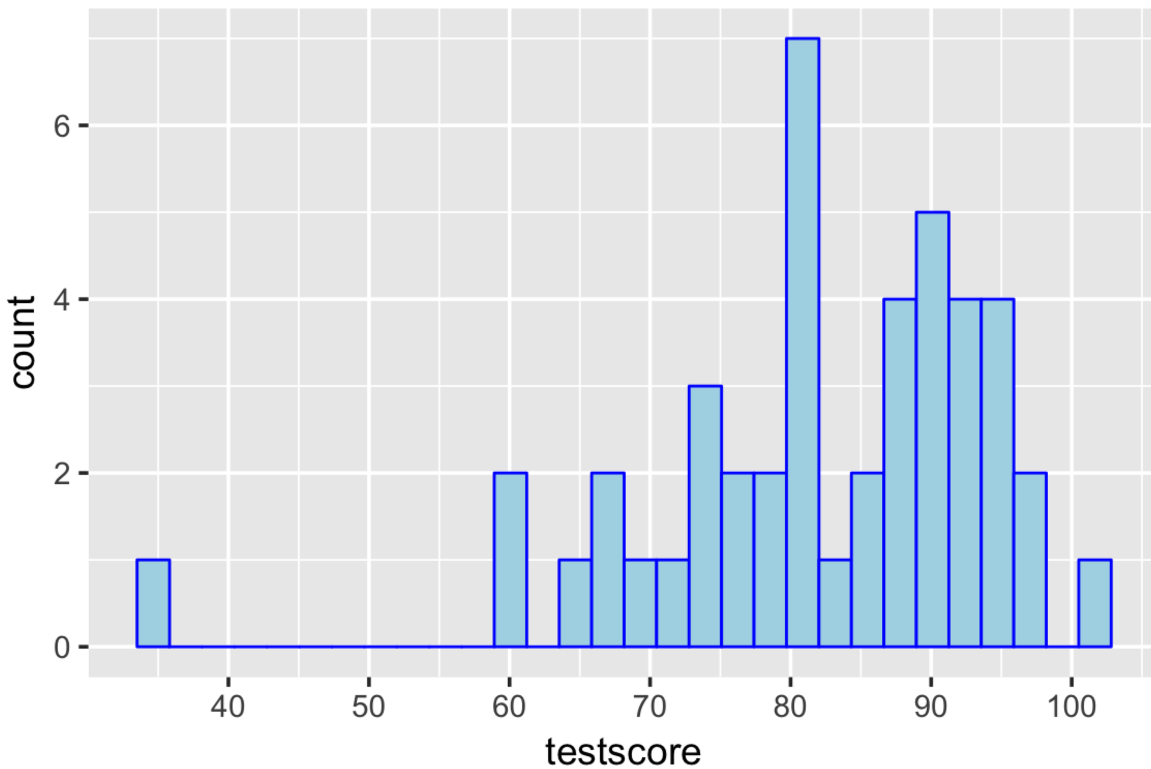
# Five number summary

1. min
2. lower fourth
3. median
4. upper fourth
5. max

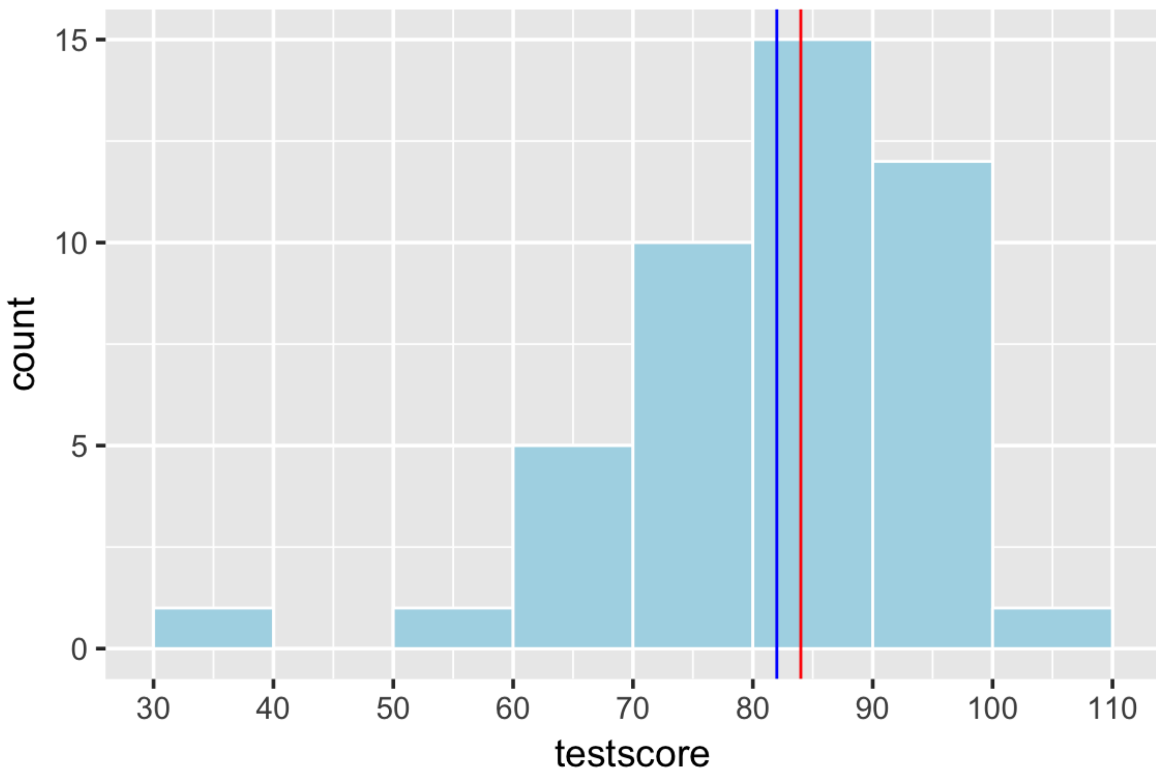
```r
summary(prices)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     379     506     572     593     699     799
```

# Test score data



# Fewer bins



# Test score dataset

Original data set of scores:
35, 59, 61, 64, 66, 66, 70, 72, 73, 74, 75, 76, 76, 78, 79, 80, 80, 81, 81, 82, 82, 82, 84, 86, 86, 88, 88, 88, 88, 89, 89, 90, 91, 91, 92, 92, 92, 92, 94, 94, 94, 94, 96, 98, 102

**Mean: 82**

**Median: 84**

Trimmed dataset (min and max removed):
59, 61, 64, 66, 66, 70, 72, 73, 74, 75, 76, 76, 78, 79, 80, 80, 81, 81, 82, 82, 82, 84, 86, 86, 88, 88, 88, 88, 89, 89, 90, 91, 91, 92, 92, 92, 92, 94, 94, 94, 94, 96, 98

**Mean: 82.63** (corrected)

**Median: 84** (corrected)

How much was trimmed? $\frac{1}{45}$ = 2.22%

# Trimmed means

Suppose we want to **trim 15%**.

