Basics of R:
Descriptive statistics

Research Methods for Human Inquiry

Andrew Perfors

Gladly's survey

Remember we loaded up Gladly's survey! Make sure you're in `gladlysurvey_modified.Rmd` — we're going to do something with the data now!

1. What year were you born?

2. What is your favourite food?

3. On a scale of 1-10, how much would you want to eat a carrot right now?

4. On a scale of 1-10, how much would you want to eat cake right now?

What can you do with data?

Two different kinds of things to do:

- Descriptive statistics:
  - Saying something about this specific data set
  - Not trying to draw any broader conclusions

- Inferential statistics:
  - Using your data to learn something more general
  - Making guesses about a broader set of people, situations or events

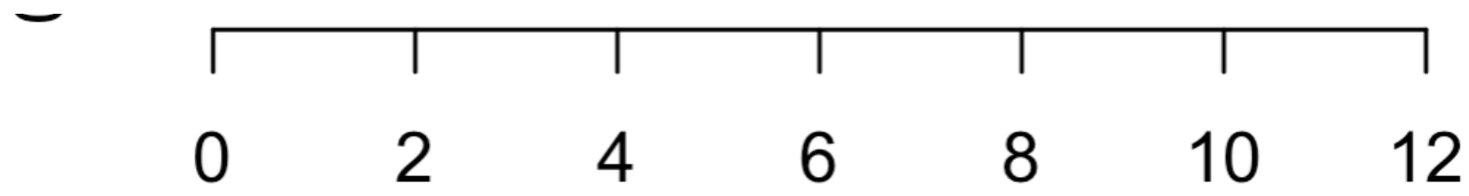
Descriptive statistics

```
> gdata$age
 [1]  8  6  3  5  7  7  3  5  5  5  7  7  6  3  4  3  3  8
[19]  4  4  6  4  4  4  2  3  5  6  5  8  2  9 10  1 11
```

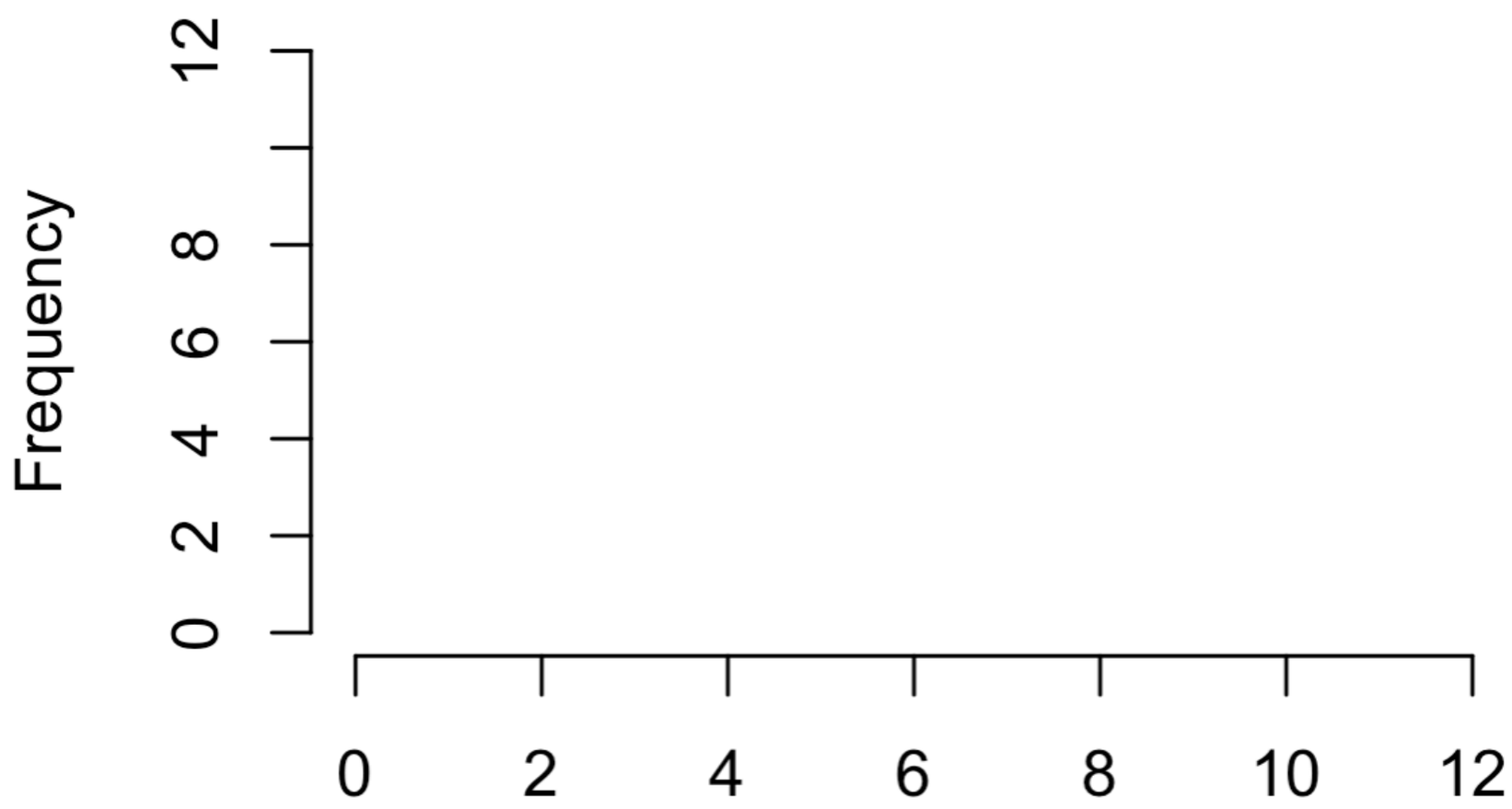
Let there be data

```
> gdata$age
 [1]  8  6  3  5  7  7  3  5  5  5  7  7  6  3  4  3  3  8
[19]  4  4  6  4  4  4  2  3  5  6  5  8  2  9 10  1 11
```

