BExam

Jeremiah Theisen

2024-09-05

1

(a)

```
names(Orange)
```

```
## [1] "Tree"          "age"           "circumference"
```

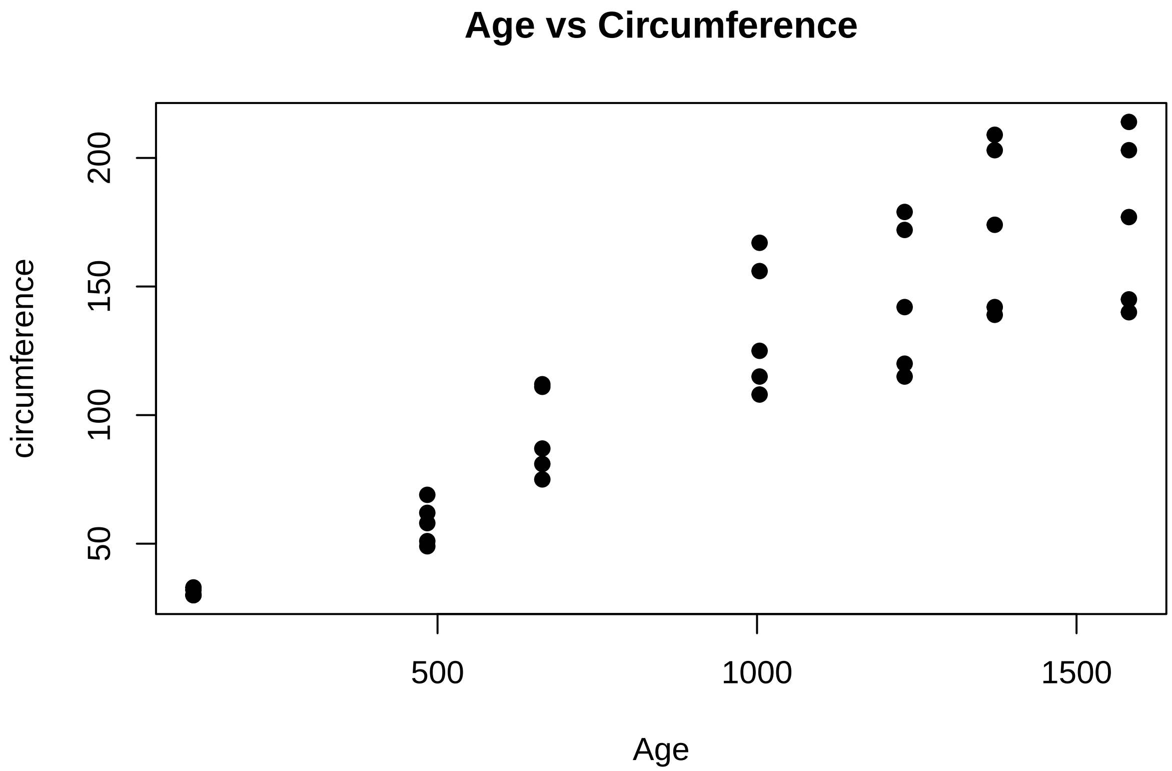
(b) Tree is Nominal, it is just the tree's name, and does not appear to have any natural ranking, even though they are numbers.

Age is Discrete, because although time is a continuous variable, the trees were all measured at certain, discrete ages (as it appears in the plot below).

Circumference is Continuous, because it is a physical measurement, so can always be more accurate, and the variable is not specific about unit.

2

```
plot(Orange$age, Orange$circumference, main=" Age vs Circumference ", xlab=" Age ", ylab=" circumferenc
```



(a) There seems to be a positive association, as the older trees are wider.

(b) This is a fairly strong association, as all of the widest trees are old, but the data set is not large enough to make any conclusive statements.