.15 x 45 = 6.75 values

**Trim 6:**

$\frac{6}{45}$ = 0.133

$\overline{x}_{tr(13.33)}$ = 83.667

**Trim 7:**

$\frac{7}{45}$ = 0.156

$\overline{x}_{tr(15.56)}$ = 83.774

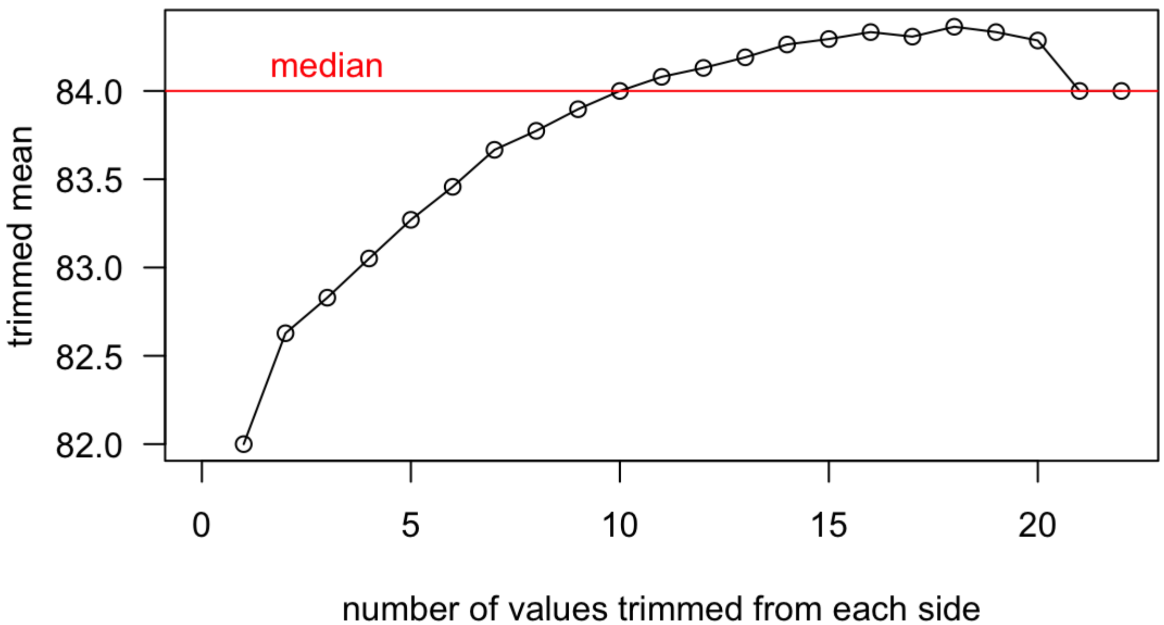
**Interpolate:**

83.667 + .75 * (difference) =

83.667 + .75 * (83.774 - 83.667) =

83.667 + .75 * (.107) = **83.747**

# Median vs. trimmed mean



# Sample and population means

population mean: $\mu$ = sum of N population values / N

sample mean: $\bar{x} = \frac{x_1+x_2+\ldots+x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$

population median: $\tilde{\mu}$

sample median: $\tilde{x}$

# Measures of variability

**deviations from the mean**

$$x_1 - \bar{x},\ x_2 - \bar{x},\ etc.$$

Data: 3, 8, 11, 14
Mean: 9

| $value$ | $deviation$ | $deviation^2$ |
|---|---|---|
| 3 | -6 | 36 |
| 8 | -1 | 1 |
| 11 | 2 | 4 |
| 14 | 5 | 25 |

**Sum of squared deviations**

$S_{xx}$: 36 + 1 + 4 + 25 = 66

**Population variance**

$\sigma^2$ = 66/4 = 16.5

$$\sigma^2 = \sum_{i=1}^{N} (x_i - \mu)^2 / N$$

# Sample variance

**Sum of squared deviations**:

$S_{xx}$: 36 + 1 + 4 + 25 = 66

**Sample variance**:

$s^2$ = 66 / **3** = 22

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

**Why n-1?**

Short answer: using **n** would result in an underestimation, since the values in the sample are closer to the sample mean than to the true population mean (which we don't know)

# Standard deviation

**Square root of variance**

- Population s.d. = $\sqrt{\sigma^2}$
- Sample s.d. = $\sqrt{s^2}$
- *same units as original values*
- Variance of test scores: 156.636
- Standard deviation of test scores: 12.515

# EXERCISE (p. 47, #62)

Consider the following information on ultimate tensile strength ($lb/in^2$) for a sample of $n = 4$ hard zirconium copper wire specimens:

$\overline{x} = 76{,}831$
$s = 180$
smallest $x_i = 76{,}683$
largest $x_i = 77{,}048$

Set up equations to determine the values of the two middle sample observations. *Do not solve.*

# EXERCISE: sd for n = 3

Find the sample mean, variance, and standard deviation:

| X1 | X2 | X3 | mean | var | sd |
|---|---|---|---|---|---|
| 1 | 2 | 3 | | | |
| 2 | 4 | 6 | | | |
| 0 | 5 | 10 | | | |
| 99 | 100 | 101 | | | |
| -8 | -5 | -2 | | | |

