Disspersion

Variance

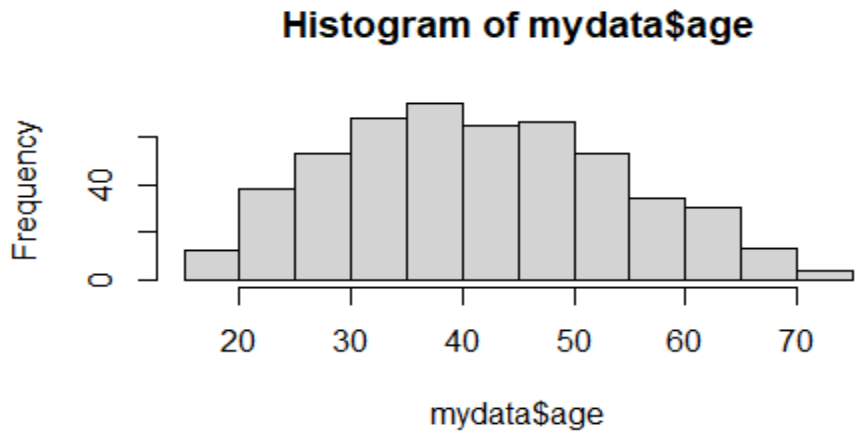
The variance is a numerical measure of how the data values is dispersed around the mean. In particular, the sample variance is defined as:

$$\sigma^2 = \frac{\sum(\chi - \mu)^2}{N}$$

`var(x)`

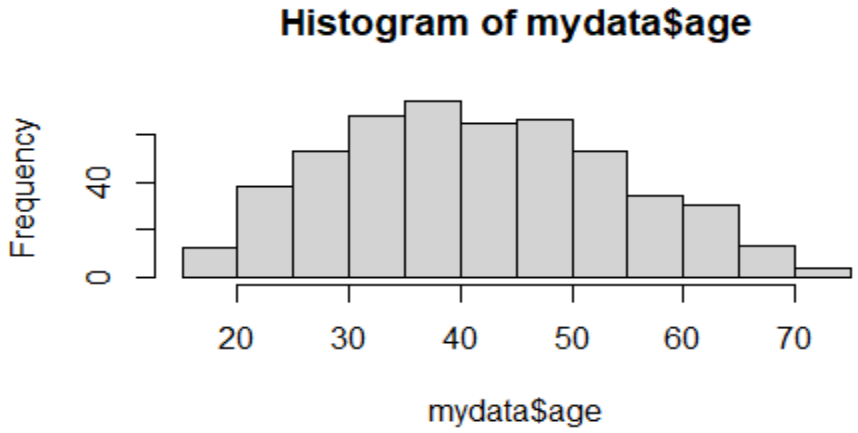
`sum((x-mean(x))^2)/(length(x))`

```
> var(mydata$age)
[1] 154.4563
> mean(mydata$age)
[1] 42.05882
```



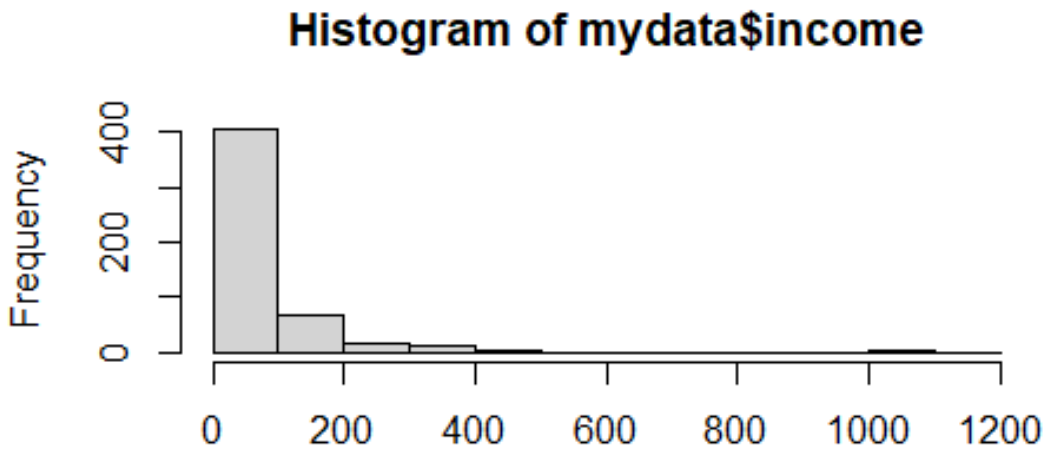
```
> stem(mydata$age)

  The decimal point is 1 digit(s) to the right of the |

  1 | 89999
  2 | 000000011111222223333334444444444
  2 | 55555555555566666666777788888888888888999999999999999999
  3 | 000000011111111111112222222222222333333334444444444444444
  3 | 55555555555555566666666666666667777777777777777888888888888899999999
  4 | 000000000000001111111111111122222222222222233333333334444444444444444
  4 | 55555555566666666666777777777777778888888888888888999999999999
  5 | 00000000011111111111122222222333333333344444444444
  5 | 555555555555566666667777888888889999999
  6 | 000000111111111111223333333334444
  6 | 5556666666678
  7 | 0001123
```

```
> var(mydata$age)
[1] 154.4563
> mean(mydata$age)
[1] 42.05882
```