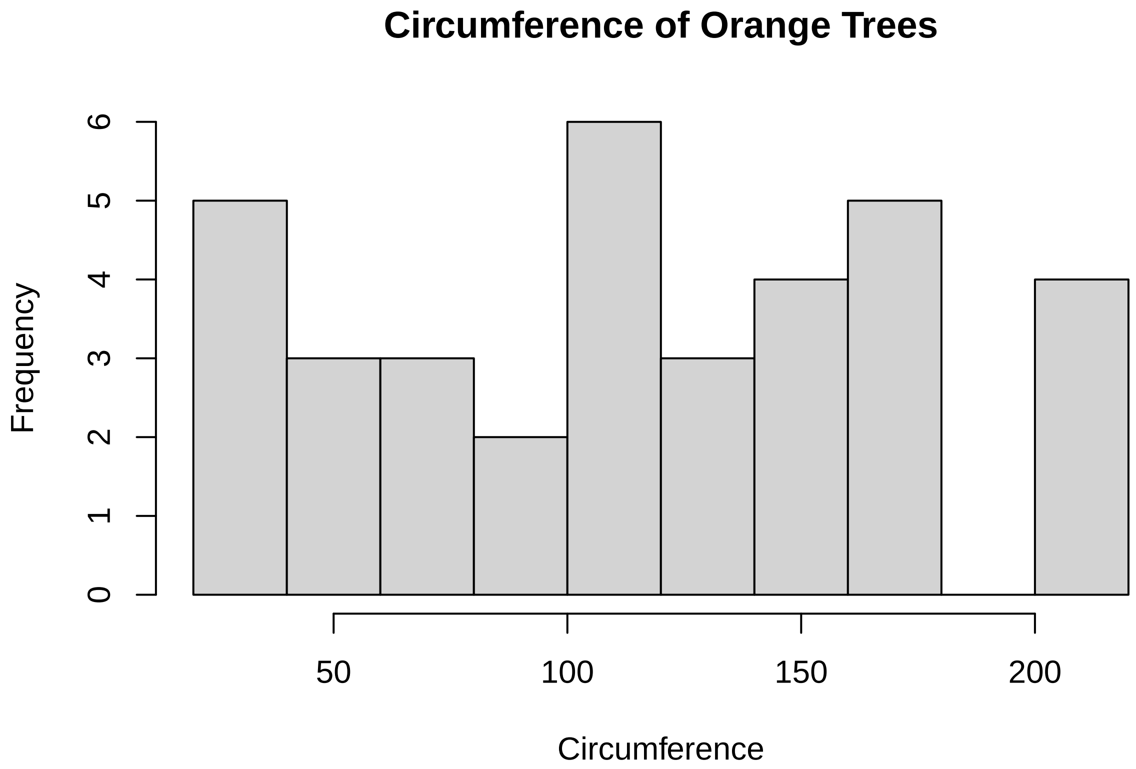
(c) This relationship can be explained by the trees growing and getting larger, as trees tend to do.

3

Stratified random sampling divides a population into strata, and then picks randomly a certain number of samples from each strata. An example would be that if someone wanted to study the average shelf life of all fruits, they would make sure to take samples of all different types of fruit, instead of randomly selecting from a group of all fruit.

4

```
hist(Orange$circumference, xlab="Circumference", main="Circumference of Orange Trees")
```



(a) Each value is a circumference range, and the height of the bars represents how many trees fall into that range.

(b) The data is distributed fairly evenly, the trees are all growing at slightly different rates. The reason would depend on a lot of factors, as the trees seem to be measured at specific times (going by the plot), it would depend on amount of sunlight, soil quality, amount of water, etc. which we can not make a conclusive statement about from only these three variables. Also, given as the data seems to be measuring only five trees as they age, I doubt that a histogram would really provide any valuable information about this data set, as it could be misleading about the total number of trees being measured.

