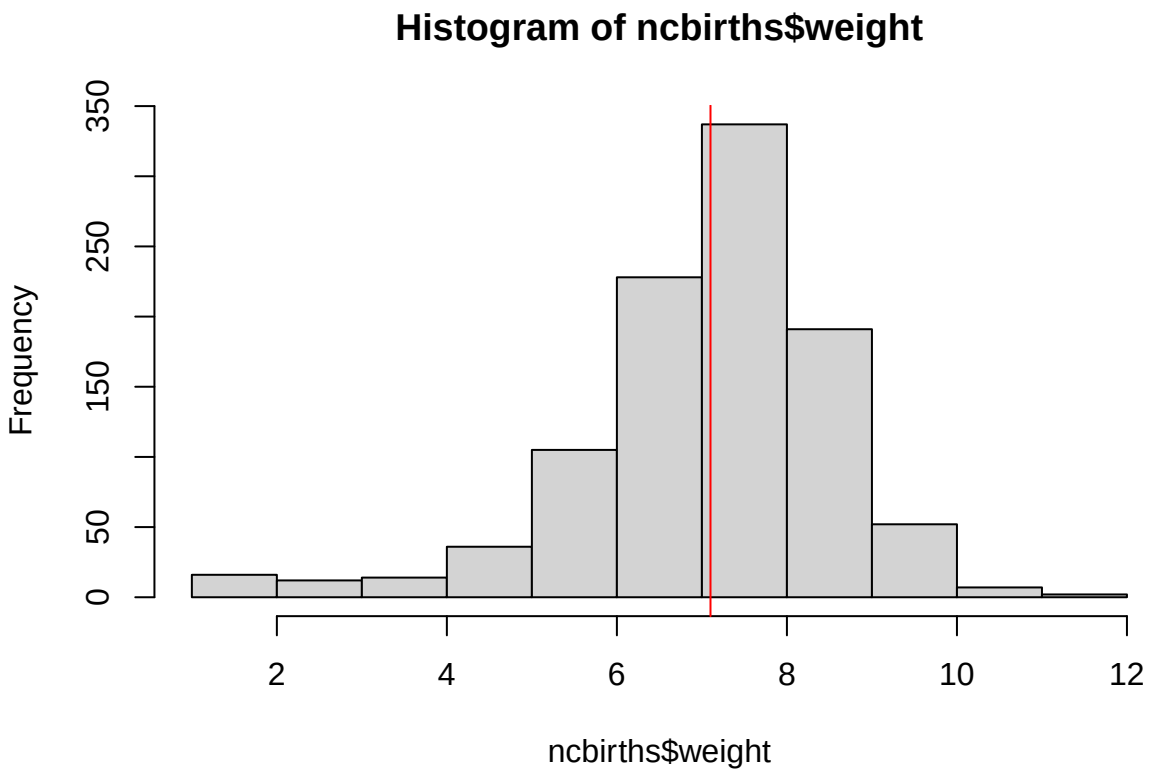
a.

```
mean(ncbirths$weight)
```

```
## [1] 7.101
```

b.

```
m=mean(ncbirths$weight)
hist(ncbirths$weight)
abline(v = m, col = "red")
```