# Standard deviation, n = 4

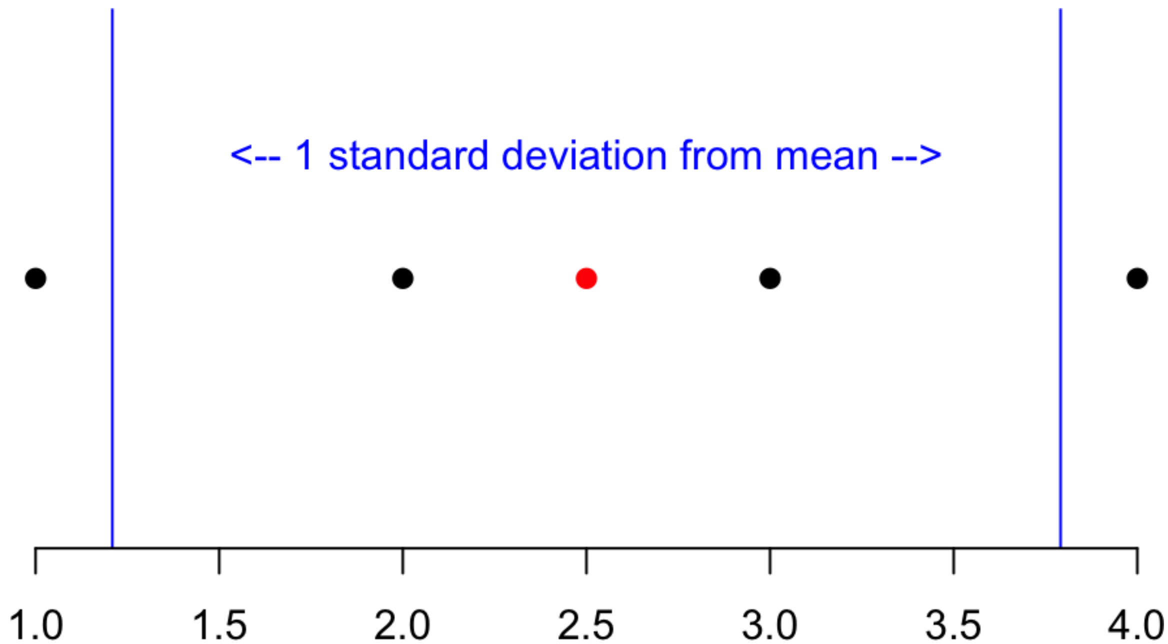
| | **X1** | **X2** | **X3** | **X4** |
|---|---|---|---|---|
| set1 | 1 | 2 | 3 | 4 |

# Standard deviation, n = 4

| | X1 | X2 | X3 | X4 | mean | var | sd |
|---|---|---|---|---|---|---|---|
| set1 | 1 | 2 | 3 | 4 | 2.5 | 1.67 | 1.29 |

