Statistical Computing
Week 4: 25. October 2019
Descriptive Statistics: Measures of dispersion

Contents

- Measures of dispersion
  - Sample variance
  - Standard Deviation
  - Range
  - Interquartile range
- Linear transformation

Introduction

When we have a numeric variable in a dataset, the values are almost never constant. The variable has **Variablility**.

Variability is a very important concept in statistics, and is a consequence of *randomness*.

In the probability world, a variable X is a random variable when its value (outcome) is unknown.

In the statistics world, we usually assume that the data come from a random sample (e.g. 20 Berlin residents chosen at random).

Before we have sampled the data, the values are unknown, they are random.

Once the sample is obtained and stored in a data file it is no longer random, but it is a **realisation** of a random process.

Dispersion (Spread, Variability)

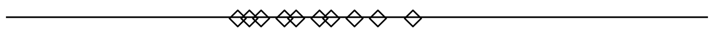
A measure of location tells us where the data sits.
A measure of dispersion tells us how much variability the data has.

Example:
Samples are obtained for two variables X and Y. Both have the same sample size and mean.

$$n_x = 10 \qquad \overline{x} = 404 \qquad n_y = 10 \qquad \overline{y} = 404$$

X: 210, 250, 340, 360, 400, 430, 440, 450, 530, 630

Y: 340, 350, 360, 380, 390, 410, 420, 440, 460, 490



The variability in X is bigger than in Y

Sample Variance

The sample variance for variable X is defined as

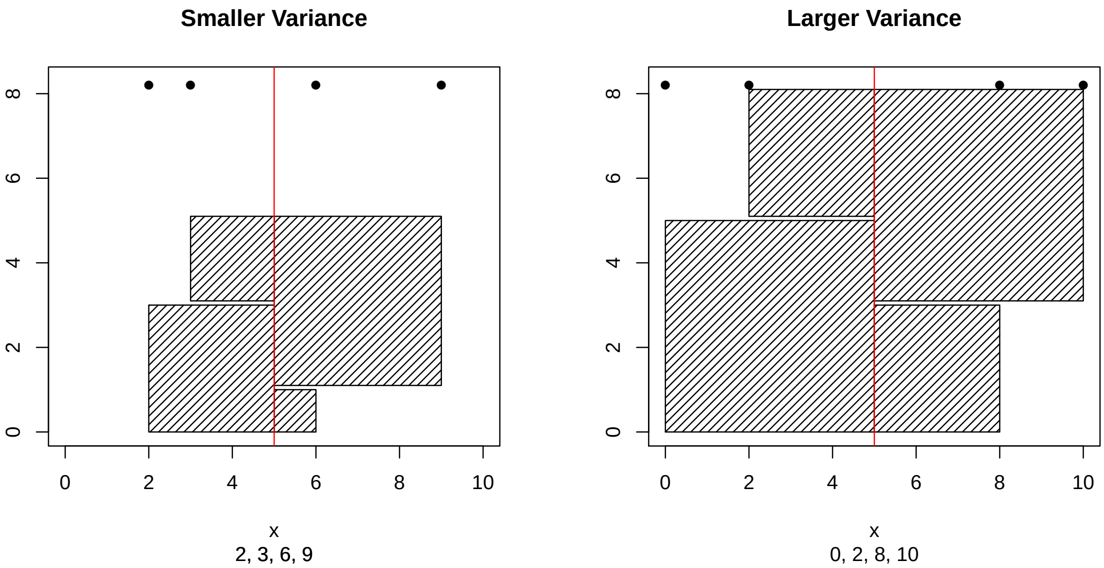
$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

$(x_1 - \overline{x})$ is the deviance from the mean for the first observation.
$(x_i - \overline{x})$ is the deviance from the mean for the i-th observation.
$(x_i - \overline{x})^2$ is the *squared* deviance from the mean for the ith observation ($\geqslant 0$).
$\sum_{i=1}^{n} (x_i - \overline{x})^2$ is the sum of all the *squared* deviances.

Two small examples:

Left sample has values 2,3,6 and 9. Right sample has values 0, 2, 8, 10.

In both cases the mean is $\overline{x} = 5$.



The area of each square corresponds to a squared deviation $(x_i - \overline{x})^2$

The sum of all the squares is bigger in the right diagram because the data points are more spread out.

In R

```
> sum((x-mean(x))^2)/(length(x)-1)
[1] 22.66667
> var(x)
[1] 22.66667
```

Sample and population variance

In the above formula we have $n - 1$ in the denominator, this is for the *sample variance*.

