Statistical Computing

Prof. Dr. Tim Downie

Week 6

13. November 2017

Basic Statistics: Variance, quantiles and Correlation

Stand: Version: 12. November 2017

Today

- Lecture: Basic Statistics
  - Population and Sample
  - Measures of location
  - Measures of dispersion (Variance)
  - Quantiles
  - Correlation
- Info for the R Test on 3 December
- Workshop
- Homework

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 the data, the vaues 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.

We can apply many of the ideas from probability theory to the collected data, called descriptive statistics.

The first example of this is measuring the amount of variability:
Probability: Variance of a random variable X
Statistics: *sample variance* of the variable $x_1, x_2, \ldots x_n$

I'll fill in a few gaps before looking into sample variance in detail.

Population and Sample

When analysing data, you should consider what the total **population** is.

This is the group which in an ideal world we would have complete data on. For example: In Week 4 we assumed that the population is *all Berlin residents*.

> This implies that we don't want to say anything about people living in Sachsen.
>
> It also implies that we do want to include people who are not Berliners/German citizens but who are resident in the city.

A sample[1] is a subset of the population.

All *elements* in the sample belong to the population.

E.g. 20 Berlin residents is a sample of the population.

In this example the **sample size** is 20, (n=20).

[1] sometimes called a "sample population". I won't use this term as it is too easy to confuse with the total population

Example: Population and Sample

- Population: All leukaemia patients in Germany
- Valid samples
  - All German patients with AML-leukaemia
  - All leukaemia patients in Charite Hospital Berlin.
  - A collection of 50 Leukaemia patients treated in German hospitals, entered into a medical study.
- Not a valid sample
  - All Patients in the Charite
  - Leukeamia patients in the EU.

Notation

Numeric Variable has the name X

The Population size is N
The Sample size is n with $n < N$

A realized sample of X from the population: $x_1, x_2, \ldots x_n$ (lower case letters)
Are the values obtained after the data has been collected.

If the sample is e obtained in the future, or the data has been sampled but we don't yet know the values, we use uppercase letters:
$X_1, X_2, \ldots X_n$
X_1 etc. are all random variables.

Ordered sample:
It is convenient to have a notation for smallest, second smallest, ... largest value in the sample. This is done by putting the index number in $(\cdot)$:
$x_{(1)}, x_{(2)}, \ldots x_{(n)}$

Measures of location: mean

An average of a numeric variable is a measure of location it tells in one number us where the data lies.

The **mean** is a the most common measure of location.

Notation: Mean of the sample from X is

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n} \qquad \text{alternative} \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

In R:

```
> sum(x)/length(x)
> mean(x)
```


Measures of location: median

The median divides the data in two.
Half of the values are greater than the median and half are smaller than the median.

Notation:

- For odd n $\qquad x_{0.5} = x_{\left(\frac{n+1}{2}\right)}$
- For even n $\qquad x_{0.5} = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)$

In R:

```
> median(x)
```

The median is less sensitive to outliers than the mean

Dispersion (Spread, Variability, deu: Streuung)

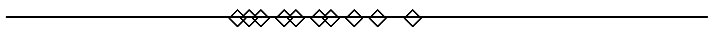
A measure of location tells us where the data sits.
A measure of dispersion tells us how much variability the data has.

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

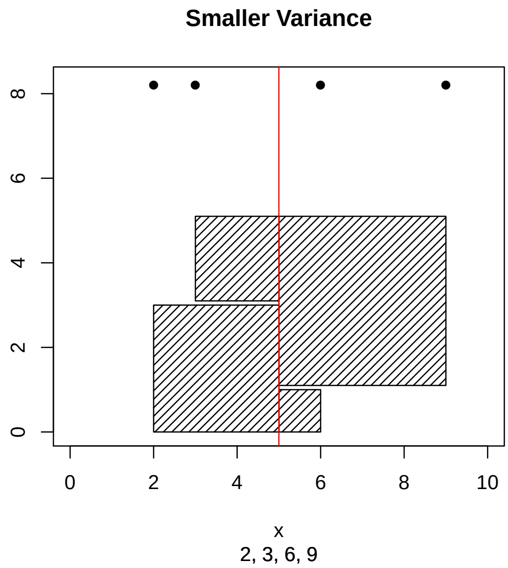
$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

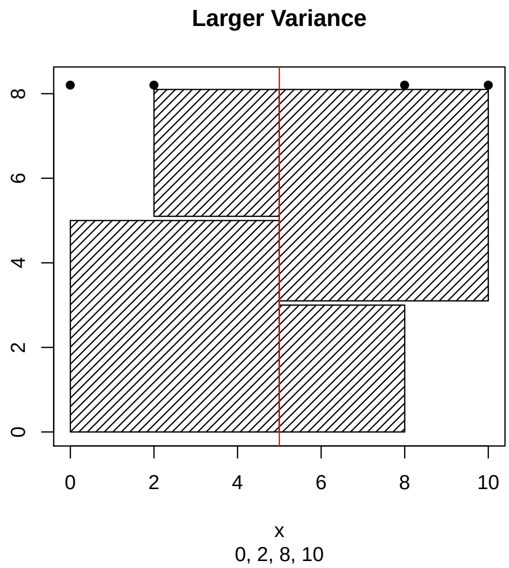
$(x_1 - \overline{x})$ is the deviance from the mean for the first observation.
$(x_i - \overline{x})$ is the deviance from the mean for the ith observation.
$(x_i - \overline{x})^2$ is the *squared* deviance from the mean for the ith observation ($\geqslant 0$).
$\sum_{i=1}^{n} (x_i - \overline{x})^2$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

Standard Deviation

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

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

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.
  This makes it difficult to interpret the variance.

Interpreting the standard deviation:

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

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

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

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

```
> x<-5:15
> quantile(x,0.5)
50%
 10
> quantile(x,0.1)
10%
  6
> quantile(x,c(0.1,0.11))
10% 11%
6.0 6.1
```


Interquartile range: IQR

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 known as 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
```

Covariance and Correlation

This subject was covered in detail in Worksheet 3

Variance for variable X

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

Covariance for variables X and Y

$$s_{xy} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$

The Correlation coefficient is standardised so that it takes values between -1 and $+1$.

$$r_{x,y} = \frac{s_{xy}}{s_x . s_y}$$

Properties:

- $-1 \leqslant r \leqslant 1$
- $r_{x,y}$ measures the linear relationship between X and Y
- $r_{x,y} = r_{y,x}$