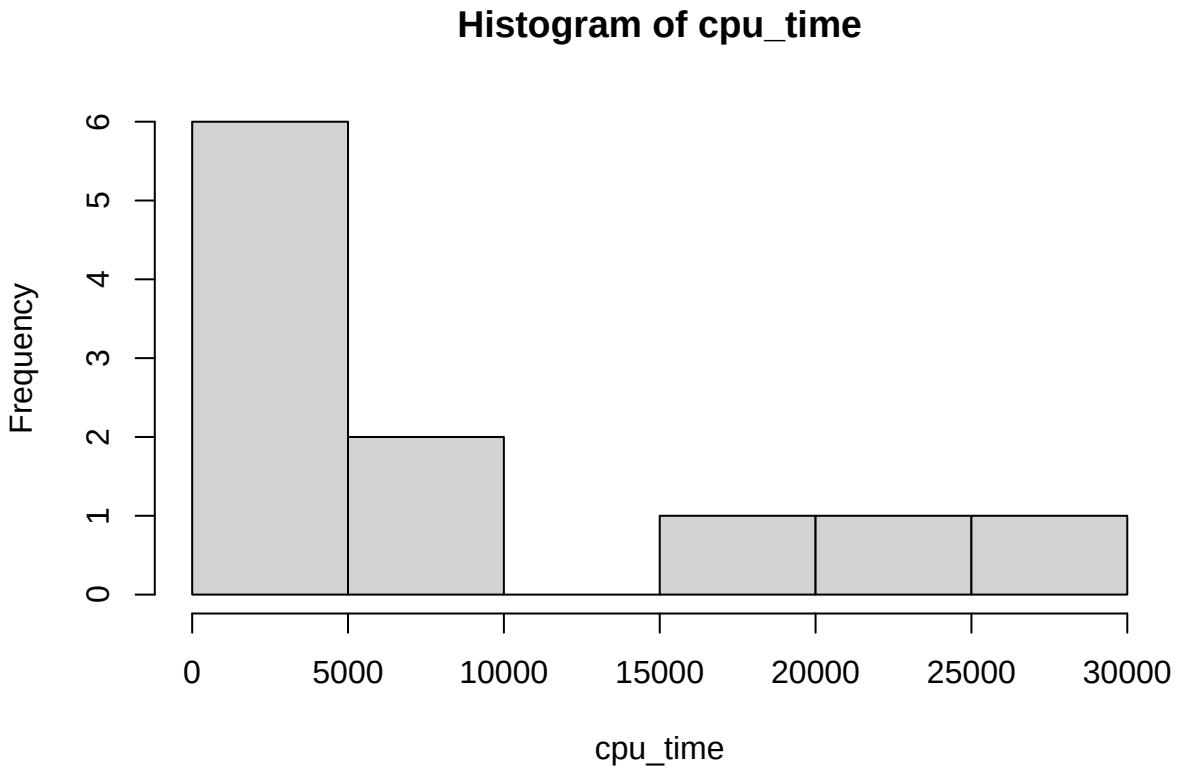
Chapter 12: Summarizing Measured Data

Nicolas Kolling Ribas

Exercises

12.10 The CPU times in milliseconds for 11 workloads on a processor are 0.74, 0.43, 0.24, 2.24, 262.08, 8960, 4720, 19740, 7360, 22,440, and 28,560. Which index of central tendency would you choose and why?

```
cpu_time <- c(0.74, 0.43, 0.24, 2.24, 262.08, 8960, 4720, 19740, 7360, 22440, 28560)
hist(cpu_time)
```



The histogram above shows the distribution of the measured CPU time. We can see a positive skew.