If we know that our data contains the whole population (i.e. all the objects that we are interested in), then we use the *population variance* with n in the denominator.

$$\sigma_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

If we are unclear whether our data contains the whole population, then use the *sample variance*.

Standard Deviation

Another common measure for variability is the standard deviation (sd), which is the square root of the variance.

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Why do we need both standard deviation and variance?

- Maths is easier using the variance.
- Understanding the data is easier with the standard deviation.
- The standard deviation has the same units as the measured variable eg cm.
- The variance has squared units e.g. cm^2, which makes it difficult to interpret the variance.

Interpreting the standard deviation:

A rough rule of thumb is: 95% of the sampled values fall in the interval

$$[\overline{x} - 2s_x; \overline{x} + 2s_x] \qquad \Leftrightarrow \qquad [\overline{x} \pm 2s_x]$$

The approximation is better when the values are roughly symmetric.

R Example.

```
> set.seed(4062875)
> x<-rnorm(100,175,12)
> xbar<-mean(x)
> xbar
[1] 175.0881
> stddev<-sd(x)
> stddev
[1] 13.45704
> xbar-2*stddev
[1] 148.1741
> xbar+2*stddev
[1] 202.0022
> sum(x>=(xbar-2*stddev) & x<=(xbar+2*stddev))
[1] 94
```

$$[\overline{x} - 2s_x; \overline{x} + 2s_x] = [148.2; 202.0]$$

There are 94 out of 100 values that lie in this interval.

Other measures of dispersion

Range Largest value minus the smallest value. $x_{(n)} - x_{(1)}$
Not good! The range is sensitive to outliers and is unstable

```
> x<-rnorm(100,175,12)
> round(diff(range(x)),1)
[1] 58
> x<-rnorm(100,175,12)
> round(diff(range(x)),1)
[1] 54.4
> x<-rnorm(100,175,12)
> round(diff(range(x)),1)
[1] 51.5
> x<-rnorm(100,175,12)
> round(diff(range(x)),1)
[1] 63.3
```

A good measure of dispersion should give similar values for each of these samples.

Interquartile range: Quantile definition

Last week you learnt about quantiles: Recap...

The median $x_{0.5}$ divides the data values into two halves.

The 0.1-quantile $x_{0.1}$ is chosen so that $p = 0.1$ (10%) of the data values are less than or equal to $x_{0.1}$

The p-quantile x_p is chosen so that the proportion p of the data values are less than or equal to x_p

Interquartile range: IQR

The quartiles are two special quantiles.

The lower quartile Q_1 is the $p = 0.25$ quantile, and
the upper quartile Q_3 is the $p = 0.75$ quantile.

The quartiles and the median divide the data values into four equally sized groups. The median is sometimes called the second or middle quartile.

The **interquartile range** is the difference between upper and lower quartiles.
$IQR = Q_1 - Q_3$

```
> x<-5:15
> diff(quantile(x,c(0.25,0.75)))
75%
  5
> IQR(x)
[1] 5
```

Measures of dispersion and linear transformation

Last week we considered what happens to the mean and median after a variable is linear transformed.

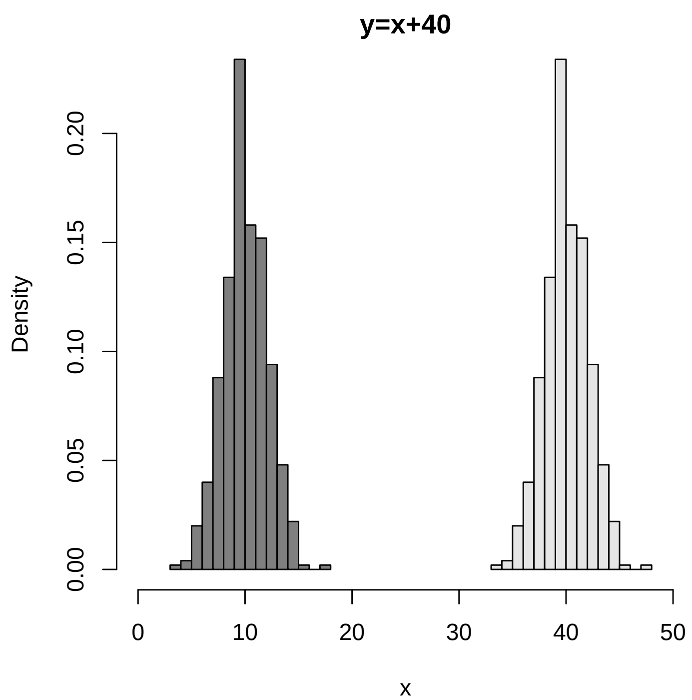
Reminder: Suppose the variable X is transformed into a new variable Y using the formula $Y = aX + b$, where a and b are known constants.