High variance

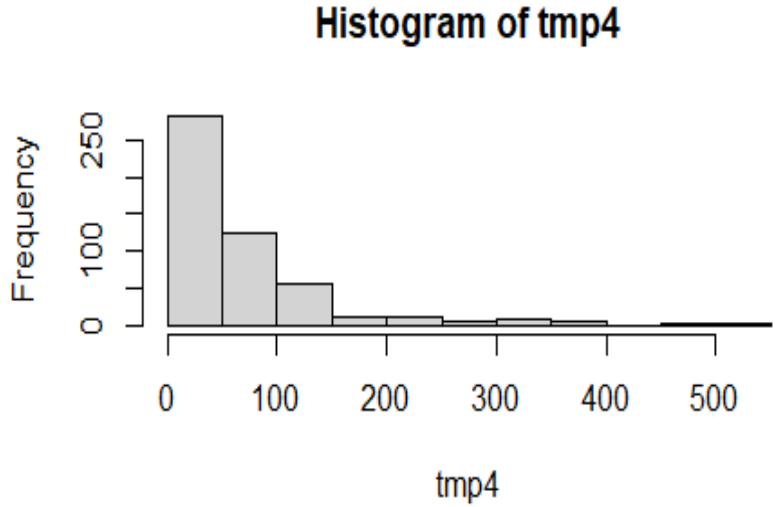
```
> var(mydata$income)
[1] 12612.51
> mean(mydata$income)
[1] 78.59412
```




```
> stem(mydata$income)

  The decimal point is 2 digit(s) to the right of the |

   0 | 1111111111111111122222222222222222222222222222222222222222222+320
   1 | 00000000000001111111122222222222233333334444444444445555555567
   2 | 000001111134444578 99
   3 | 0112355555789
   4 | 8
   5 | 04
   6 |
   7 |
   8 | 4
   9 |
  10 | 57
  11 | 2
```

```
> tmp4=mydata$income[mydata$income<600]
> str(tmp4)
 int [1:506] 72 153 28 26 23 76 40 57 24 89 ...
```

