Measures of Spread

John Karuitha

4/6/2021

Introduction

- Measures of spread capture how far apart or close together a set of data points are.
- The most prominent measures of spread we capture in this lesson.
  - The variance.
  - The standard deviation.
  - The mean absolute deviation.
  - The inter-quartile range.

The data

- We use the following two datasets to examine these measures.

```
set.seed(100, sample.kind = "Rounding")

my_id <- 1:30 %>% as.character()

my_x <- rnorm(30, mean = 100, sd = 4)


my_y <- rnorm(30, mean = 100, sd = 10)


our_data <- data_frame(my_id, my_x, my_y) %>%

  pivot_longer(-my_id, names_to = "sample", values_to = "measure")

head(our_data)
```

```
## # A tibble: 6 x 3
##   my_id sample measure
##   <chr> <chr>    <dbl>
## 1 1     my_x      98.0
## 2 1     my_y      99.1
## 3 2     my_x     101.
## 4 2     my_y     118.
## 5 3     my_x      99.7
## 6 3     my_y      98.6
```

- Do not get scared of the R code. What I am doing is generating TWO random sample of 30 numbers.
- Each sample has a mean of 100.

- However, the samples differ in terms of standard deviation (sd) with y having a higher sd than x.
- The set.seed argument is there to ensure that if you run this code on your machine, you and I will get the same samples.
- Given that these are random numbers, without the set.seed argument, you would get different samples every time you run the code. Try removing the set.seed argument and run the code several times and see you get diddering data points.
- The head function gives the first six rows of data. Type ?head in your R console then run for details.
- I DO NOT expect you to memorise the code.

Visualizing the data

- The easiest way to see the spread of a dataset is to draw a histogram.

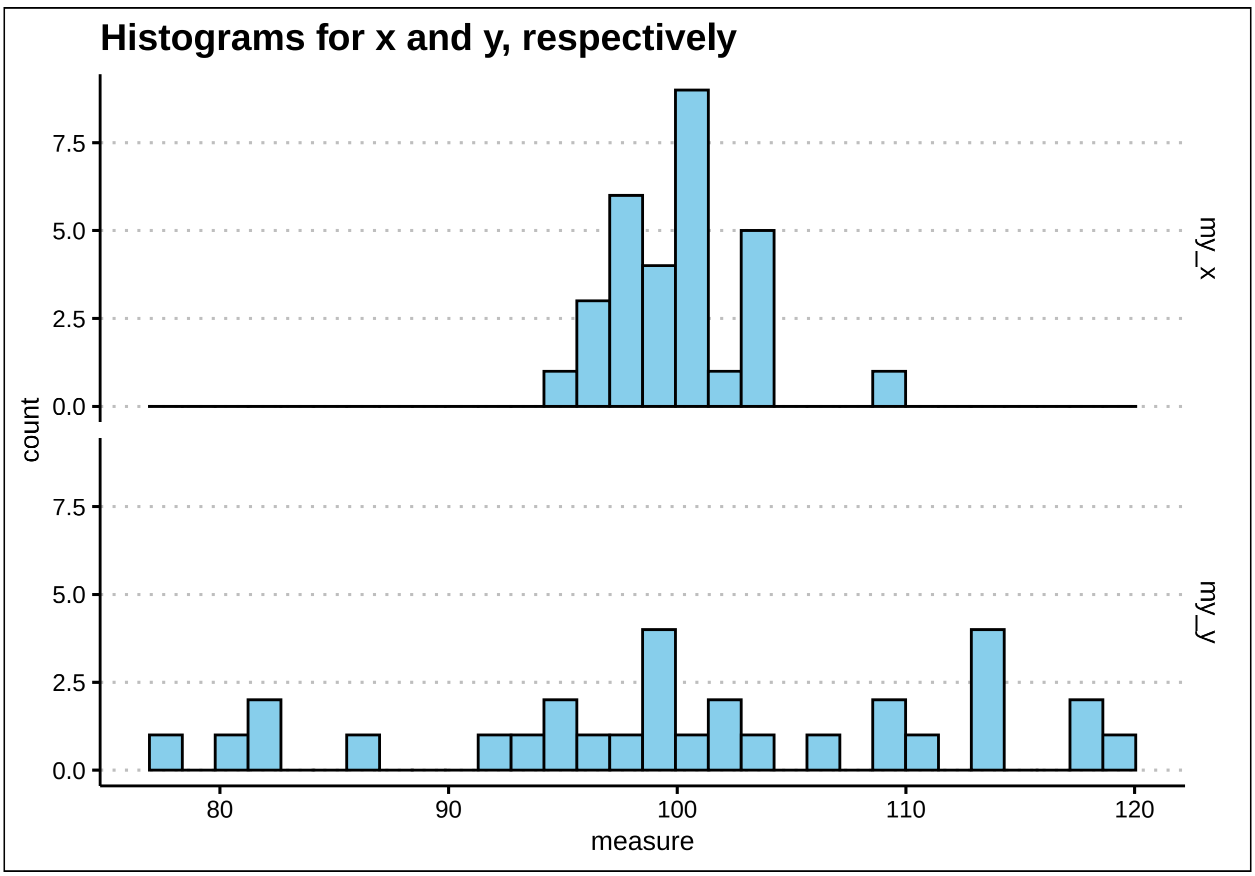
```
our_data %>%

  ggplot(aes(x = measure)) + geom_histogram(col = "black", fill = "skyblue") +

  facet_grid(sample ~ .) +

  ggthemes::theme_clean() +

  labs(title = "Histograms for x and y, respectively")
```



- Note that because x had a lower standard deviation, the values are closer together than those of y which had a larger standard deviation.