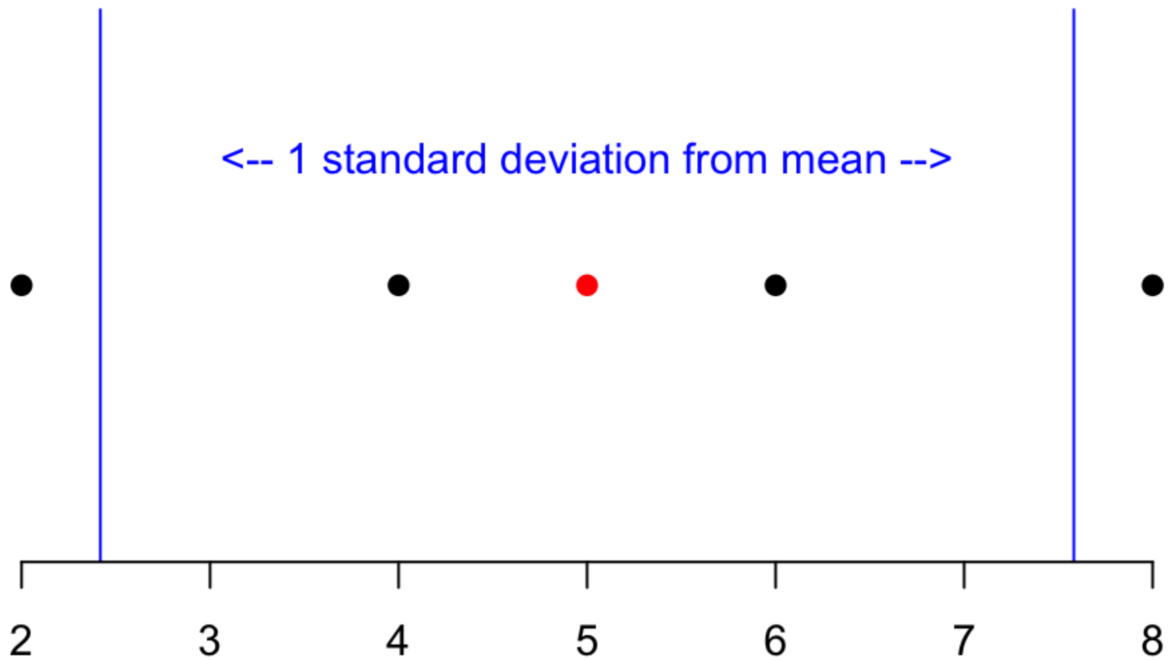
# Standard deviation, n = 4

| | X1 | X2 | X3 | X4 | mean | var | sd |
|---|---|---|---|---|---|---|---|
| set1 | 1 | 2 | 3 | 4 | 2.5 | 1.67 | 1.29 |



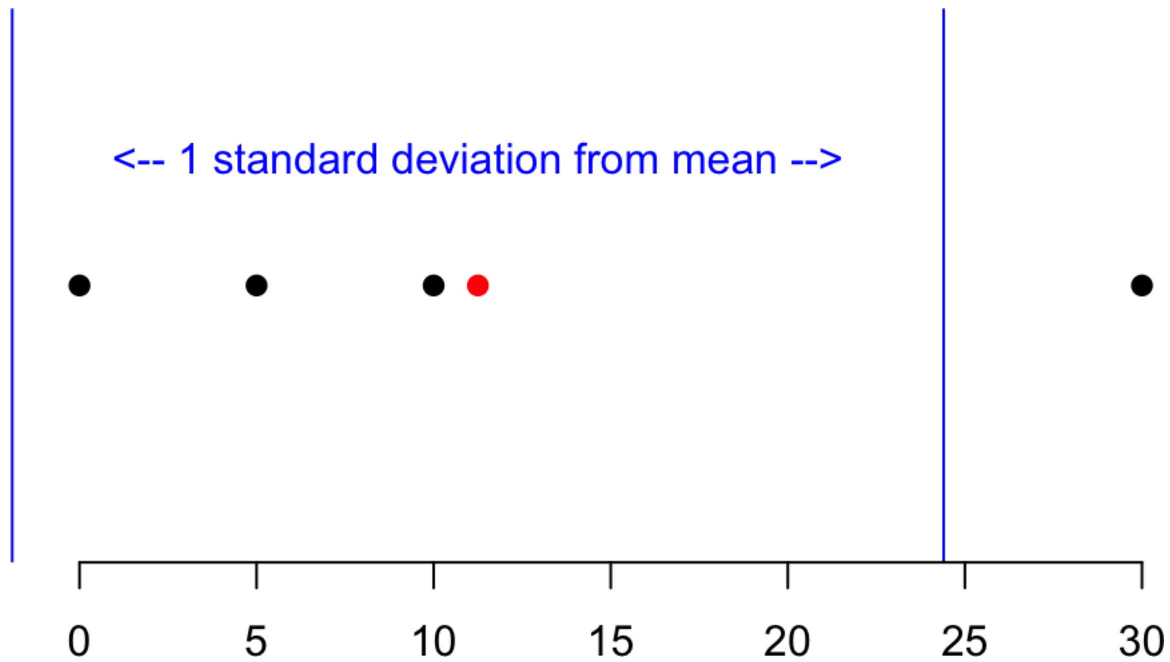
# Standard deviation, n = 4

| | **X1** | **X2** | **X3** | **X4** | **mean** | **var** | **sd** |
|---|---|---|---|---|---|---|---|
| set2 | 2 | 4 | 6 | 8 | 5 | 6.67 | 2.58 |