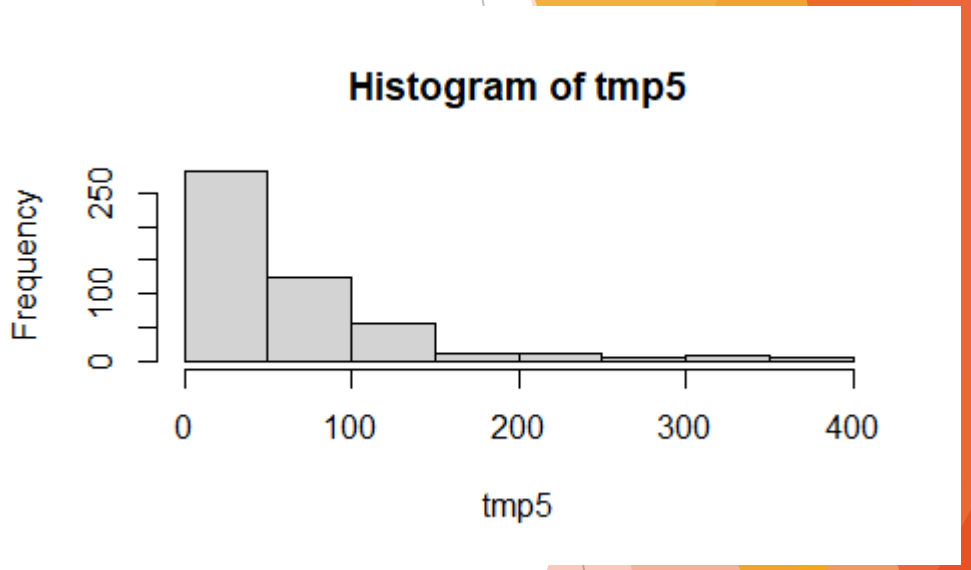
```
> stem(tmp4)

  The decimal point is 2 digit(s) to the right of the |

  0 | 11111111111111111222222222222222222222222222222222222222222222222+173
  0 | 55555555555555555555555555555555555555555566666666666666666666666666+67
  1 | 00000000000001111111122222222222233333344444444444444
  1 | 555555556788
  2 | 000001111134444
  2 | 57899
  3 | 01123
  3 | 55555789
  4 |
  4 | 8
  5 | 04
```




```
> var(tmp4)
[1] 5591.5
> mean(tmp4)
[1] 71.17589
```

```
> tmp5=mydata$income[mydata$income<400]
> str(tmp5)
 int [1:503] 72 153 28 26 23 76 40 57 24 89 ...
```

```
> stem(tmp5)

  The decimal point is 1 digit(s) to the right of the |

   0 | 990001111113333444555556666677788999999999
   2 | 000000011111111222222222333333333333344444444445555555556666666666+94
   4 | 000000000000011111112222222223333333455555666666777889999990000000011+21
   6 | 000011111223344556667788888999990001111222334456677[illegible]
   8 | 11123456666678899999111234578
  10 | 00122223344579991244567778888 9
  12 | 02336990034557789
  14 | 00133334567122349
  16 | 76
  18 | 1668
  20 | 1478903
  22 | 758
  24 | 024
  26 | 2
  28 | 0688
  30 | 74
  32 | 13
  34 | 56924
  36 | 16
  38 | 3
```




```
> var(tmp5)
[1] 4483.626
> mean(tmp5)
[1] 68.5825
```

Standard deviation

The standard deviation is the square root of its variance.

```
sd(dat1)
[1] 22.06706
```

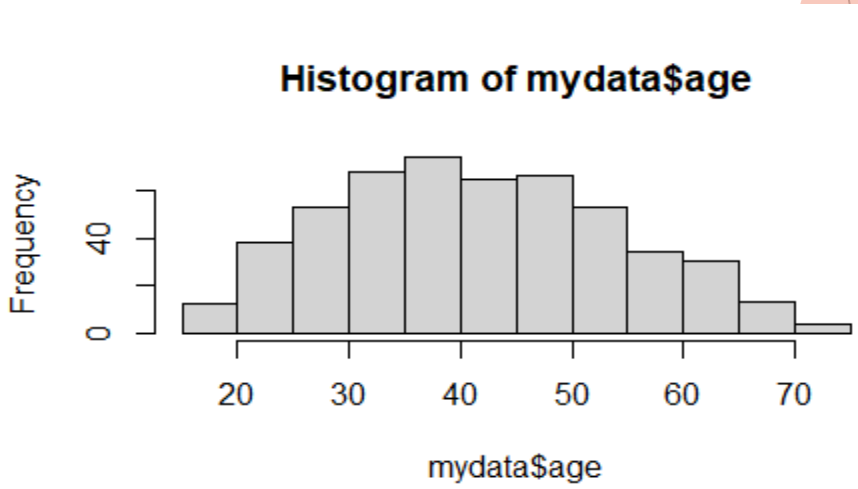
$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

The standard deviation is a value in the same units as the original values, which makes easier to understand