- A higher standard deviation and variance means the data is more spread out.
- Again, do not memorize the code, just look at the graphs.

Visulaizing the data

- Another good visuaf for spread is the box plot.

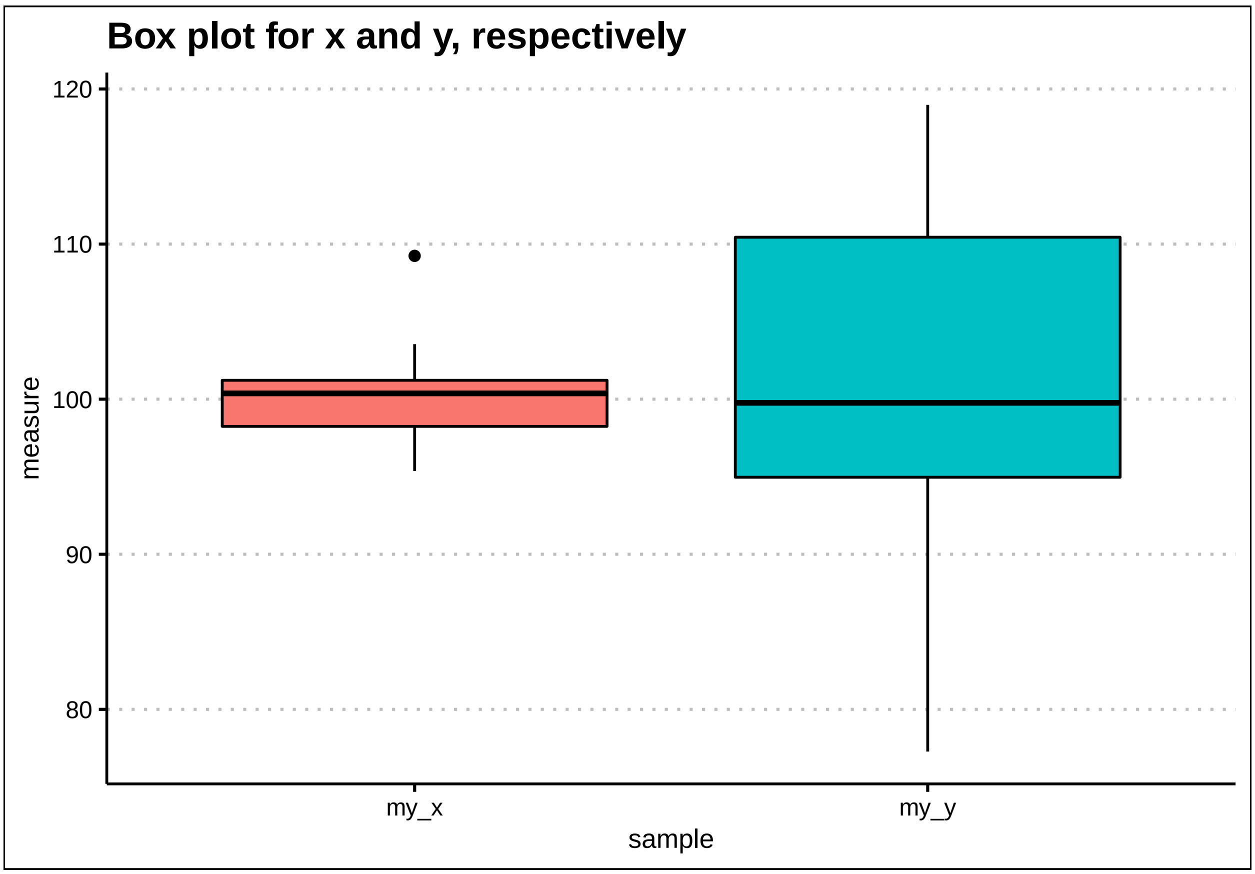
```
our_data %>%

  ggplot(aes(x = sample, y = measure, fill = sample)) + geom_boxplot(col = "black") +

  ggthemes::theme_clean() +

  labs(title = "Box plot for x and y, respectively") +

  theme(legend.position = "none")
```



- can you interprest this boxplot? What five number summaries does it provide?

The variance (var)

- The variance is the average distance each point in the dataset is from the arithmetic mean.
- The formula is as follows.

$\sum_{n=1}^{k} \frac{(x_k - \bar{x})^2}{n-1}$

- The steps for computing the variance are as follows.