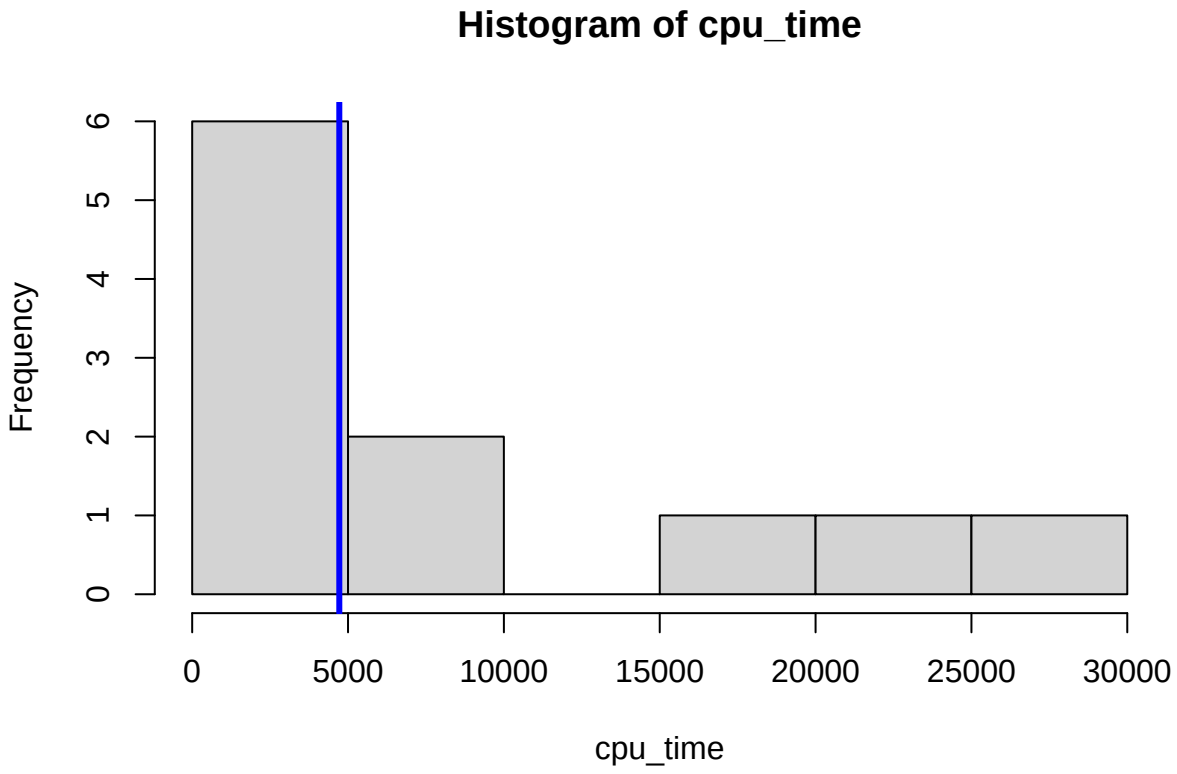
> If the histogram is skewed, the median is more representative of a typical observation than the mean.

So median is chosen.

```
print(paste("Median = ", median(cpu_time)))
```

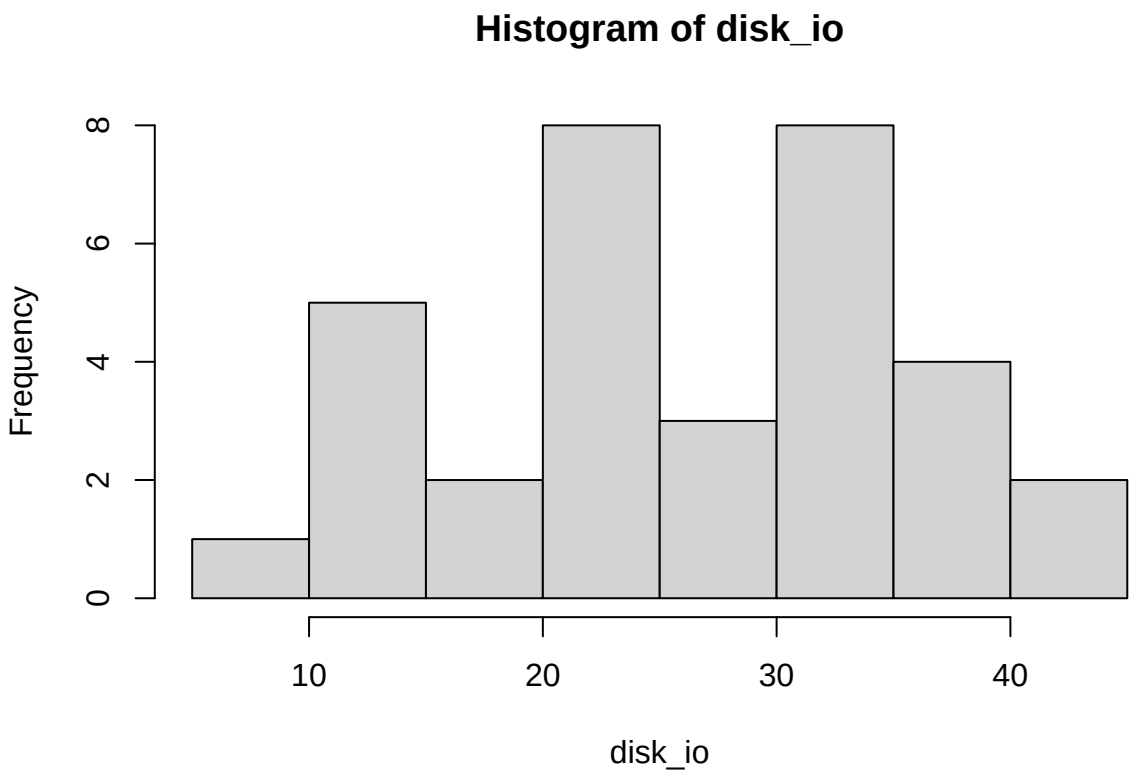
```
## [1] "Median =  4720"
```

```
hist(cpu_time)
abline(v = median(cpu_time), col = "blue", lwd = 3)
```