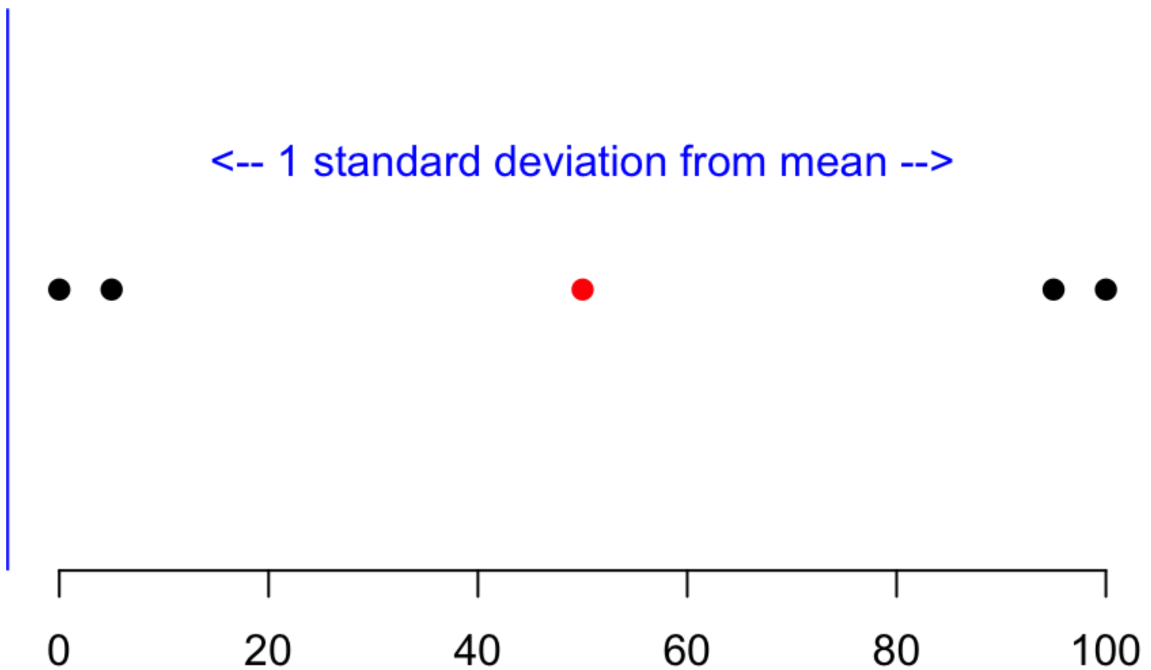
# Standard deviation, n = 4

| | X1 | X2 | X3 | X4 | mean | var | sd |
|---|---|---|---|---|---|---|---|
| set3 | 0 | 5 | 10 | 30 | 11.2 | 173 | 13.2 |



# Standard deviation, n = 4

| | X1 | X2 | X3 | X4 | mean | var | sd |
|---|---|---|---|---|---|---|---|
| set4 | 0 | 5 | 95 | 100 | 50 | 3017 | 54.9 |



# STAT UN1201 – Chapter 2

Prof. Joyce Robbins

# Probability

In 1654, writer Antoine Gombaud “Chevalier de Méré” wanted to know if the following bets are profitable:

- getting at least one six on 4 dice rolls
- getting at least one double-six on 24 dice rolls

# Vocabulary (2.1)

- **experiment** – process whose outcome is subject to uncertainty
  (ex. rolling a die)
- **sample space** – set of all possible outcomes of an experiment
  S = {1, 2, 3, 4, 5, 6}

# Experiment with an infinite sample space

- ex. flip a coin until you get tails
- **sample space**
  S = {T, HT, HHT, HHHT, …}
- **event**
  you get tails in less than 8 flips
  A = {T, HT, HHT, HHHT, HHHHT, HHHHHT, HHHHHHT}

# Vocabulary (2.1)

- **event** – *collection* of outcomes contained in the sample space
- **simple event** – one outcome (ex. getting a 5)
  A = {5}
- **compound event** – more than one outcome (ex. rolling > 3)
  B = {4, 5, 6}