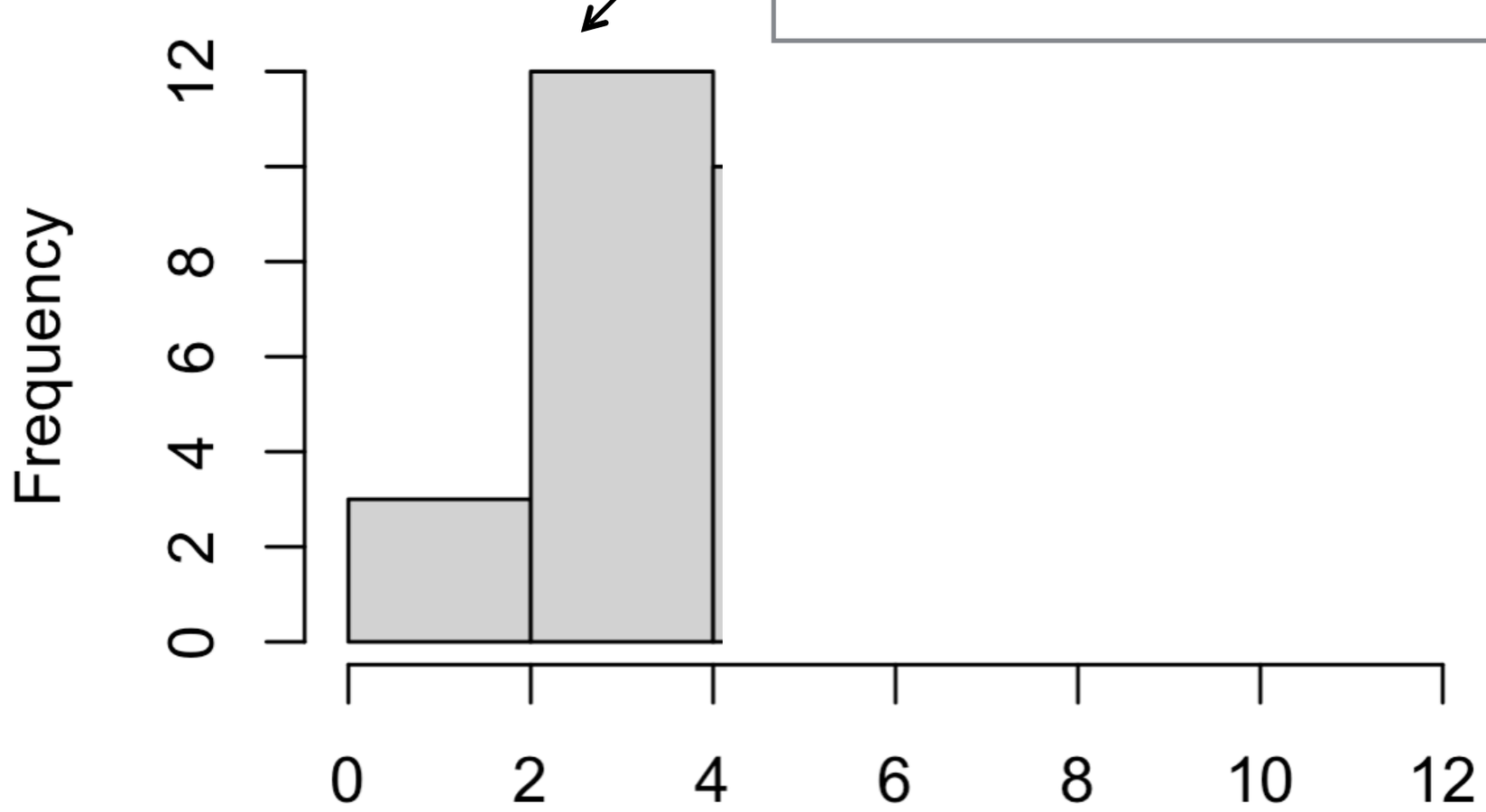
We divide it up into **bins**



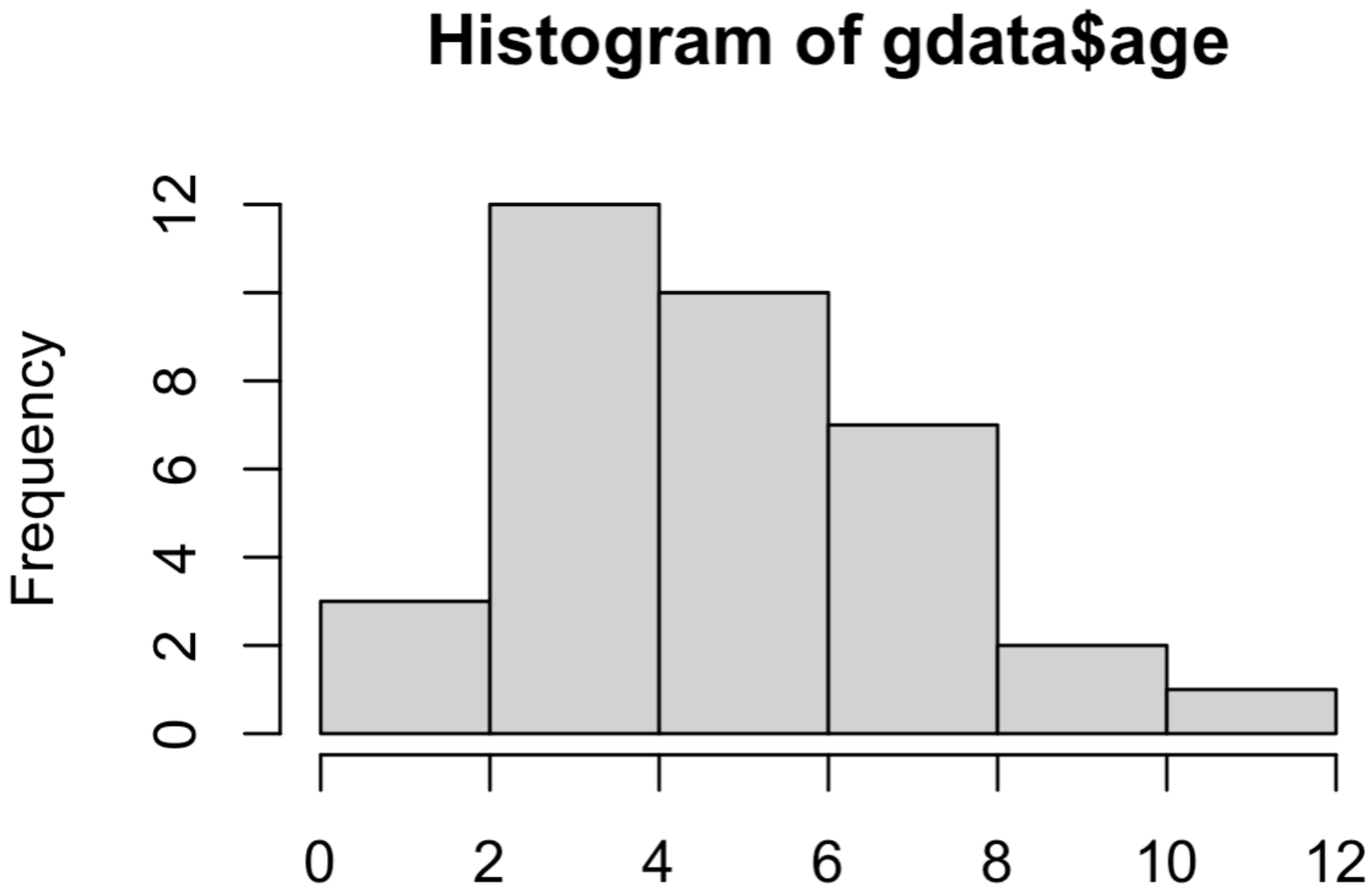
We count the number of observations that fall within each bin, called the **frequency**



Height of the bars describe the frequency with which observations fall in the bin



Here's the one I made in R



Histograms in R

- The function we need is `hist()`
- Some arguments we can specify:

  `x` = the data to be plotted **(mandatory!)**

  `breaks` = how many "breaks" between bins?

  `xlab` = label on the x-axis

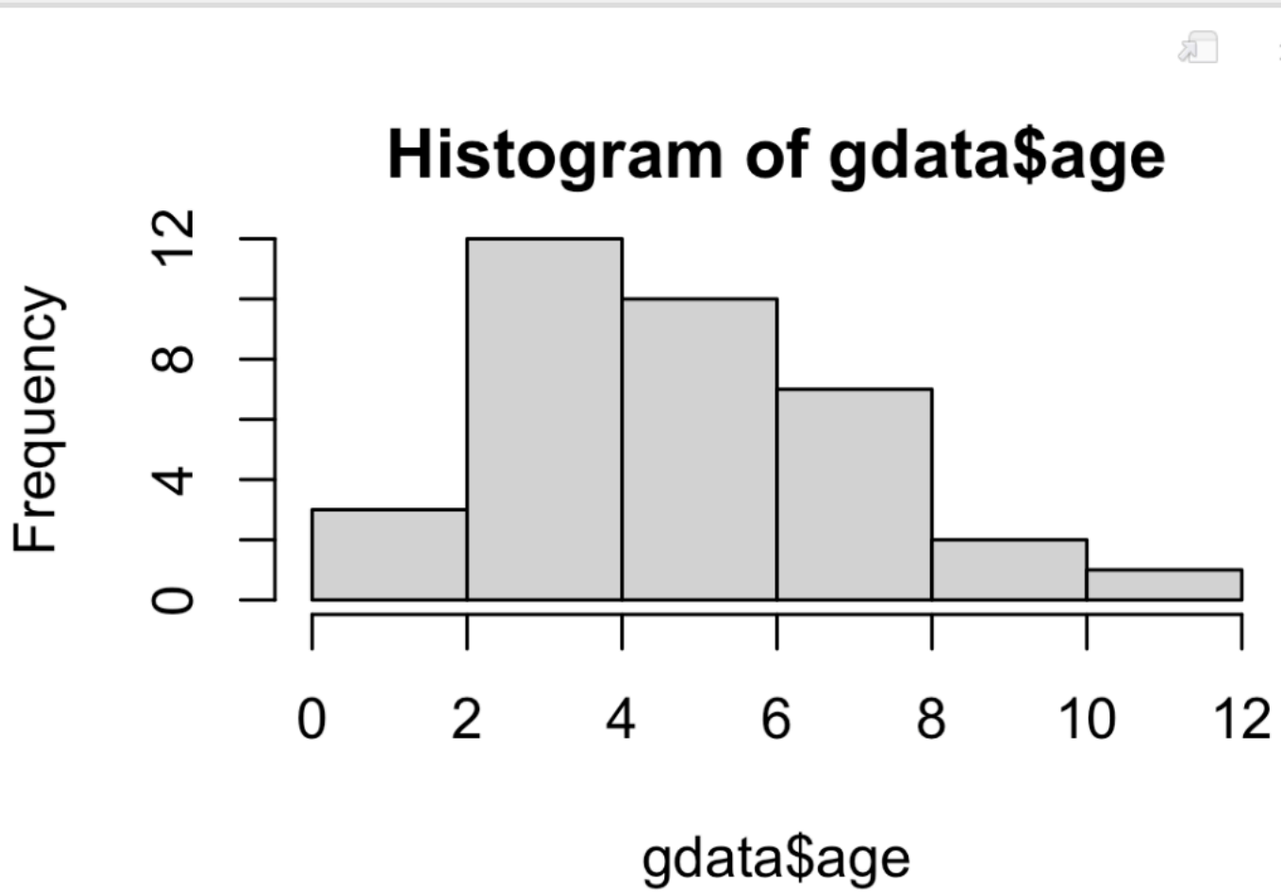
  `ylab` = label on the y-axis

  `main` = title of the plot

  `col` = colour of the bars

Add it to your Markdown file

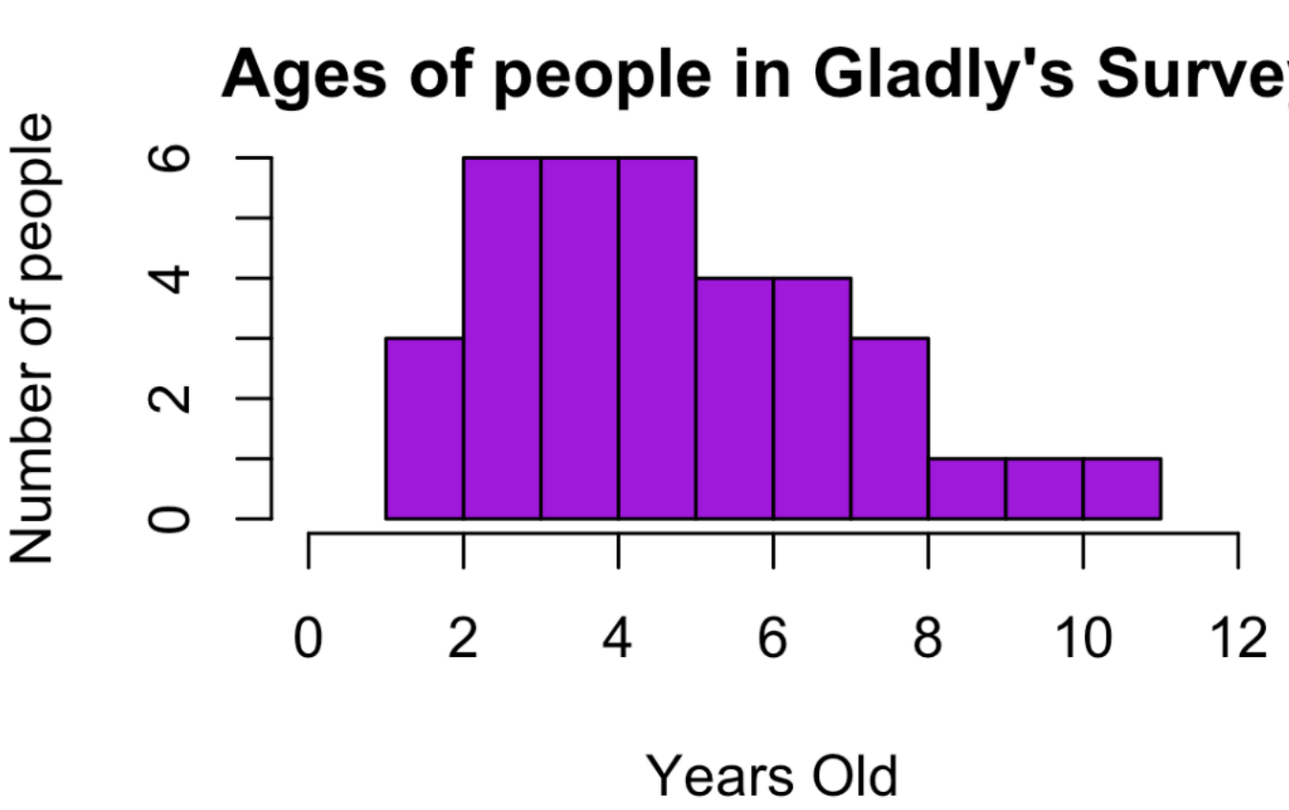
```{r gdhistogram}
hist(x=gdata$age)
```



When you press ‘Play’, it should show the figure!