5

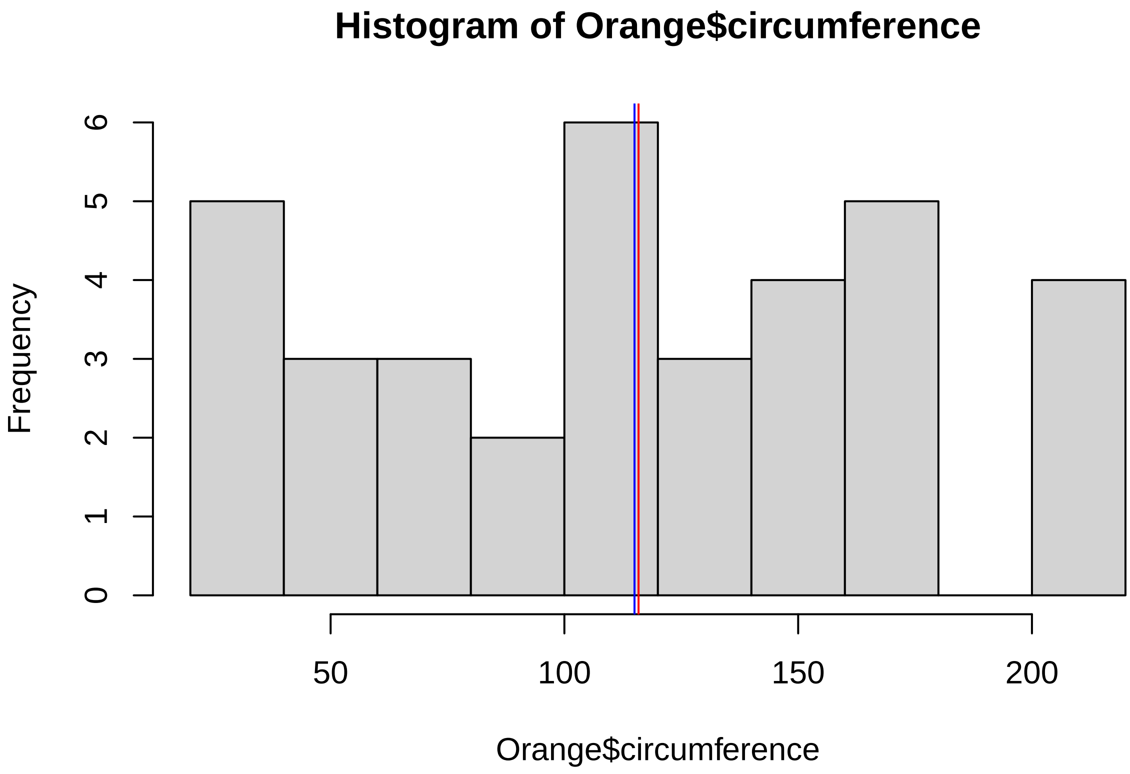
(a) A histogram is best used for numeric variables.
(b) A bar graph is best used for categorical variables.

6

(a)

```
m=mean(Orange$circumference)

md=median(Orange$circumference)
```

```r
hist(Orange$circumference)

abline(v = m, col = "red")

abline(v = md, col = "blue")
```



(b) The data has no statistically significant skew.

7

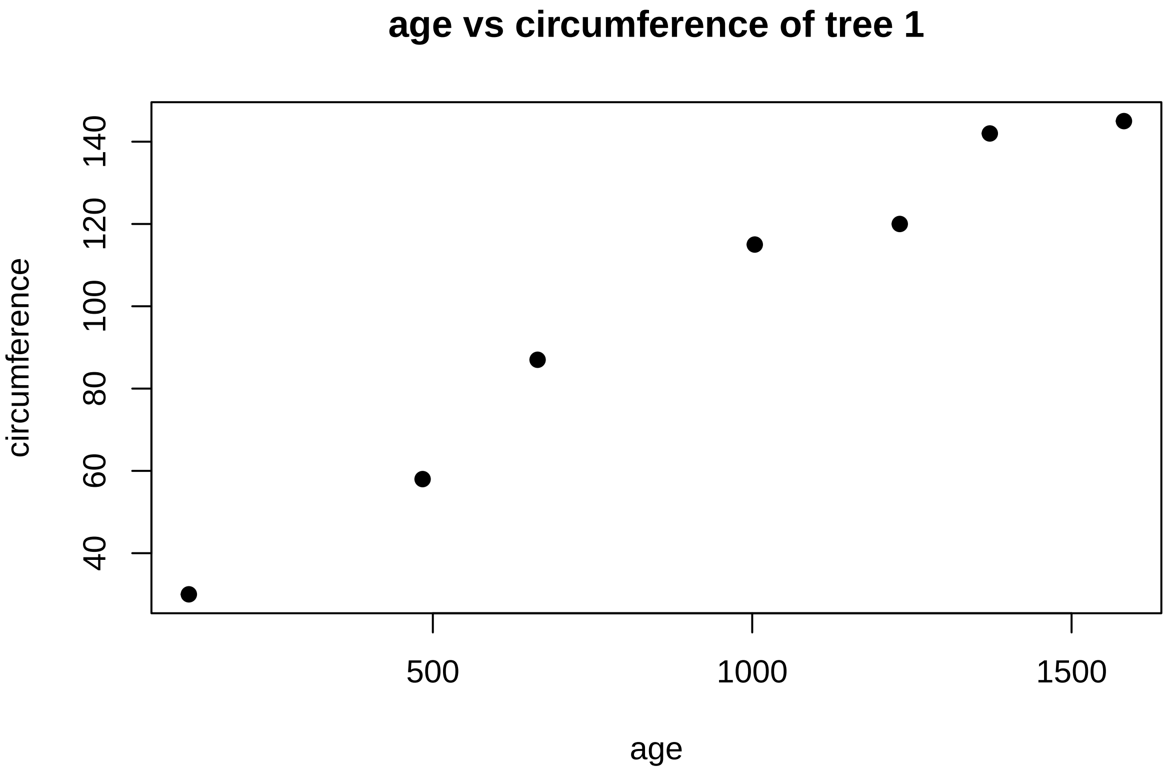
(a)

```r
t1 = subset(Orange, Tree=="1")
t1
```

```
##   Tree  age circumference
## 1    1  118            30
## 2    1  484            58
## 3    1  664            87
## 4    1 1004           115
## 5    1 1231           120
## 6    1 1372           142
## 7    1 1582           145
```