- First get the arithmetic mean.
- Take each value and subtract the mean you just calculated to get $(x - \bar{x})$.
- Square the above to get $(x - \bar{x})^2$. Why do we square?
- Sum up the $(x - \bar{x})^2$ and divide by number of observations minus one $(n - 1)$.

Example 1

let us have a simple dataset with 5 observations

```
## # A tibble: 5 x 3
##   X     `(X-X_bar)` `(X - X_bar)^2`
##   <chr> <chr>       <chr>
## 1 1     1-3 = -2    4
## 2 2     2-3 = -1    1
## 3 3     3-3 = 0     0
## 4 4     4-3 = 1     1
## 5 5     5-3 = 2     4
```

- The mean in this case is (1 + 2 + 3 + 4 + 5) / 5 = 15/(5) = 3.
- From every value of x, we subtract 3 to get $(x - \bar{x})$.
- Can you try adding up the $(x - \bar{x})$. Do you see why we need to square?
- Now, for each value of $(x - \bar{x})$, let us square, then sum up; we get 10 $(4 + 1 + 0 + 1 + 4)$.
- Note that the number of observations equals to 5. So our n = 5.
- Remember our variance formula is $\sum_{n=1}^{k} \frac{(x_k - \bar{x})^2}{n-1}$.
- The numerator is 10; the denominator = 5 - 1 = 4
- The variance is $10/4 = 2.5$.
- The standard deviation (sd) is the square root of the variance. $sd = 1.581139$.
- Let us confirm these results using R.

```
variance_example_data <- 1:5

## The colon is a shortcut for writing a series of numbers in R.

## Try typing 1:10 in your console ans see what happens.

## var and sd are the functions for variance and standard deviation in R.

var(variance_example_data)
```

```
## [1] 2.5
```

```
sd(variance_example_data)
```

```
## [1] 1.581139
```

Exercise 1 (10 minutes)

You are given the folowing datasets, compute the variance and standard deviation for each.

```
##  [1]  1  4  7 10 13 16 19 22 25 28
```

```
## [1] 100.74985  92.62990  45.14678  76.03329 111.78181 115.59177
```

The mean absolute deviation (MAD)

- We have seen that the sum of $(x - \bar{x}) = 0$.
- This is why we square in computing variance.
- An alternative would be to ignore the signs.
- Let us ignore the signs in example 1 that we did.
- The mean in this case is $(1 + 2 + 3 + 4 + 5)$ / $5 = 15/(5) = 3$.
- From every value of x, we subtract 3 to get $(x - \bar{x})$.
- Can you try adding up the $(x - \bar{x})$. In computing variance we square this.
- For MAD, instead of squaring, we ignore the signs and take the absolute values of $(x - \bar{x})$. See below

```
## # A tibble: 5 x 3
##   X     `(X-X_bar)` `abs(X - X_bar)`
##   <chr> <chr>       <chr>
## 1 1     1-3 = -2    2
## 2 2     2-3 = -1    1
## 3 3     3-3 = 0     0
## 4 4     4-3 = 1     1
## 5 5     5-3 = 2     2
```

- We add up the $|x - \hat{x}|$ to get $2 + 1 + 0 + 1 + 2 = 6$.
- The numerator is 6; the denominator = 5 - 1 = 4
- The MAD is $6/4 = 1.5$.
- MAD is not as popular as the variance though.

Exercise (10 minutes)

- Compute the MAD for the exercise 1 above.

The Quartiles and the interquatile range

- The quartiles are the two values that occupy the position that divides the dataset into three equal parts.
- NB: As is the case for the median, you must first arrange the data in ascending order.
- Let us revisit example 1 to get the quatiles.
- In example 1, we are lucky that the data is already in ascending order.

```
1:5
```

```
## [1] 1 2 3 4 5
```

- Note that 2 and 4 are the numbers that divide this dataset into three equal parts.
- 2 is the first quatile (or 25th percentile), while 4 is the 3rd quartile (75th percentile).
- 3 is a special kind of a quantile called the median (the 50th percentile). remember this?
- The median divides the data into 2 equal parts.
- Lets confirm this in R. We use the function quantile() and set the probabilities.

```
quantile(1:5, probs = 0.25) # 0.25 gives first quantile
```

```
## 25%
##   2
```

```
quantile(1:5, probs = 0.5) # 0.5 gives second quantile or median.
```

```
## 50%
##   3
```

```
median(1:5) ## Just to confirm median is the second quartile
```

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

```
quantile(1:5, probs = 0.75) # 0.75 gives third quantile.
```

```
## 75%
##   4
```

- The interquantile range is the difference between the 3rd quatile and the first quartile.
- In this case the interquartile range is $4 - 2 = 2$.

exercise 10 minutes

- For exercise 1 above, compute the quatiles and the interquantile range manually.
- For exercise 1 above, compute the quatiles using R. Just copy and paste the code in R as the starting point.
- Compute the measures of center for the datasets, both manually and using R.