Add optional arguments to make it nicer

```{r gdhistogrammod}
hist(x=gdata$age,main="Ages of people in Gladly's
Survey",xlab="Years Old",ylab="Number of
people",col="darkviolet",breaks=11,xlim=c(0,12))
```

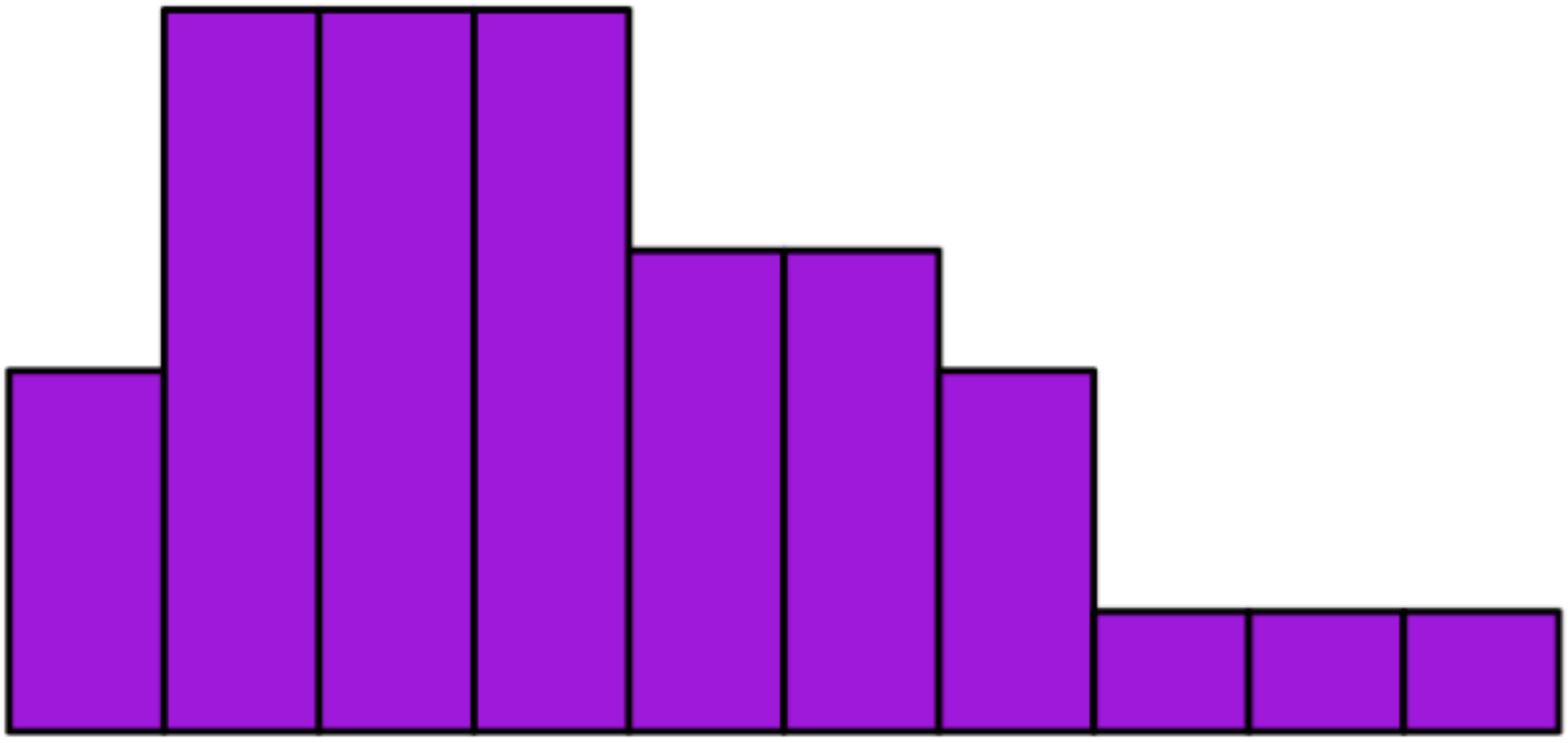


But how do you know what the optional arguments are?

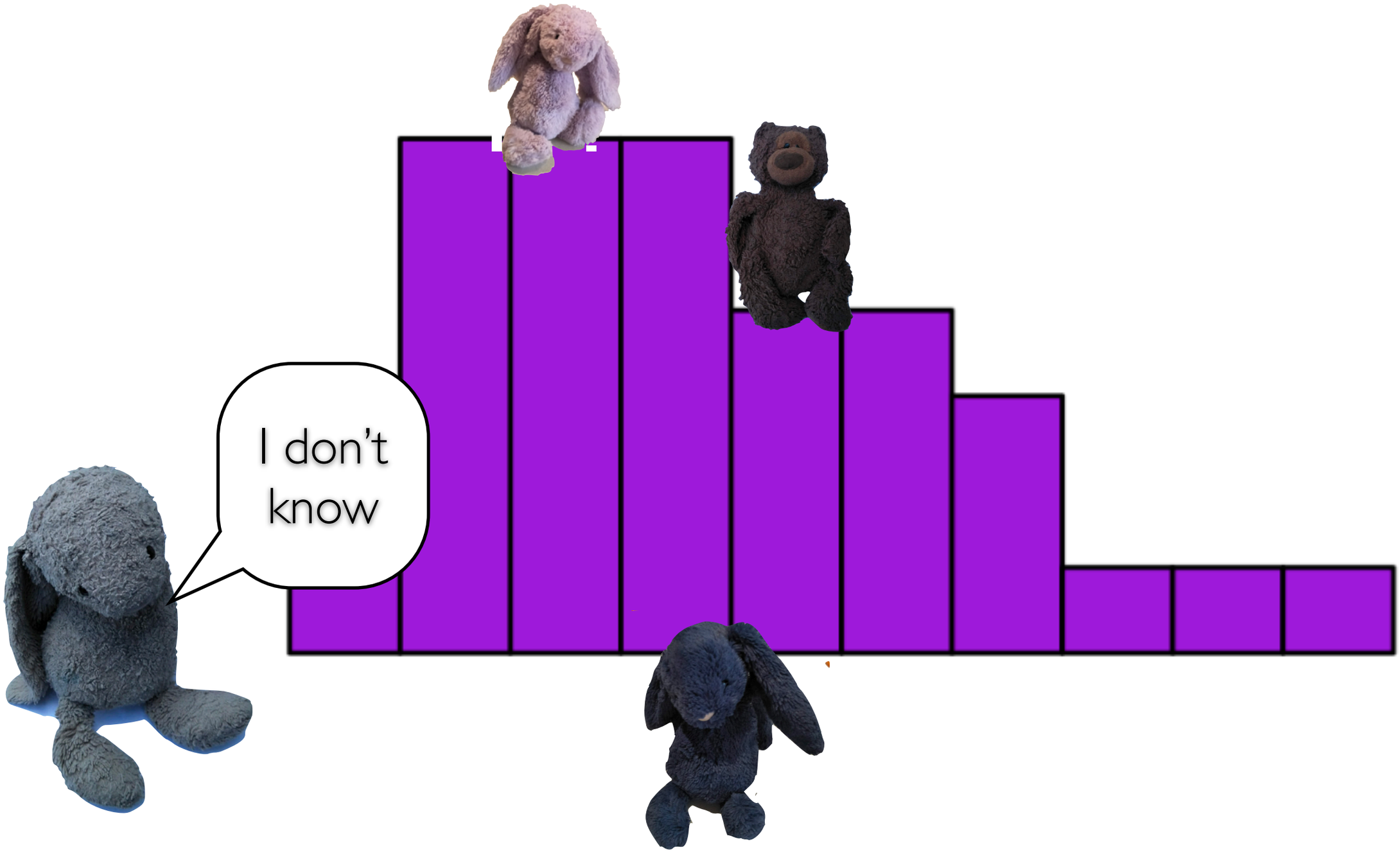
Suppose I want to tell you about the dataset without showing you the entire histogram.

Suppose I just want a number. Or maybe two.

Where is the centre part of the dataset?

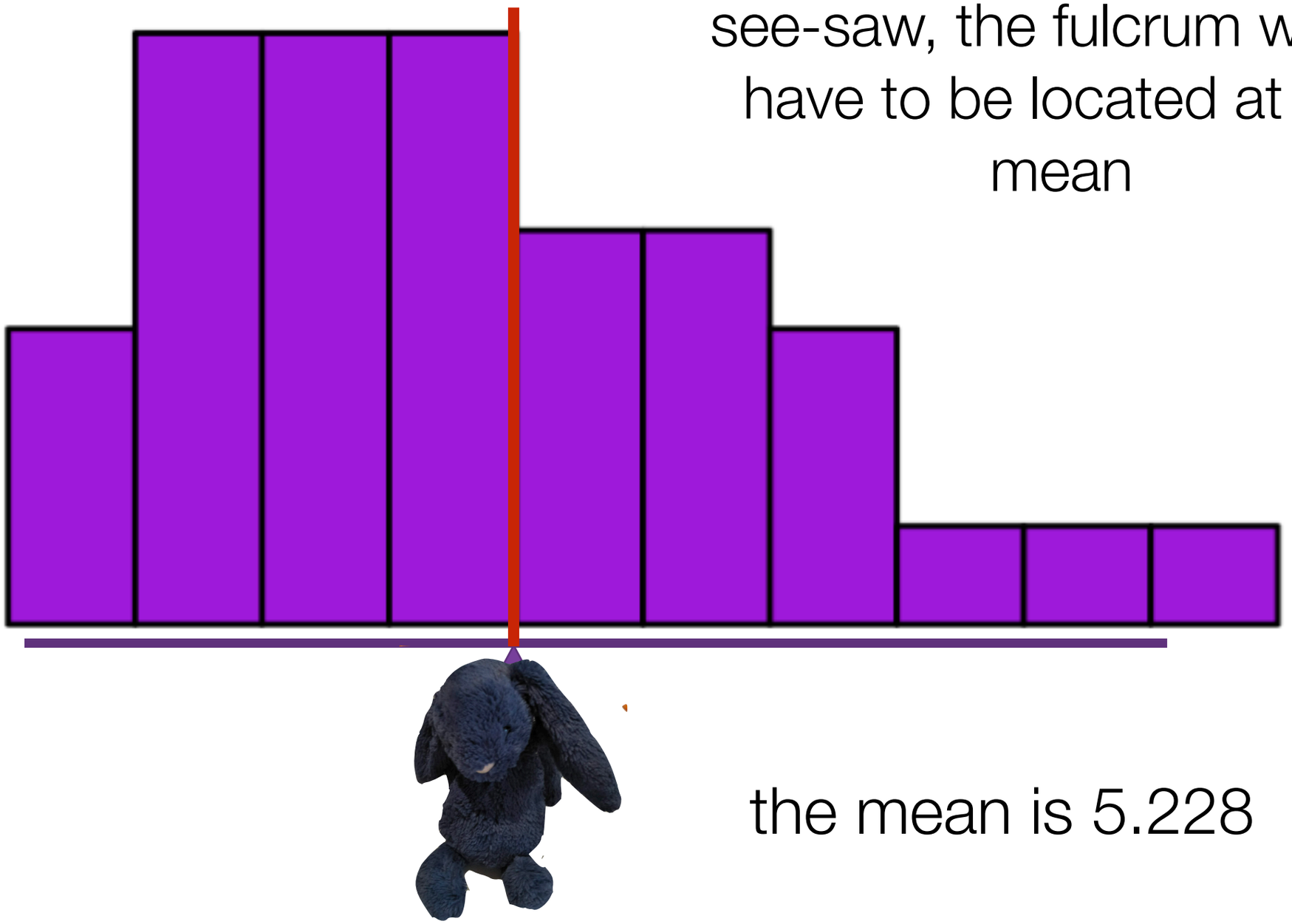


Where is the centre part of the dataset?