1.4.4

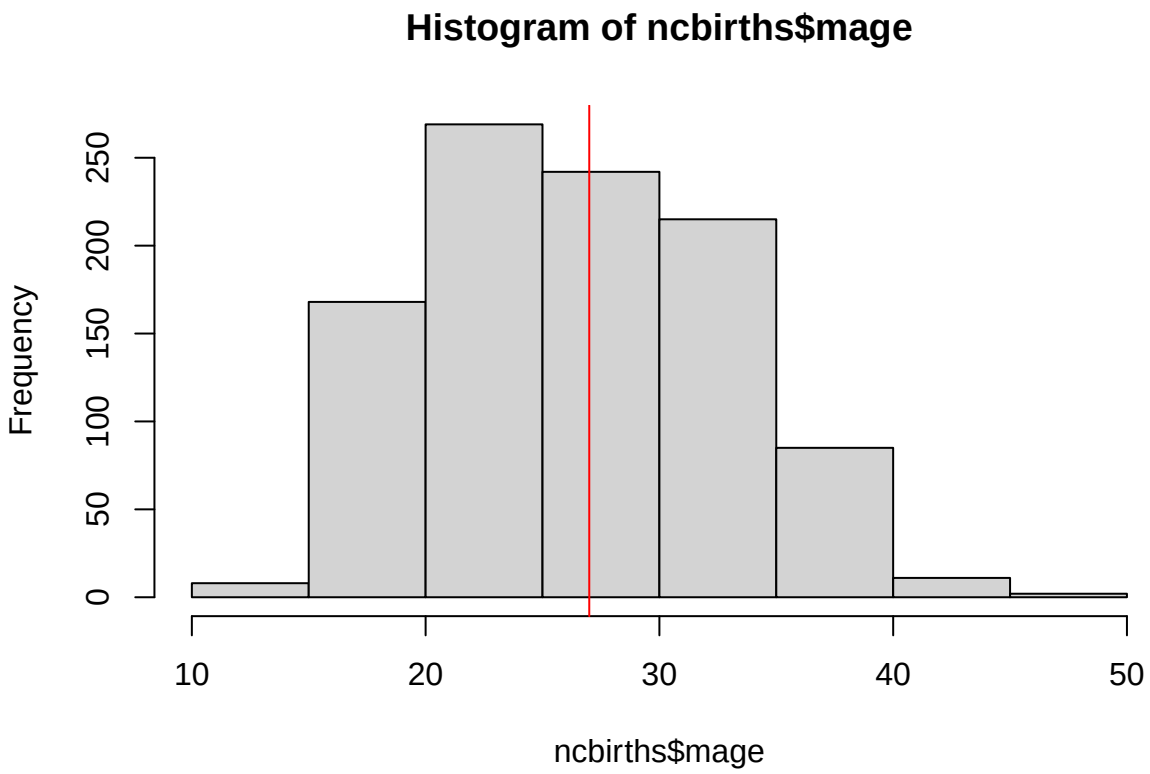
a.
b.

```
mean(ncbirths$mage)
```

```
## [1] 27
```