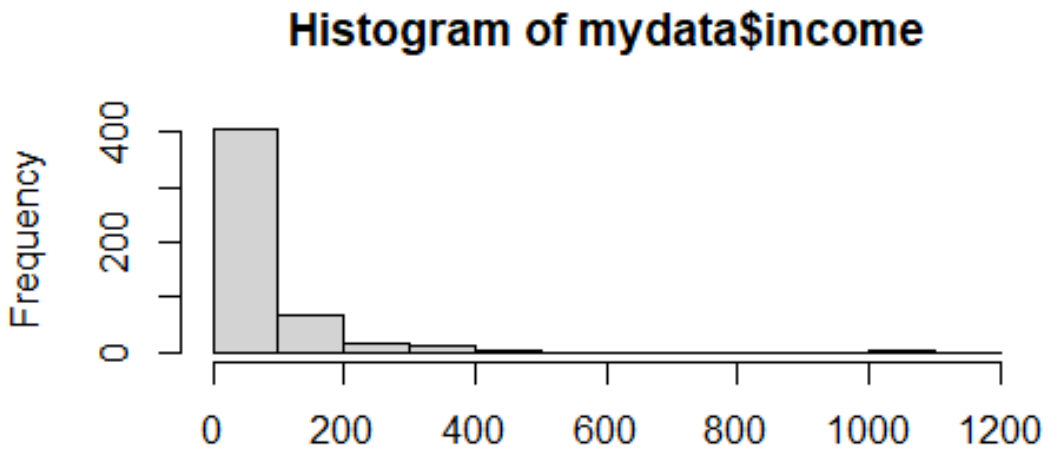
```
> stem(mydata$age)

  The decimal point is 1 digit(s) to the right of the |

  1 | 89999
  2 | 00000001111122222333333444444444
  2 | 55555555555566666666777788888888888888999999999999999999
  3 | 000000011111111111112222222222222233333333444444444444444444
  3 | 55555555555555566666666666666667777777777777777888888888888899999999
  4 | 000000000000001111111111111122222222222222233333333334444444444444444
  4 | 55555555566666666667777777777777788888888888888888999999999999
  5 | 00000000011111111111122222222333333333344444444444
  5 | 555555555555566666667777888888889999999
  6 | 000000111111111111223333333334444
  6 | 5556666666678
  7 | 0001123
```

```
> var(mydata$age)
[1] 154.4563
> mean(mydata$age)
[1] 42.05882
```

```
> sd(mydata$age)
[1] 12.42804
```



```
> var(mydata$income)
[1] 12612.51
> mean(mydata$income)
[1] 78.59412
```

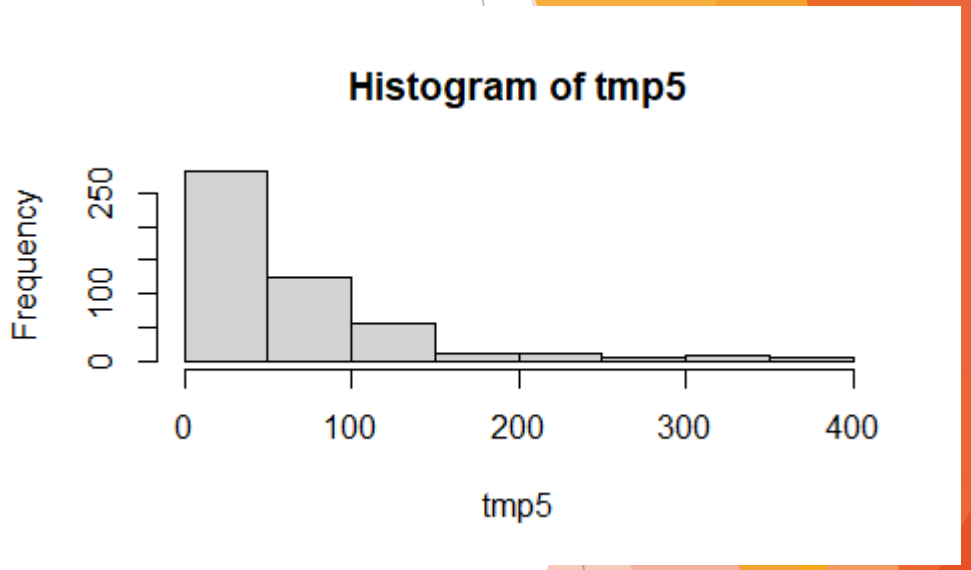



```
> stem(mydata$income)

The decimal point is 2 digit(s) to the right of the |

  0 | 1111111111111111122222222222222222222222222222222222222222222+320
  1 | 00000000000001111111122222222222233333334444444444445555555567
  2 | 000001111134444457899
  3 | 0112355555789
  4 | 8
  5 | 04
  6 |
  7 |
  8 | 4
  9 |
 10 | 57
 11 | 2
```

```
> sd(mydata$income)
[1] 112.3054
```

```
> tmp5=mydata$income[mydata$income<400]
> str(tmp5)
 int [1:503] 72 153 28 26 23 76 40 57 24 89 ...
```

```
> stem(tmp5)

  The decimal point is 1 digit(s) to the right of the |

  0 | 990001111113333444555556666677788999999999
  2 | 000000011111111222222222333333333333344444444445555555556666666666+94
  4 | 000000000000011111112222222223333333455555666666777889999990000000011+21
  6 | 000011111223344556667788888999990001111222334456677888888
  8 | 11123456666678899999111234578
 10 | 00122223344579991244567778888 9
 12 | 02336990034557789
 14 | 00133334567122349
 16 | 76
 18 | 1668
 20 | 1478903
 22 | 758
 24 | 024
 26 | 2
 28 | 0688
 30 | 74
 32 | 13
 34 | 56924
 36 | 16
 38 | 3
```




```
> sd(tmp5)
[1] 66.95988
```

```
> var(tmp5)
[1] 4483.626
> mean(tmp5)
[1] 68.5825
```