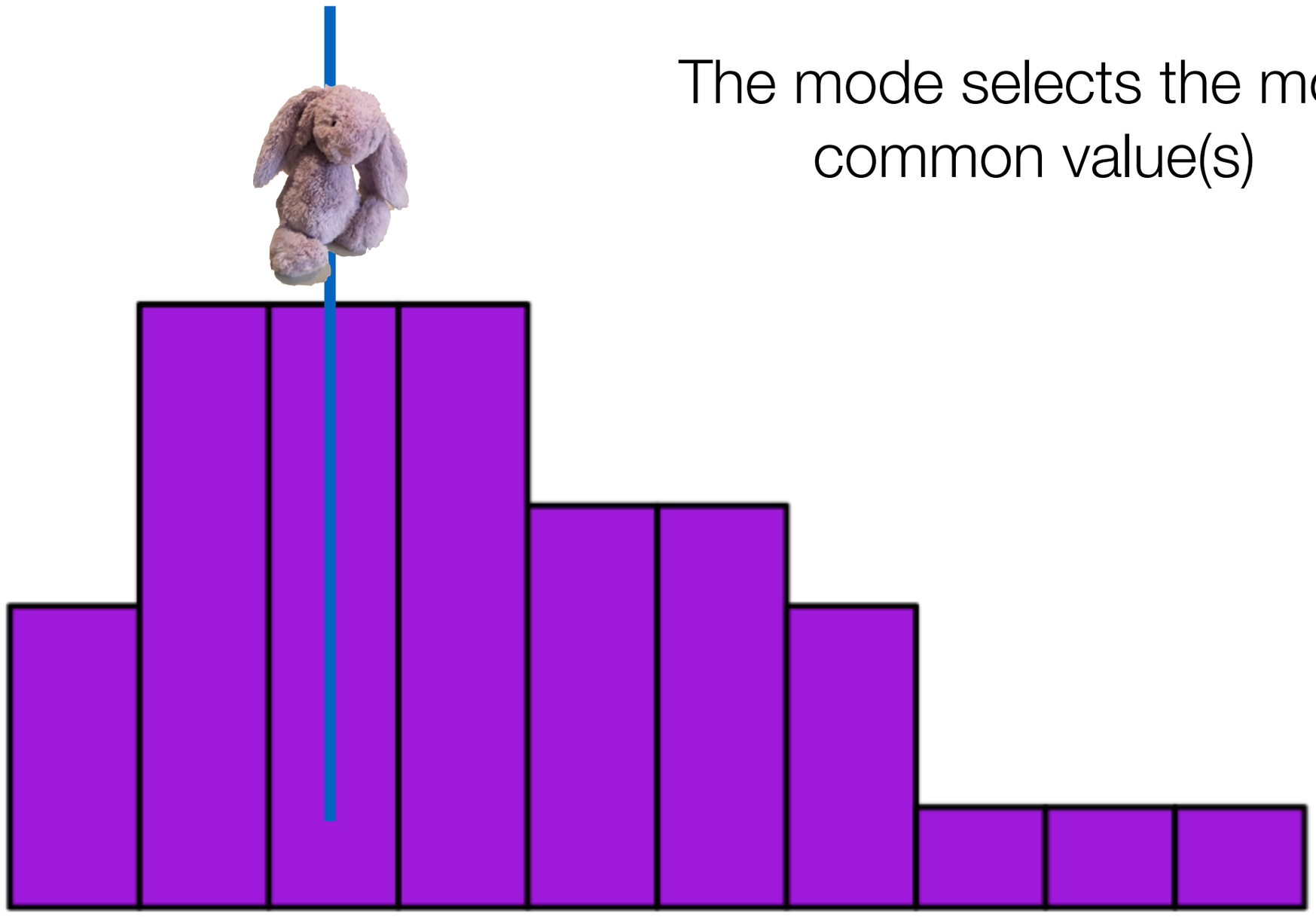
Idea #1: centre of mass (mean)

If the histogram were a solid object and you wanted to balance it on a see-saw, the fulcrum would have to be located at the mean



the mean is 5.228

Idea #2: most common point or points (mode)



The mode selects the most common value(s)

the modes are 3, 4, and 5

Idea #3: halfway point (median)

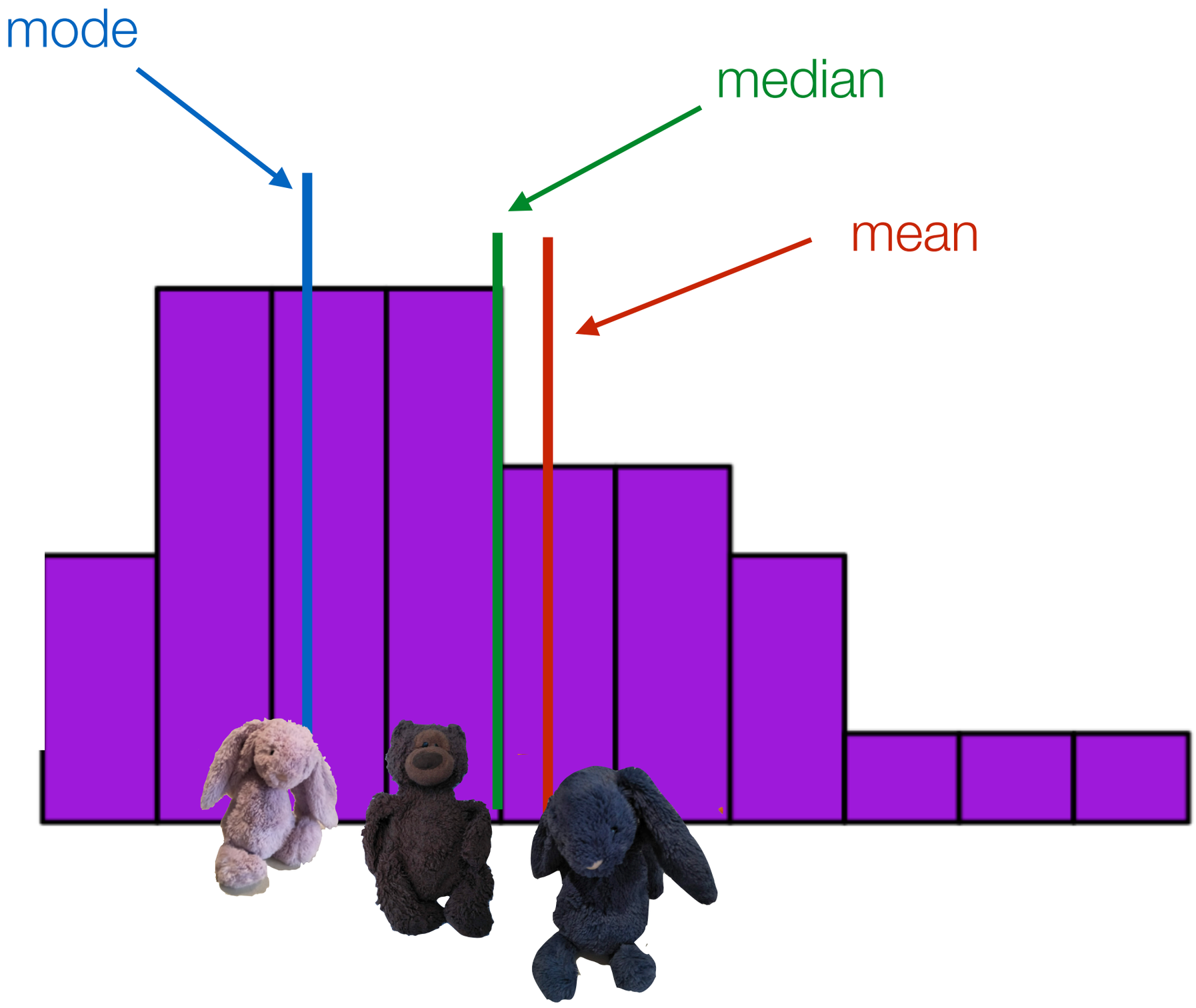
The median splits it so half your data is on one side and half is on another

50% of the data is below the median… the median is the "50th **percentile**" (a.k.a. "50th **quantile**")

```
> sort(gdata$age)
 [1]  1  2  2  3  3  3  3  3
      3  4  4  4  4  4  4  5
      5  5  5  5  5  5  6  6
      6  6  6  6  7  7  7  7
      8  8  8  9 10 11
```

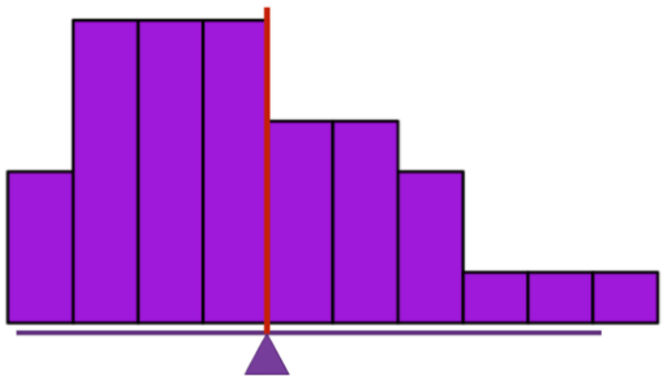
the median is 5

For many datasets, they aren't the same



So… how do we calculate these?

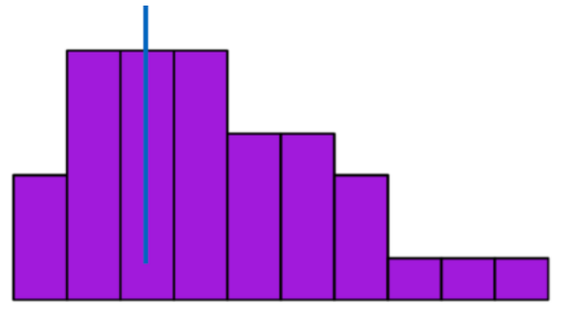
mean



median

1 2 2 3 3 3 3 3
3 4 4 4 4 4 4 5
5 5 5 5 5 5 6 6
6 6 6 6 7 7 7 7
8 8 8 9 10 11

mode



Notation. Yay!

the observation	its symbol	the observed value
Friend 1 of Gladly	X_1	8
Friend 2 of Gladly	X_2	6
Friend 3 of Gladly	X_3	3
Friend 4 of Gladly	X_4	5
Friend 5 of Gladly	X_5	7

In general… X_i is the i-th observation

The **mean** is just an average

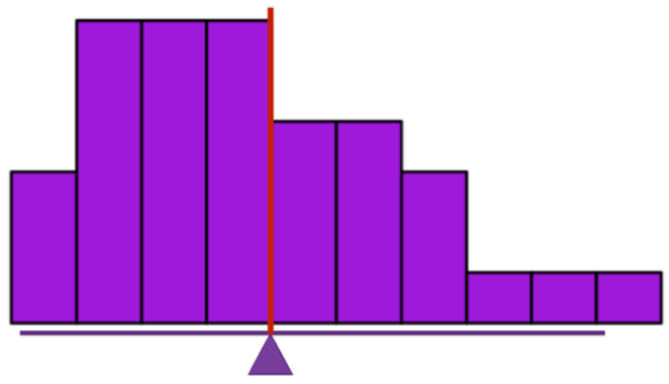
- Basic idea: add up each of the data points, and divide by the number of data points total (*N*)

Calculating the mean involves adding up all the observed values and dividing by the number of such values

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_{N-1} + X_N}{N}$$

The mean of the X values is generally written as “X-bar”

The number of observations is called the “sample size”



Some more notation

These are the same thing

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_{N-1} + X_N}{N}$$

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

This is the summation sign. It means “add up all the X_i values”

How do we do this in R?

```
> mean(gdata$age)
[1] 5.228571
```

```
> sum(gdata$age)/length(gdata$age)
[1] 5.228571
```

What if some data is missing?

```
# remove the 20th item
gdata$age[20] <- NA

# calculate the mean of the new vector
mean(gdata$age)
[1] NA
```

Oh noes!!
What can we do?!!

What if some data is missing?

```
# remove the 20th item
gdata$age[20] <- NA

# calculate the mean of the new vector
mean(gdata$age,na.rm=TRUE)
[1] 5.264706
```

It works!