c.

```
m=mean(ncbirths$mage)
hist(ncbirths$mage)
abline(v = m, col = "red")
```



1.4.5

a. 0, 2, 3, 9, 9, 11, 12 median is 9

b. 1, 2, 5, 6, 7, 8

1.4.6

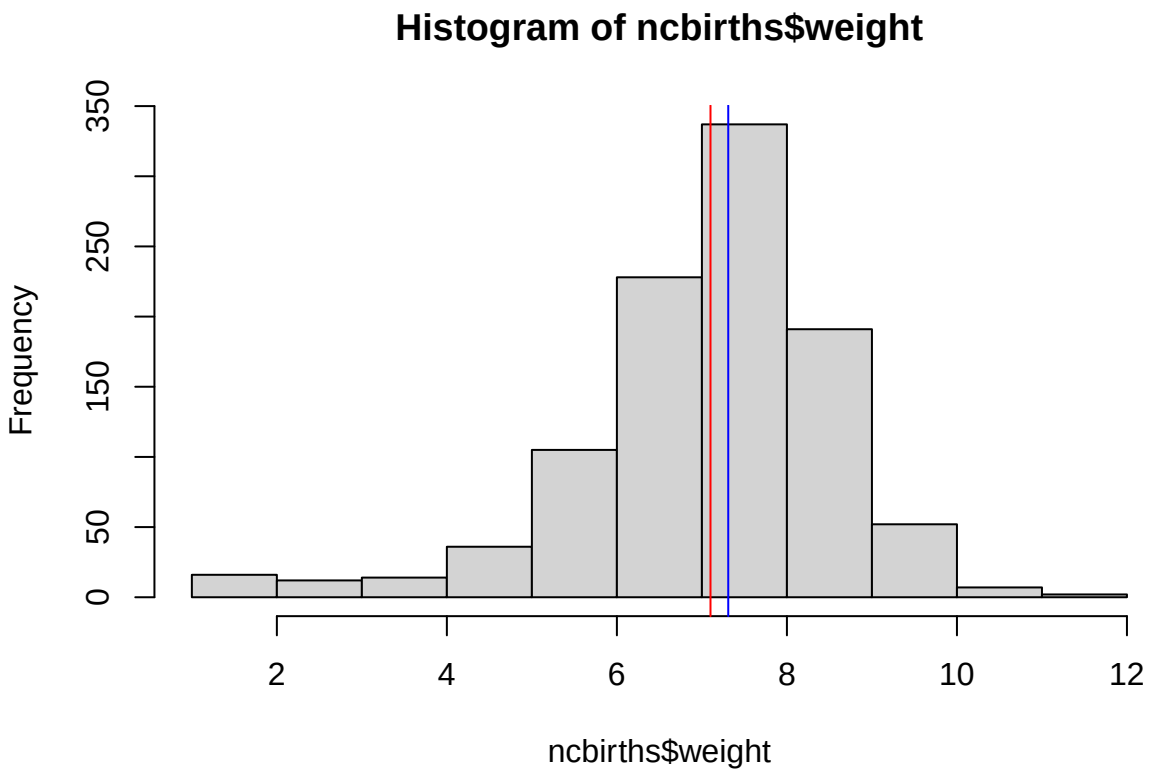
a.

```
median(ncbirths$weight)
```

```
## [1] 7.31
```

b.

```
md=median(ncbirths$weight)
m=mean(ncbirths$weight)
hist(ncbirths$weight)
abline(v = md, col = "blue")
abline(v = m, col ='red')
```




```
print(m)
```

```
## [1] 7.101
```

```
print(md)
```

```
## [1] 7.31
```

1.4.7

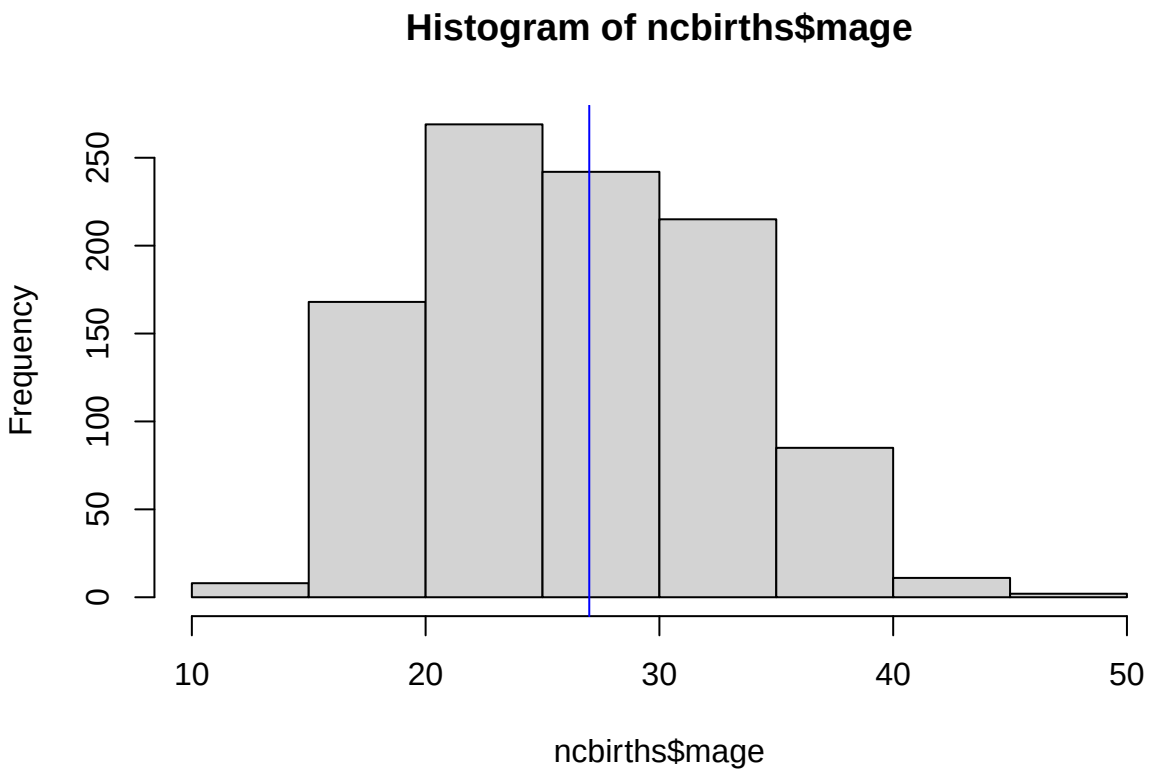
a.
b.

```
median(ncbirths$mage)
```

```
## [1] 27
```