Coefficient of variation

Coefficient of variation

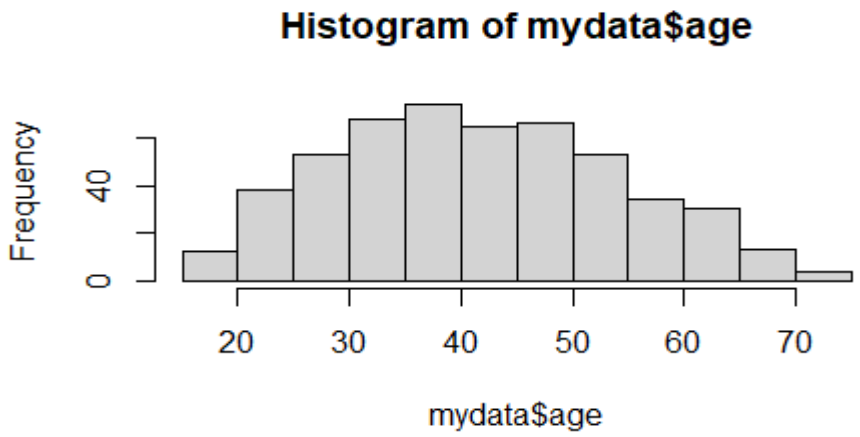
$$= \frac{\text{Standar deviation}}{\text{mean}}$$

```
> stem(mydata$age)

  The decimal point is 1 digit(s) to the right of the |

  1 | 89999
  2 | 00000001111122222333333444444444
  2 | 55555555555566666666777788888888888888999999999999999999
  3 | 00000001111111111111222222222222233333333444444444444444
  3 | 55555555555555566666666666666677777777777777788888888888899999999
  4 | 00000000000000111111111111122222222222222333333333344444444444444
  4 | 55555555566666666666777777777777778888888888888888999999999999
  5 | 00000000011111111111122222223333333333444444444
  5 | 55555555555556666666777788888888999999999
  6 | 00000011111111111122333333333444444
  6 | 5556666666678
  7 | 0001123
```

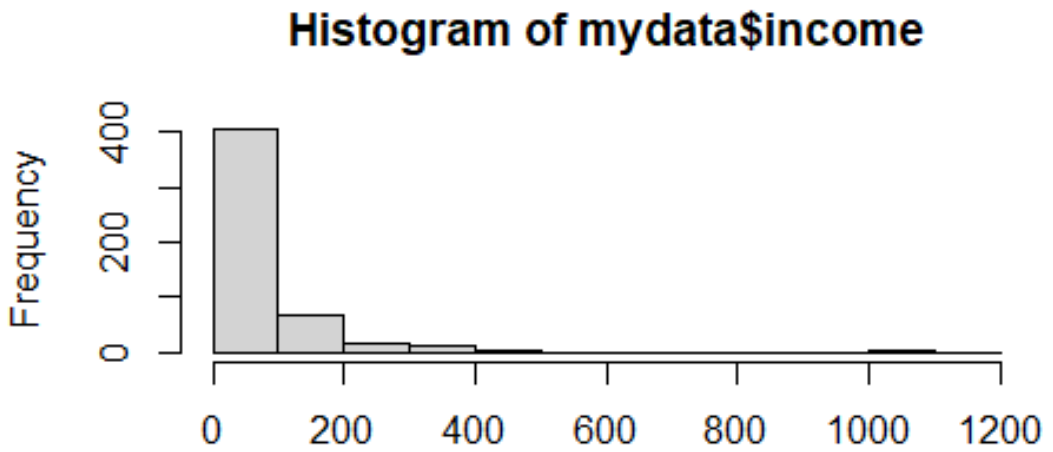
```
> var(mydata$age)
[1] 154.4563
> mean(mydata$age)
[1] 42.05882
```

```
> sd(mydata$age)
[1] 12.42804
```

```
> sd(mydata$age)/mean(mydata$age)
[1] 0.2954919
```



```
> var(mydata$income)
[1] 12612.51
> mean(mydata$income)
[1] 78.59412
```