Tell it to remove the NA values when calculating mean

```
# now let's put the data back
gdata$age[20] <- 4
```

The **median**

1 2 2 3 3 3 3 3
3 4 4 4 4 4 4 5
5 5 5 5 5 5 6 6
6 6 6 6 7 7 7 7
8 8 8 9 10 11

- The median is the “middle observation”
  - Half of the observations are bigger than the median
  - Half of the observations are smaller than the median
  - e.g., with 5 observations, the median is the 3rd largest
  - e.g., with 8 observations, the median is the average of the 4th and 5th

8, 31, 32, 56, 56

median is 32

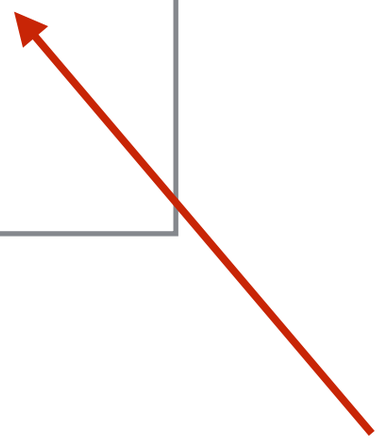
3, 8, 15, 31, 32, 41, 56, 56

median is 31.5

Calculating medians in R

```
> median(gdata$age)
[1] 5
```

```
> quantile(gdata$age,0.5)
  50%
[1] 5
```



50% quantile is the median

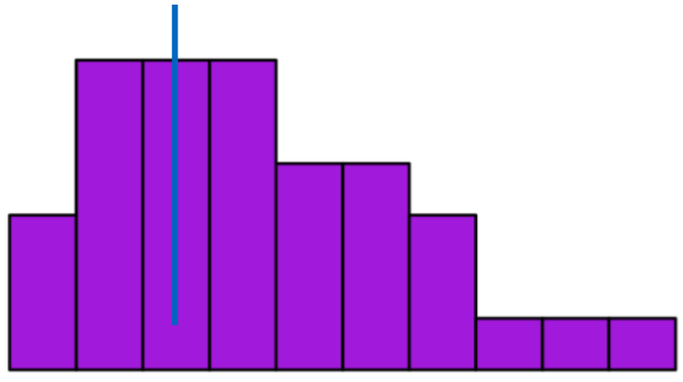
The **mode**

- The mode is the most frequently observed value(s) in the data
- It's quite often used in connection with nominal scale data
- The library lsr contains the functions we need

```
> modeOf(gdata$age)
[1] 3 5 4

> maxFreq(gdata$age)
[1] 6
```

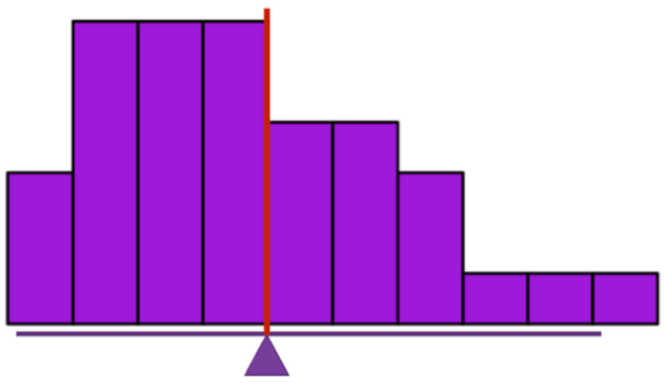
the most common ages are 3, 4, and 5

There are six people who are three, four, or five years old



Measures of central tendency

mean



$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_{N-1} + X_N}{N}$$

`mean(data)`

or

`sum(data)/N`

median

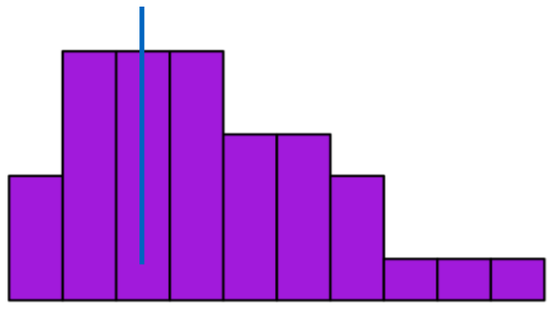
1 2 2 3 3 3 3 3
3 4 4 4 4 4 4 5
5 5 **5** 5 5 5 6 6
6 6 6 6 7 7 7 7
8 8 8 9 10 11

`median(data)`

or

`quantile(data,0.5)`

mode



```
library(lsr)
modeOf(data)
maxFreq(data)
```

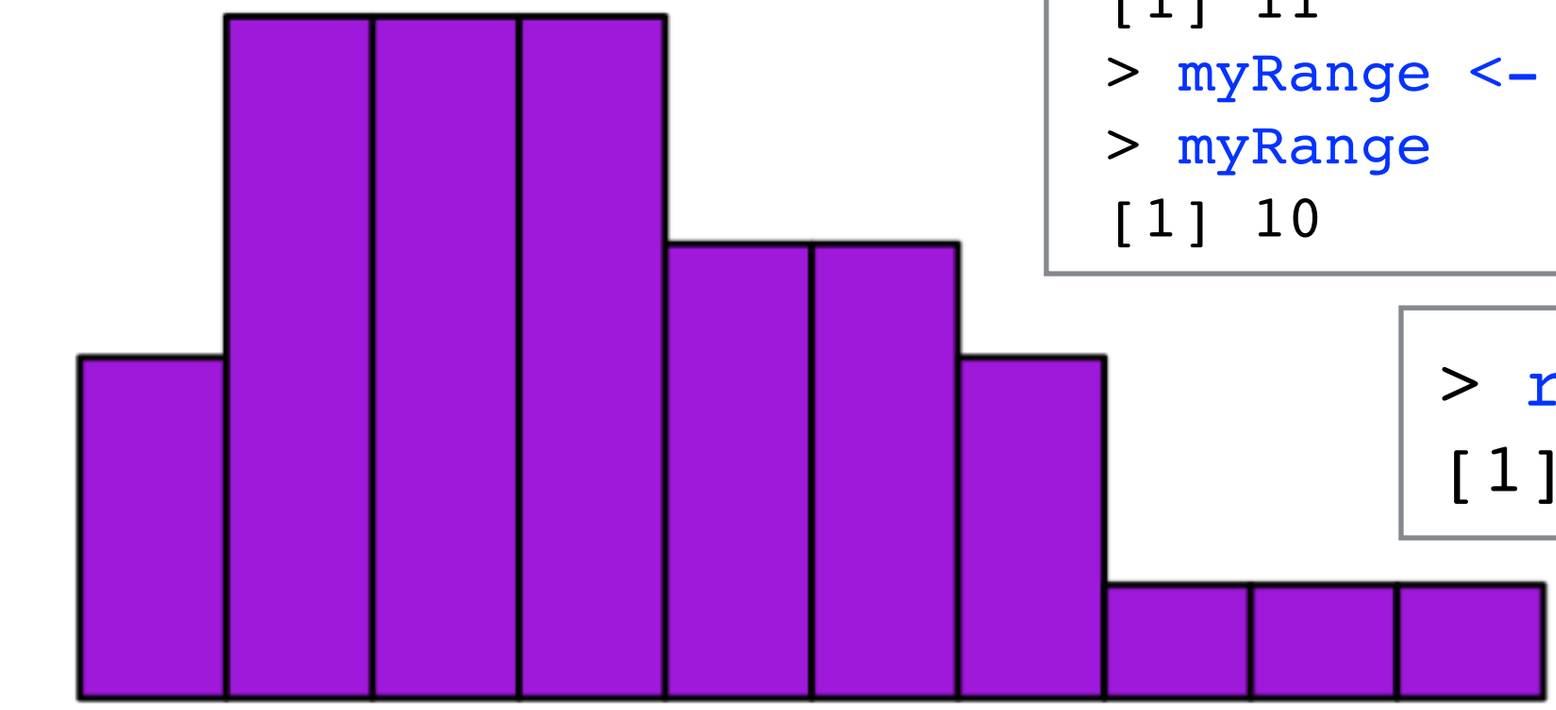
Measures of spread

range

interquartile range

standard deviation

Idea #1: the two ends (range)



range = max - min

min

max

```
> minimum <- min(gdata$age)
> minimum
[1] 1
> maximum <- max(gdata$age)
> maximum
[1] 11
> myRange <- maximum - minimum
> myRange
[1] 10
```

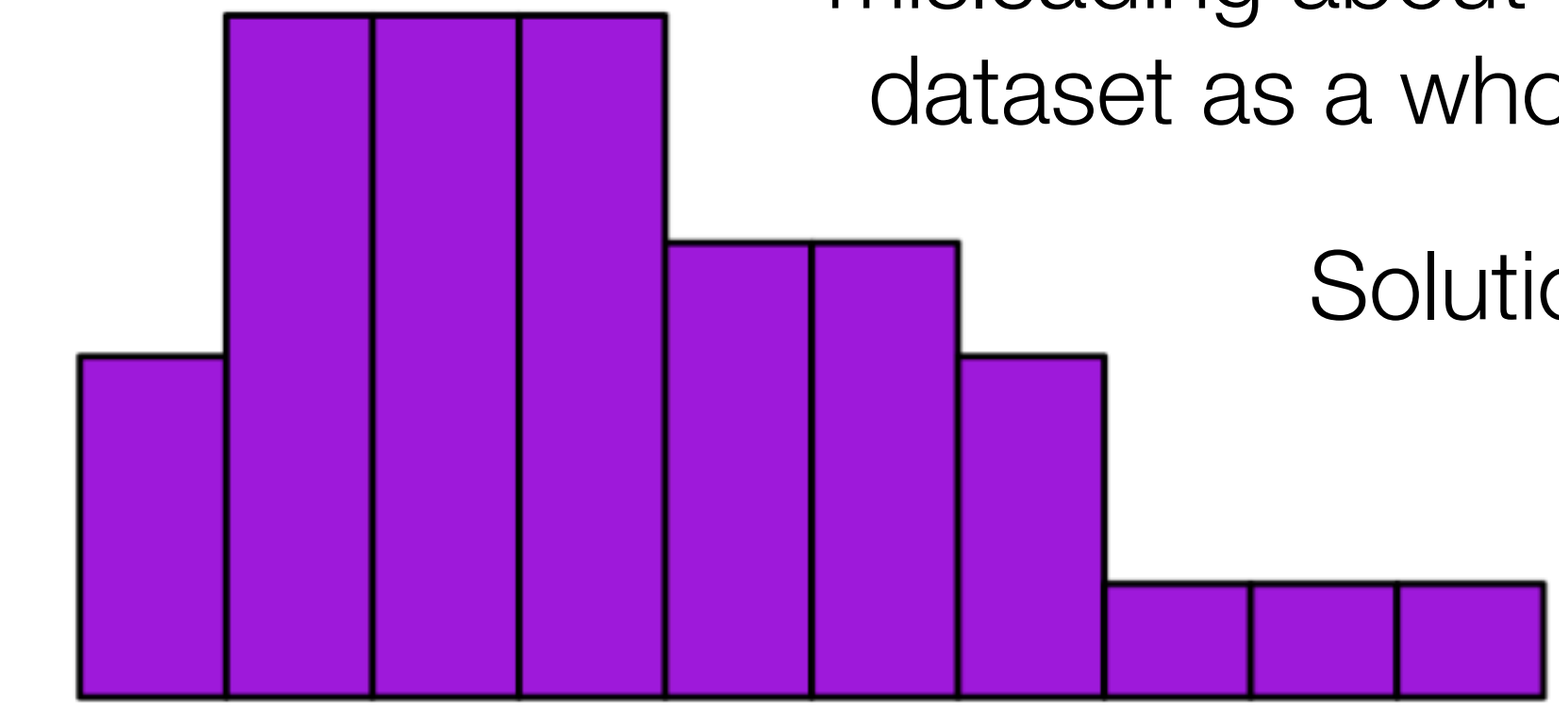
```
> range(gdata$age)
[1]   1 11
```

Idea #1: the two ends (range)