If $\quad y_i = ax_i + b \quad$ for $i = 1, \ldots, n$,

then $\quad \overline{y} = a\overline{x} + b$

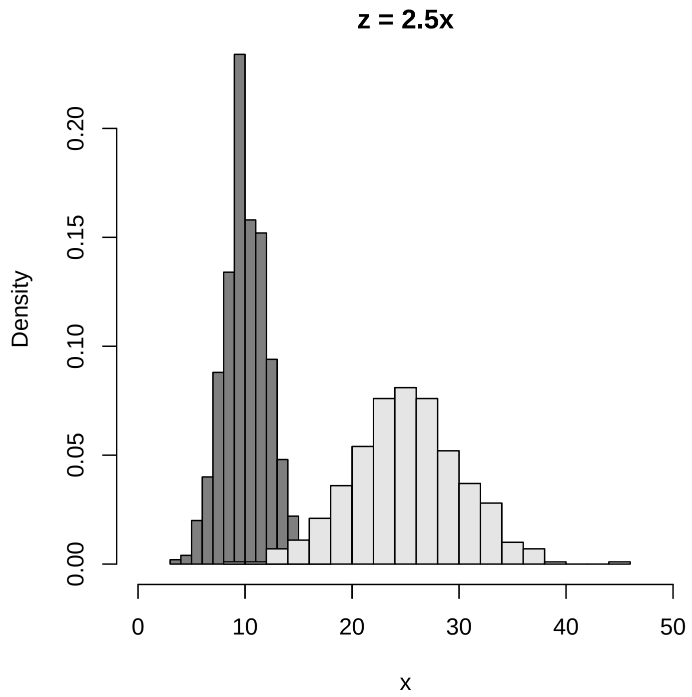
and $\quad y_{0.5} = ax_{0.5} + b.$

For the standard deviation there is a similar formula.
It is important to understand that the shifting the position of data has no effect on how spread out the data is.



The sd of x and y is 2.

But multiplying the data by a factor does change the spread.

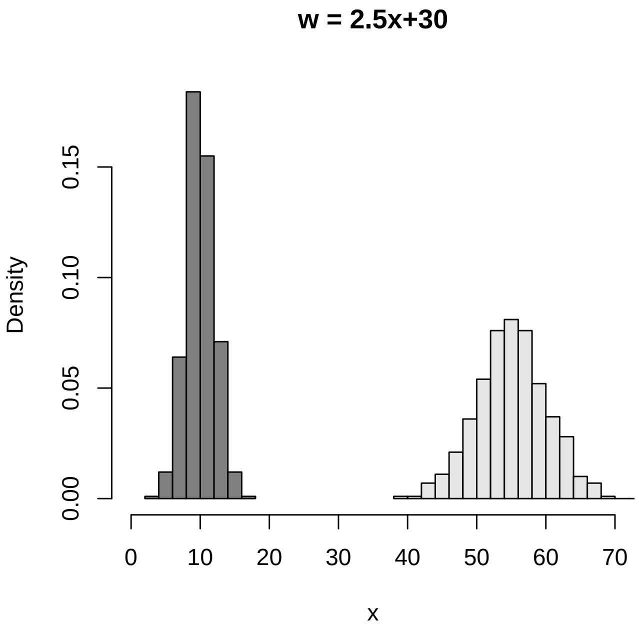


The s.d. of x is 2 the s.d. of z is 5.

Which gives the formula

If	$y_i = ax_i + b$	for $i = 1, \ldots, n,$
then	$s_y = as_x$	

Note that b has no effect on the s.d.



The s.d. of x is 2 the s.d. of w is $2.5 \cdot 2 = 5$.

To summarise: if the variable Y is a **linear transformation** of variable X, which means there relationship can be written using formula

$$y_i = ax_i + b,$$

with a and b constants, then the following statistics can be directly transformed using these Formulae:

Mean	$\bar{y} = a\bar{x} + b$
Median	$y_{0.5} = ax_{0.5} + b$
Standard deviation	$s_y = as_x$
Variance	$s_y^2 = a^2 s_x^2$
Range	$\text{Range}_y = a\text{Range}_x$
IQR	$\text{IQR}_y = a\text{IQR}_x$

RHD Short Workshop on Measures of dispersion