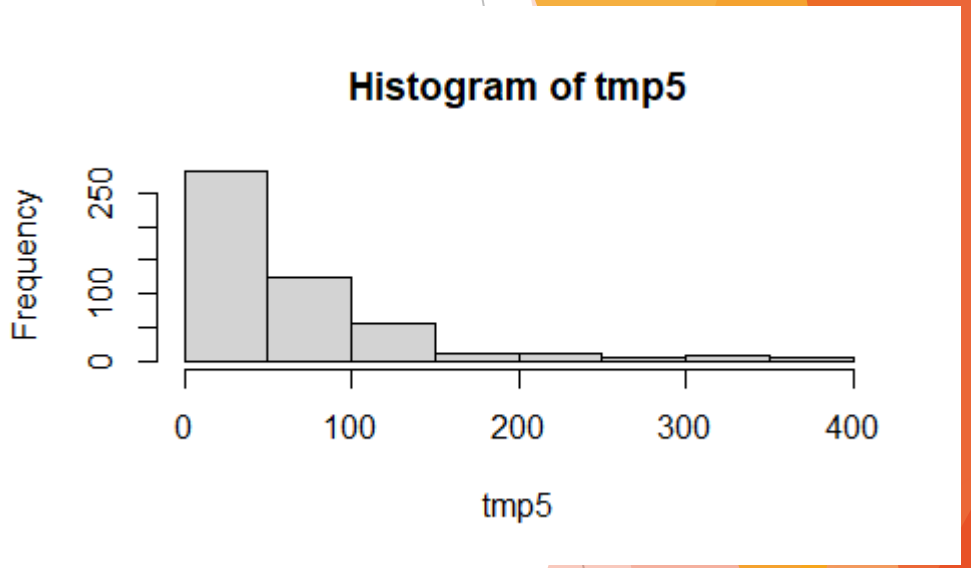

```
> stem(mydata$income)

The decimal point is 2 digit(s) to the right of the |

  0 | 11111111111111111222222222222222222222222222222222222222222222+320
  1 | 00000000000001111111122222222222233333334444444444445555555567
  2 | 000001111134444457899
  3 | 0112355555789
  4 | 8
  5 | 04
  6 |
  7 |
  8 | 4
  9 |
 10 | 57
 11 | 2
```

```
> sd(mydata$income)/mean(mydata$income)
[1] 1.428929
```

```
> sd(mydata$income)
[1] 112.3054
```

```
> tmp5=mydata$income[mydata$income<400]
> str(tmp5)
 int [1:503] 72 153 28 26 23 76 40 57 24 89 ...
```

```
> stem(tmp5)

  The decimal point is 1 digit(s) to the right of the |

   0 | 990001111113333444555556666677788999999999
   2 | 000000011111111222222222333333333333344444444445555555556666666666+94
   4 | 00000000000001111111222222222333333345555566666677788999999000000011+21
   6 | 000011111223344556667788888999990001111222334456677888 8
   8 | 11123456666678899999111234578
  10 | 00122223344579991244567778888 9
  12 | 02336990034557789
  14 | 00133334567122349
  16 | 76
  18 | 1668
  20 | 1478903
  22 | 758
  24 | 024
  26 | 2
  28 | 0688
  30 | 74
  32 | 13
  34 | 56924
  36 | 16
  38 | 3
```




```
> sd(tmp5)/mean(tmp5)
[1] 0.9763406
```

```
> var(tmp5)
[1] 4483.626
> mean(tmp5)
[1] 68.5825
```

```
> sd(tmp5)
[1] 66.95988
```

CV

$$\frac{10}{168} = 0.059$$

$$\frac{10}{17} = 0.58$$

Same standard deviation, different meaning over the data

Neural networks