A problem - very sensitive to *outliers*

Range is now fairly misleading about the dataset as a whole

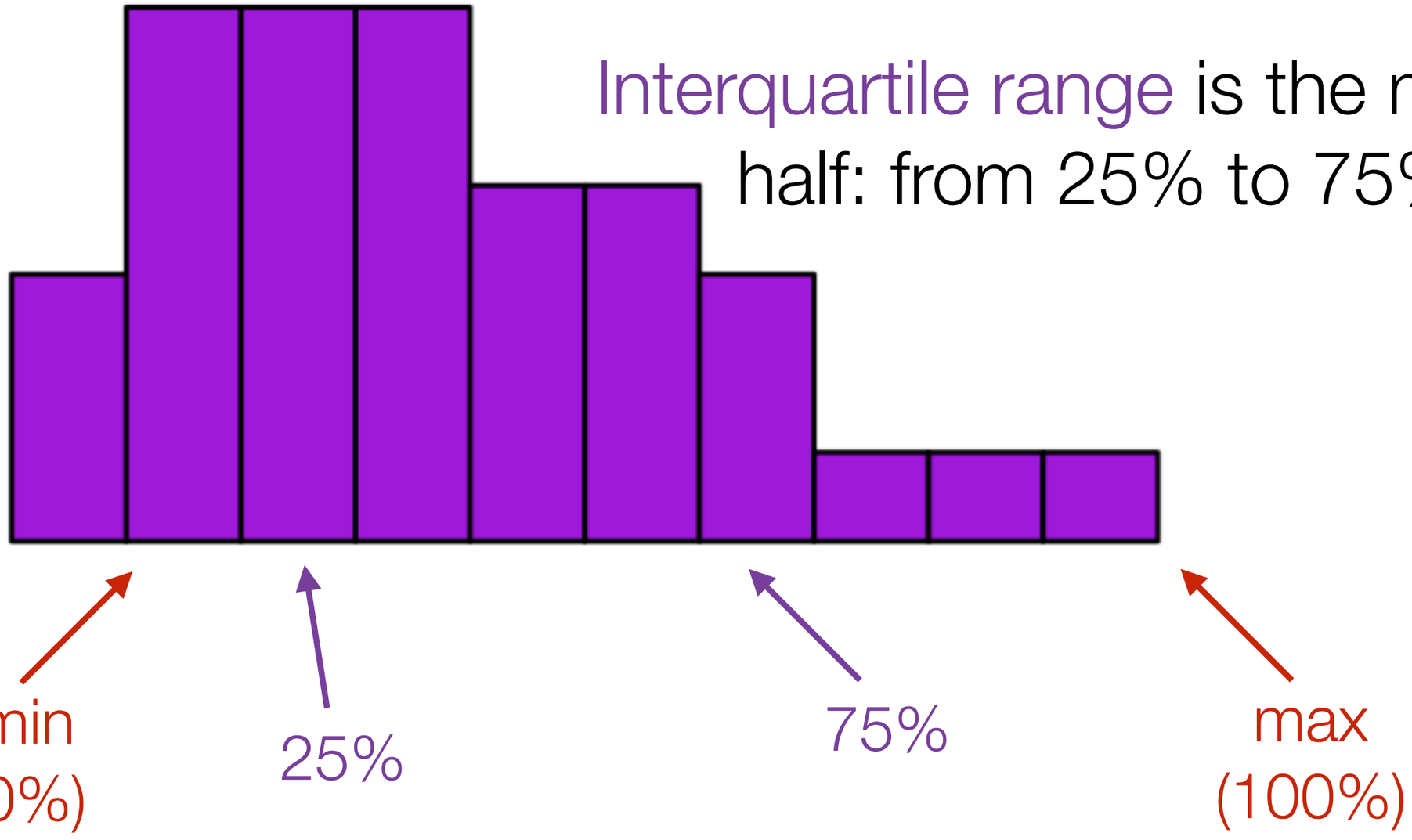
Solution: **interquartile range**



Idea #2: the middle half (interquartile range)

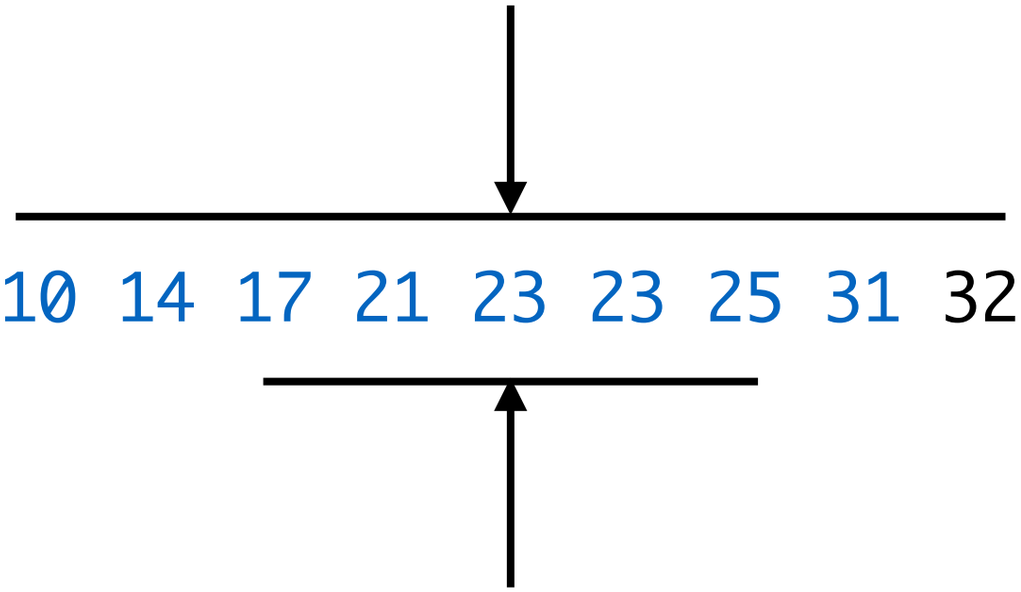
Range is from 0% (minimum) to 100% (maximum)

Interquartile range is the middle half: from 25% to 75%



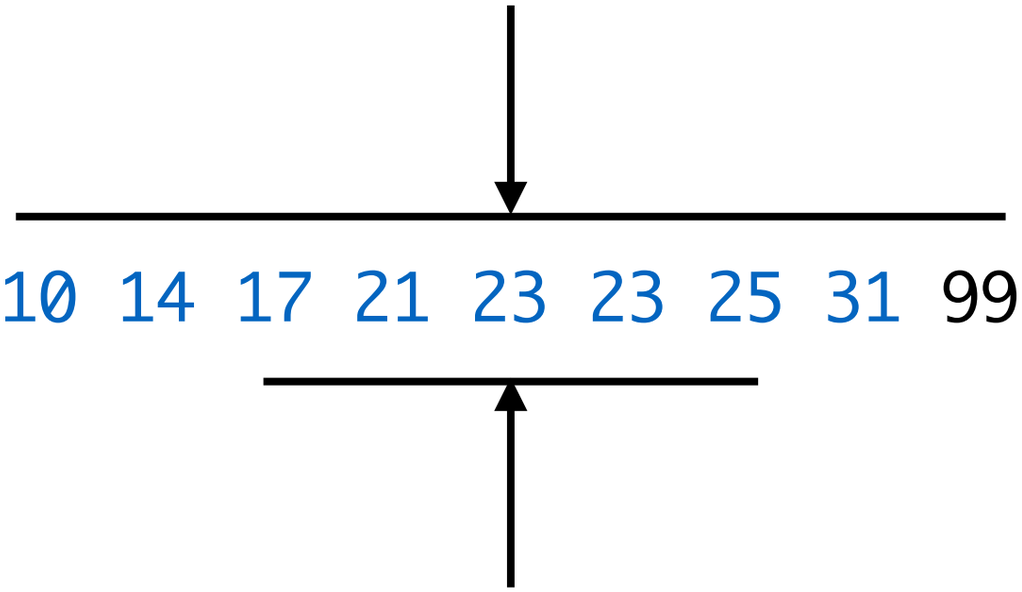
Idea #2: the middle half (interquartile range)

The range is 22



The the interquartile range is 8

The range is 89



The the interquartile range is 8

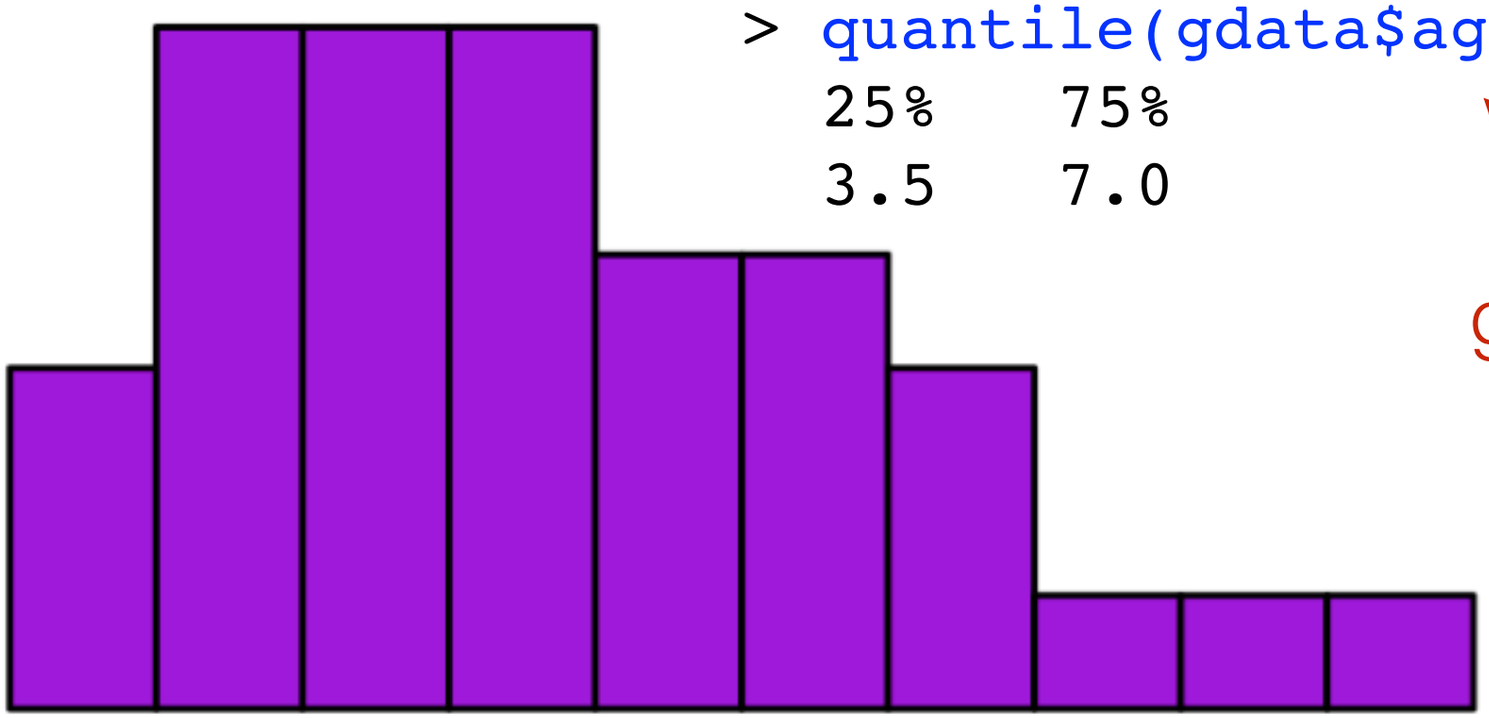
Idea #2: the middle half (interquartile range)

```
> IQR(gdata$age)
[1] 3.5
```