(b)

```
plot(t1$age, t1$circumference, main=" age vs circumference of tree 1 ", xlab=" age ", ylab=" circumferen
```



(c) The data appears to be going linearly up, because as the tree gets older, it is growing, and getting larger, although growth does slow down after 1000, as it reaches the largest size that it can get given its genetics and/or its location.

8

(a)

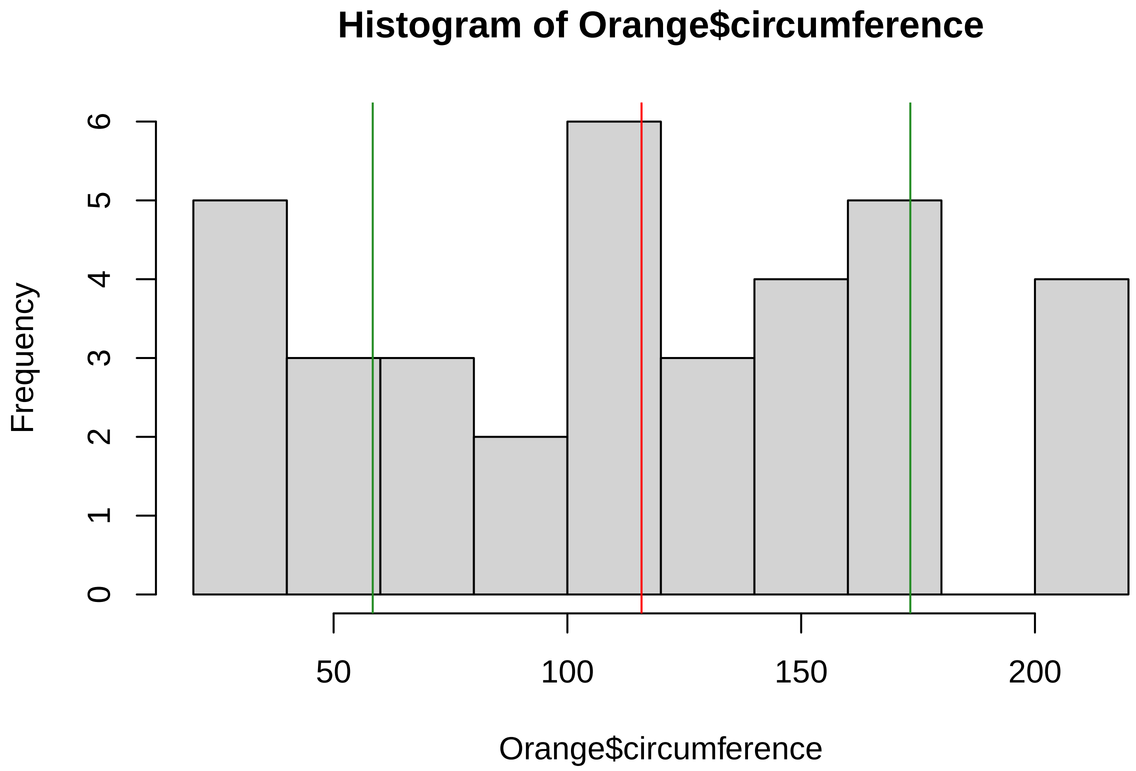
```
sd(Orange$circumference)
```

```
## [1] 57.48818
```

(b)

```
m=mean(Orange$circumference)

s=sd(Orange$circumference)

hist(Orange$circumference)
```

```
abline(v = m, col = "red")

abline(v = m + s, col = "forestgreen")

abline(v = m - s, col = "forestgreen")
```


Histogram of Orange$circumference
Frequency
Orange$circumference