12.11 The number of disk I/O's performed by a number of programs were measured as follows: {23, 33, 14, 15, 42, 28, 33, 45, 23, 34, 39, 21, 36, 23, 34, 36, 25, 9, 11, 19, 35, 24, 31, 29, 16, 23, 34, 24, 38, 15, 13, 35, 28}. Which index of central tendency would you choose and why?

```
disk_io <- c(23, 33, 14, 15, 42, 28, 33, 45, 23, 34, 39, 21, 36, 23, 34, 36, 25, 9, 11, 19, 35, 24, 31,
hist(disk_io)
```

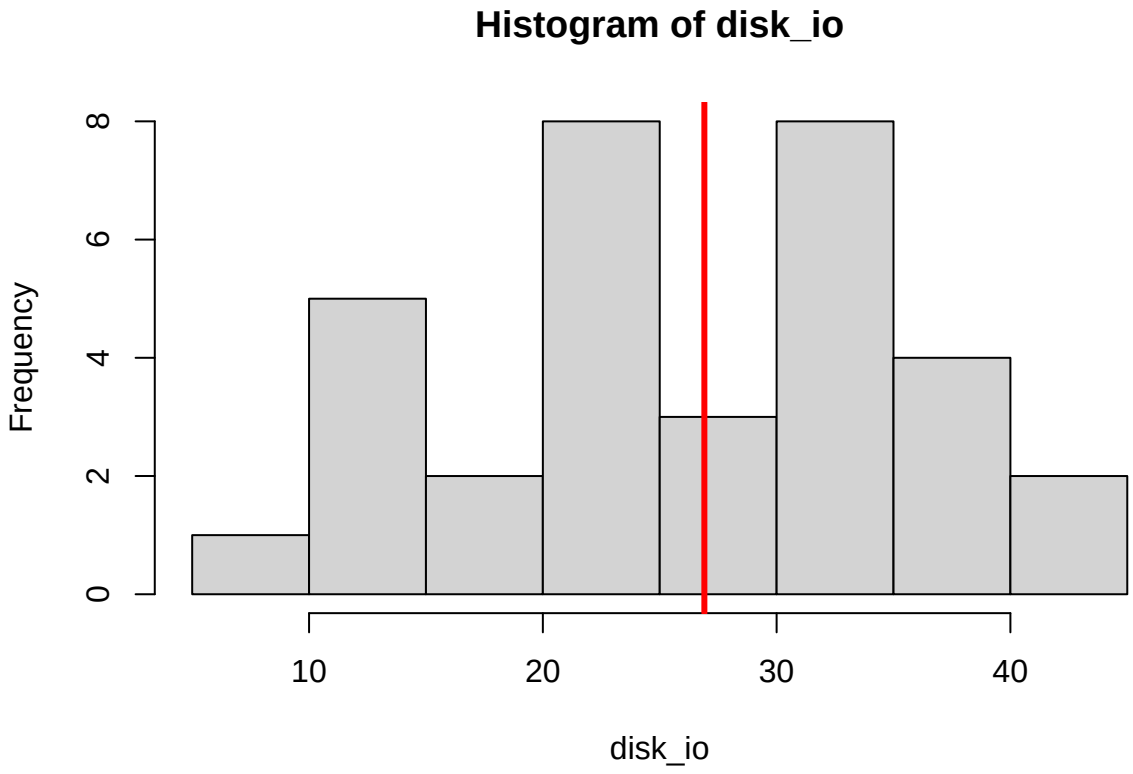


Distribution is not so skewed so mean is chosen as central tendency index

```
print(paste("Mean = ", mean(disk_io)))
```

```
## [1] "Mean =  26.9090909090909"
```

```
hist(disk_io)
abline(v = mean(disk_io), col = "red", lwd = 3)
```



12.13 For the data of Exercise 12.10, which index of dispersion would you choose and why?

Semi-interquantile range. For the same reason the median was chosen as index of central tendency. The distribution of the data has a positive skewed.

```
print(paste("Semi-interquantile range = ", IQR(cpu_time)/2))
```

```
## [1] "Semi-interquantile range =  7174.255"
```

12.14 For the data of Exercise 12.11, compute all possible indices of dispersion. Which index would you choose and why?

```
print(paste("Range = ", range(disk_io)))
```

```
## [1] "Range =  9"  "Range =  45"
```

```
print(paste("Standard deviation = ", sd(disk_io)))
```

```
## [1] "Standard deviation =  9.49461569905424"
```

```
print(paste("C.O.V. = ", sd(disk_io) / mean(disk_io) * 100))
```

```
## [1] "C.O.V. =  35.2840448275664"
```

```
print(paste("Semi-interquantile range = ", IQR(disk_io)/2))
```

```
## [1] "Semi-interquantile range =  6.5"
```

The data is not bounded, so a range doesn't make sense. The distribution is symmetrical, so there is no need for SIQR. Finally, I would choose standard deviation as index of dispersion.