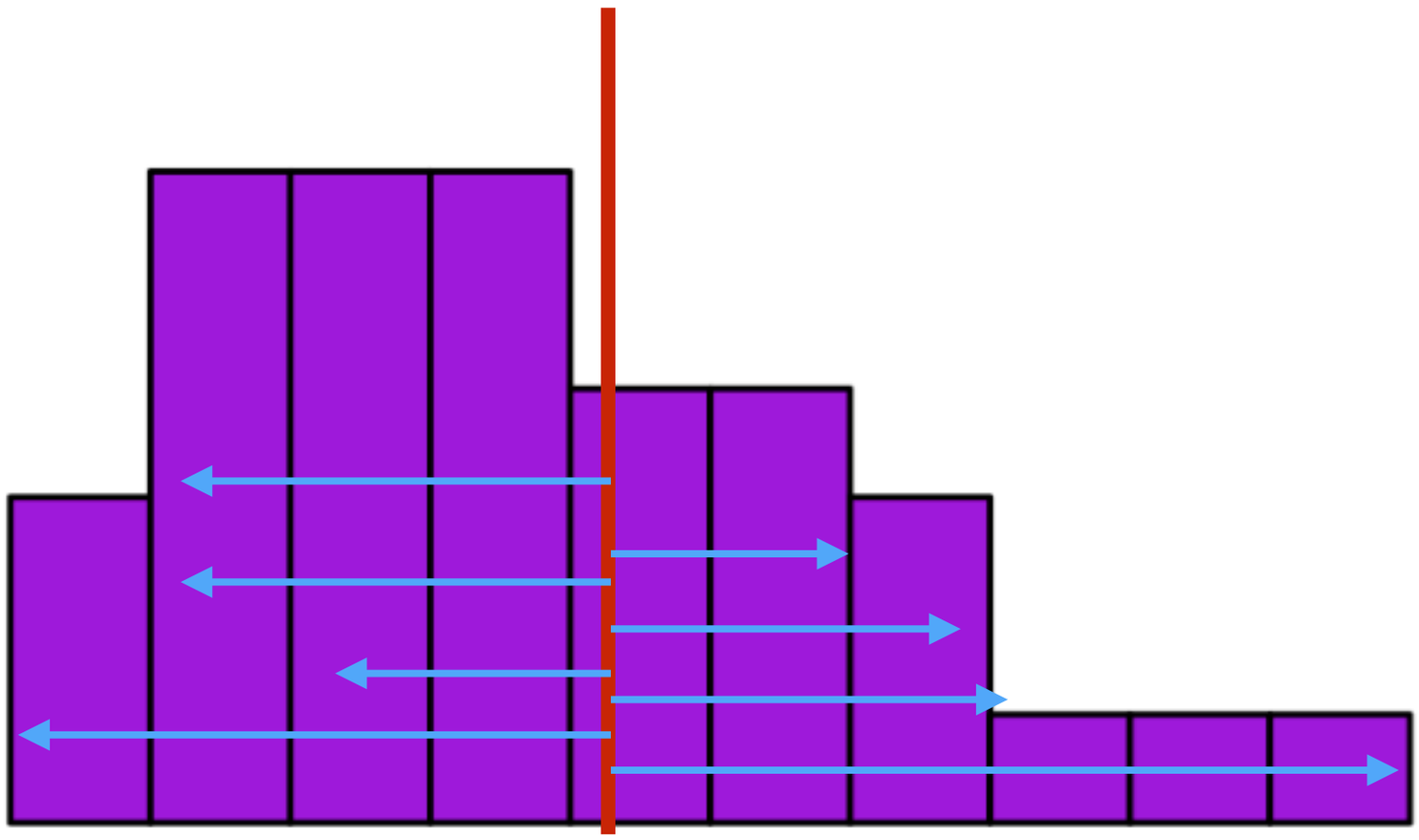
gives the 75th-25th

```
> quantile(gdata$age,c(0.25,0.75))
  25%   75%
  3.5   7.0
```

gives the 25% and 75% edges

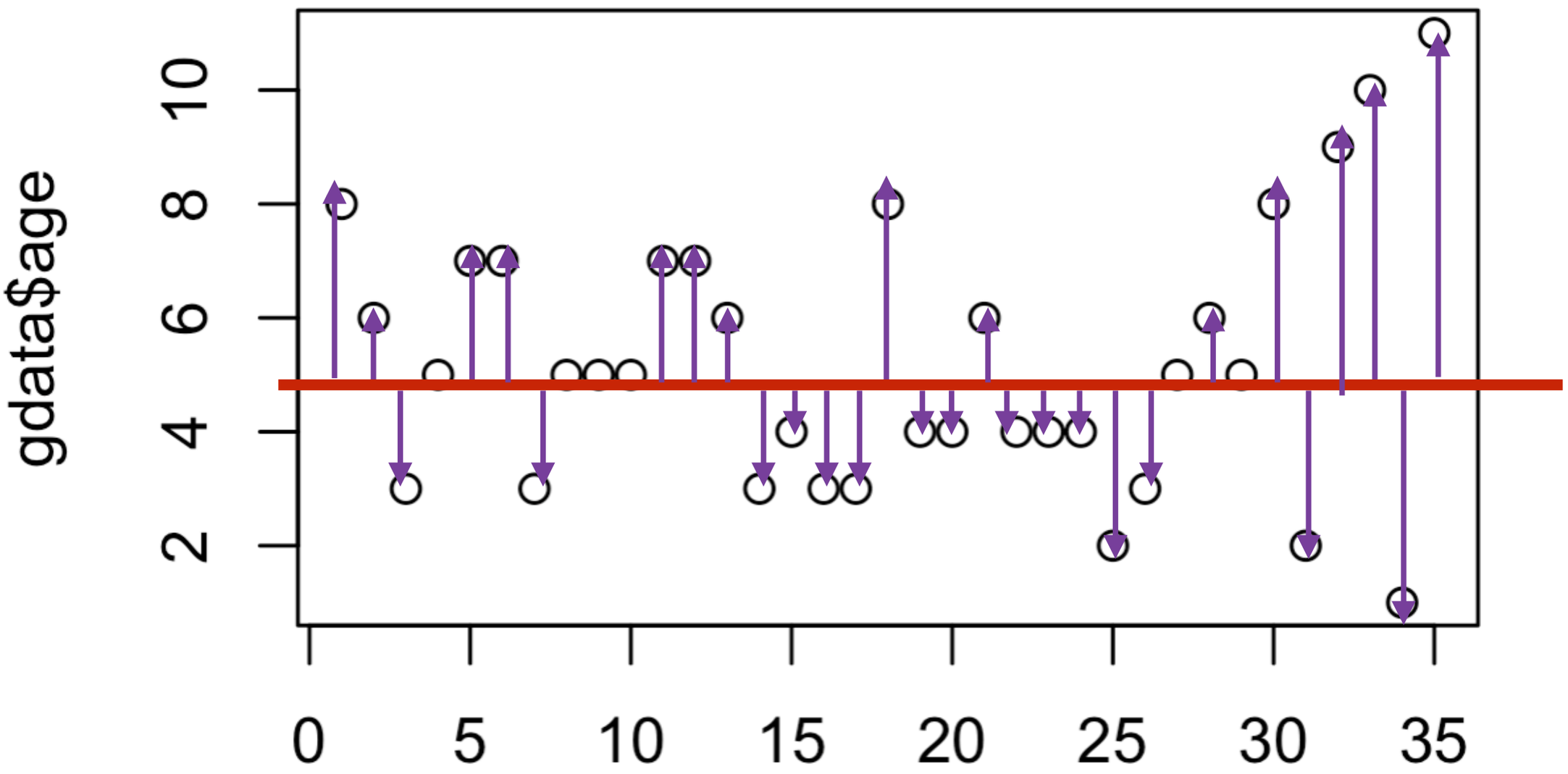


Idea #3: how far each point is from the centre of mass (standard deviation)



Idea #3: how far each point is from the centre of mass (standard deviation)

```
> plot(gdata$age)
```



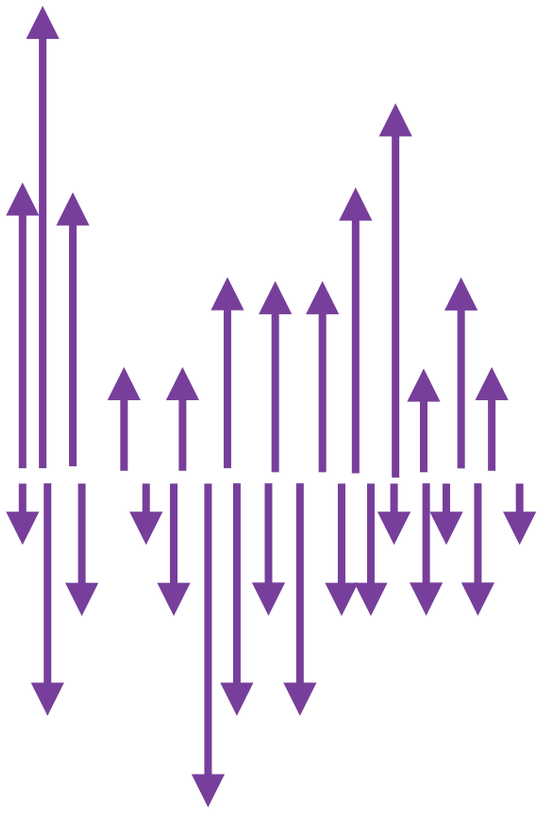
Idea #3: how far each point is from the centre of mass (standard deviation)

```
> plot(gdata$age)
```

Might be sensible to add these up and take the average

Idea #3: how far each point is from the centre of mass (standard deviation)

```
> plot(gdata$age)
```



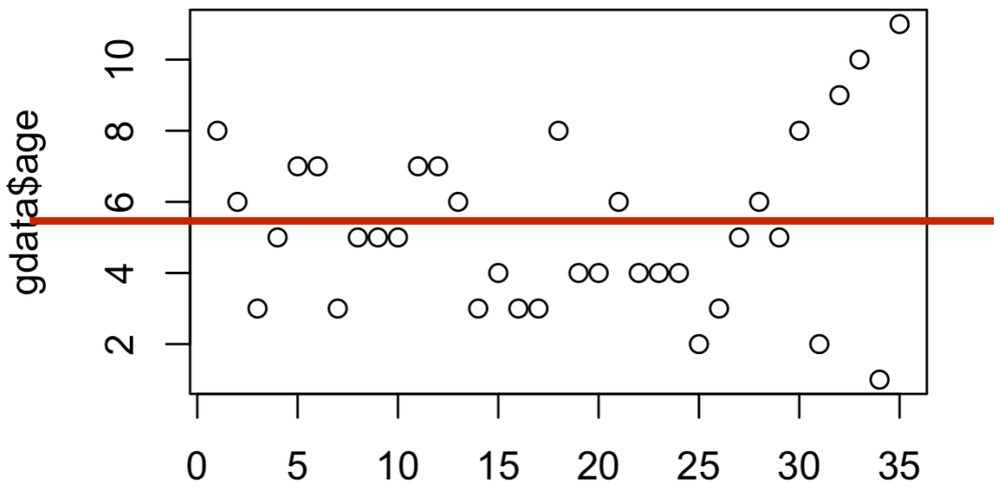
Problem: the positive and negative cancel each other out!