c.

```
md=median(ncbirths$mage)
hist(ncbirths$mage)
abline(v = md, col = "blue")
```



1.4.8

a. The difference is 0.444, which has a moderate impact, as the range is 1.5-7.5

b. The difference is 1.3, which is very significant, since most of the data is in the range of 0-3.

c. https://www.desmos.com/calculator/ifd1eescig

d. https://www.desmos.com/calculator/nwojupanzh

e. The median is less than the mean when there is a large number as an outlier. The median is larger than the mean when there is a small or negative number as an outlier.

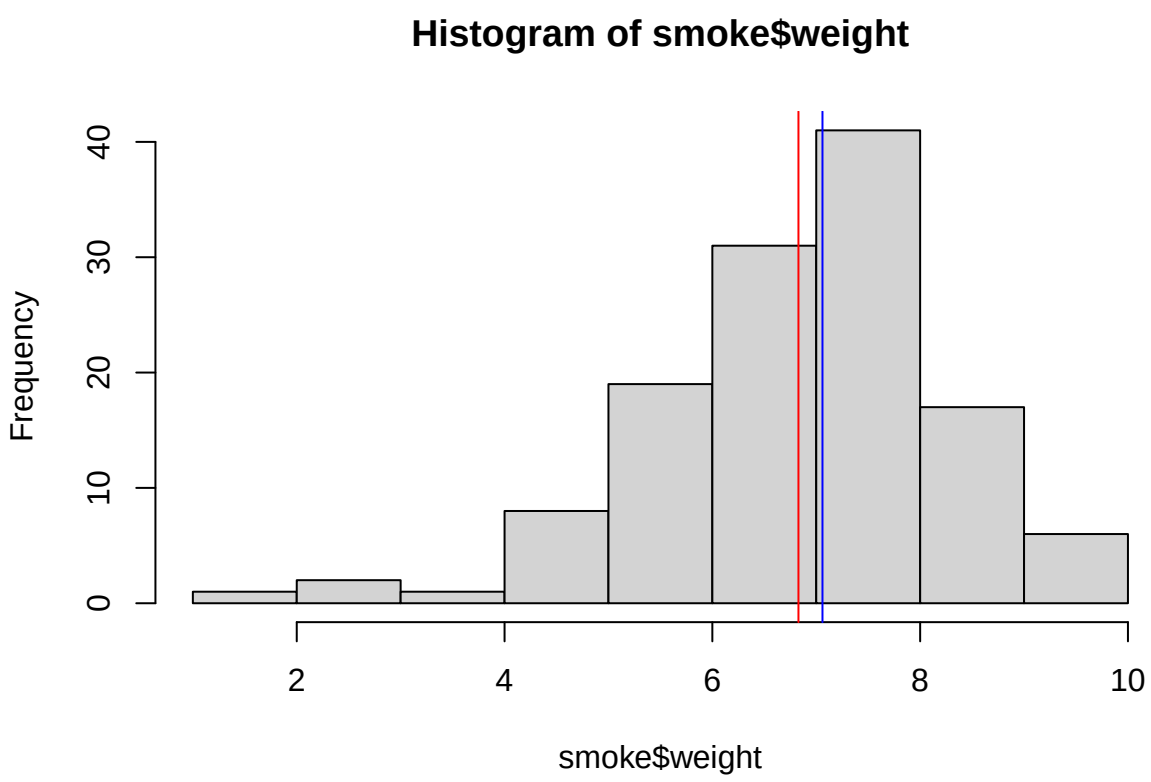
1.4.9

a.

```
smoke=subset(ncbirths, habit=="smoker")
```

b.

```
m=mean(smoke$weight)
md=median(smoke$weight)
hist(smoke$weight)
abline(v = m, col = "red")
abline(v = md, col = "blue")
```




```
print(m)
```

```
## [1] 6.82873
```

```
print(md)
```

```
## [1] 7.06
```

c. The mean is skewed slightly left

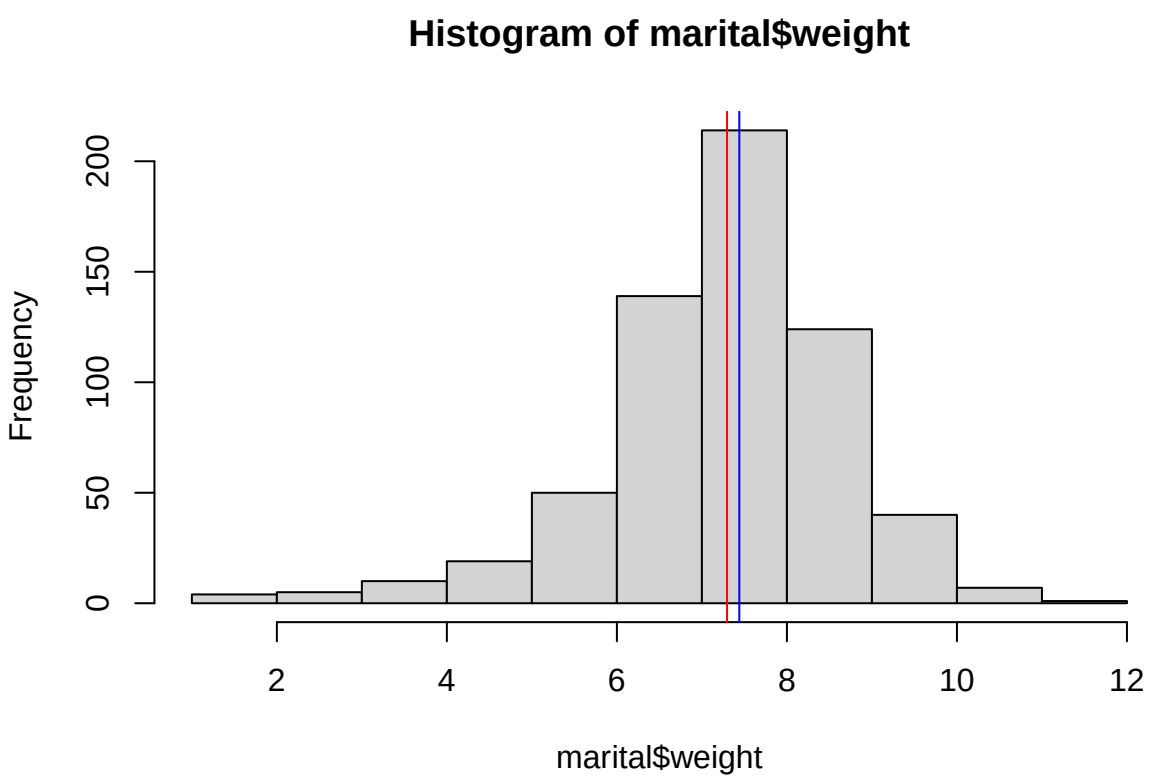
1.4.10

a.

```
marital=subset(ncbirths, marital=="married")
```

b.

```
m=mean(marital$weight)
md=median(marital$weight)
hist(marital$weight)
abline(v = m, col = "red")
abline(v = md, col = "blue")
```




```
print(m)
```

```
## [1] 7.295759
```

```
print(md)
```

```
## [1] 7.44
```

c. The mean is skewed slightly left

1.4.11

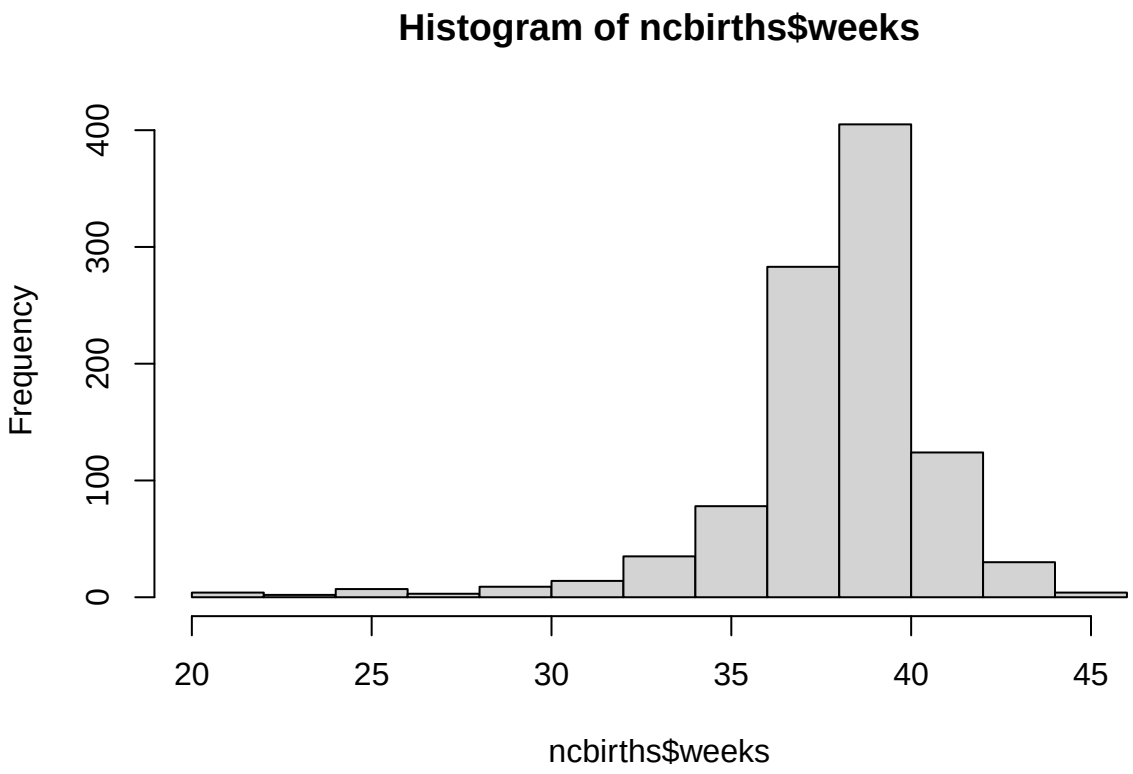
a. The mode is 1

b. 4, 4, 4, 7, 7, 7, 2, 3, 8, 9, 0, 0, 1

1.4.12

a.

```
hist(ncbirths$weeks)
```



b. The mode is the value which has the highest bar, because it has the highest frequency, making it the most common.

1.4.13

a.For categorical variables, mean makes no sense, as dividing and adding categories is not meaningful. Median makes sense unless there is an even number of variables and division is required between two different categories. Mode makes sense, as it is simply which value is most common.