To deal with this issue we do a little trick with squaring

Might be sensible to add these up and take the average

Idea #3: how far each point is from the centre of mass (standard deviation)

Let's see it in terms of the equation



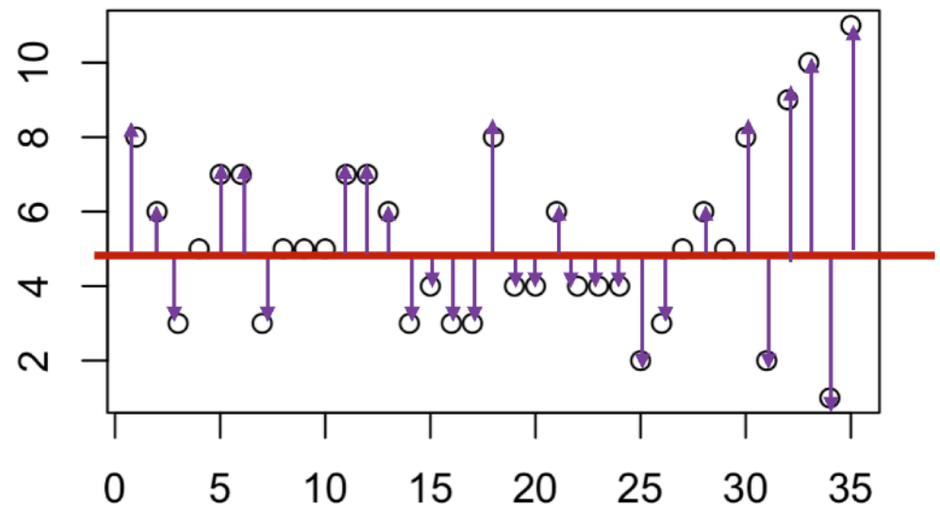
This is the **mean** of the sample

$$\bar{X}$$

NOTE: you do *not* need to have memorised this equation for the exam. The important thing is to understand what it's doing.

Idea #3: how far each point is from the centre of mass (standard deviation)

Let's see it in terms of the equation



For each of the i items, I **calculate the distance between it and the mean**

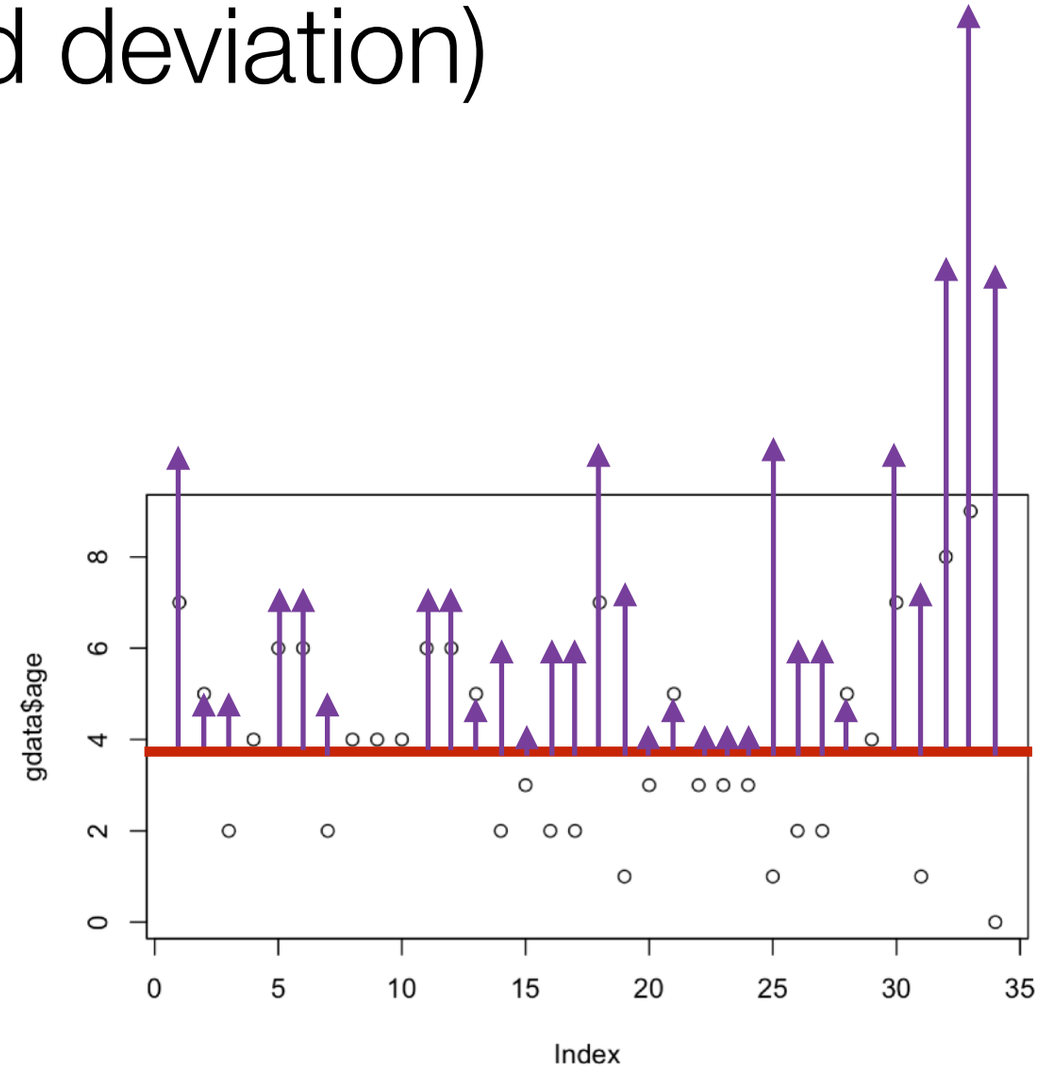
$$X_i - \bar{X}$$

NOTE: you do *not* need to have memorised this equation for the exam. The important thing is to understand what it's doing.

Idea #3: how far each point is from the centre of mass (standard deviation)

Let's see it in terms of the equation

Squaring them removes the problem of negatives, since squaring things makes them positive… but it also makes them bigger



$$(X_i - \bar{X})^2$$

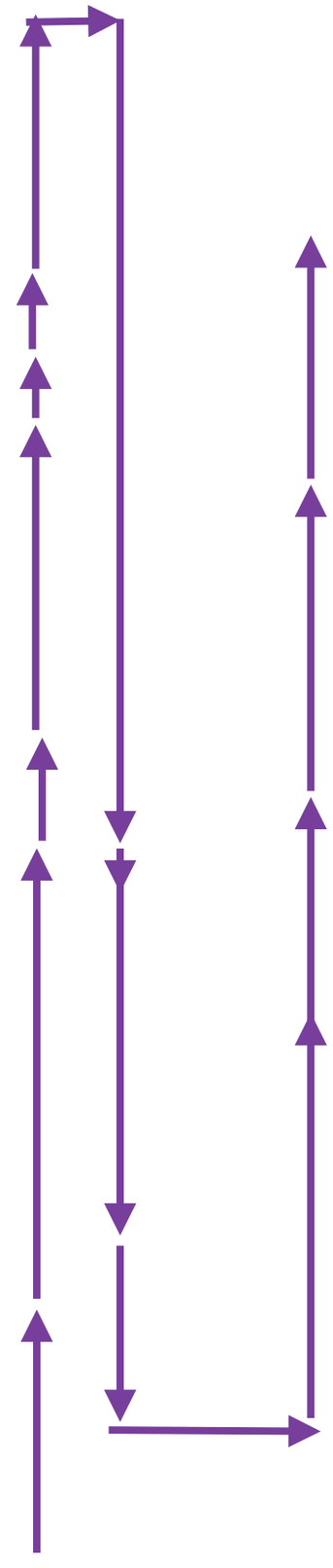
NOTE: you do *not* need to have memorised this equation for the exam. The important thing is to understand what it's doing.

Idea #3: how far each point is from the centre of mass (standard deviation)

Let's see it in terms of the equation

Add all N of them up

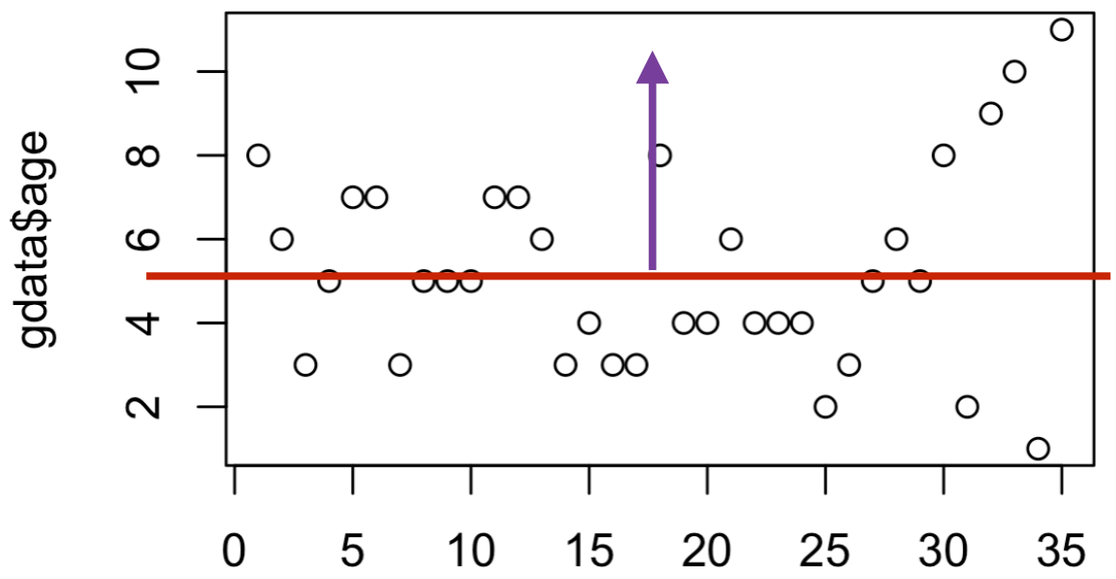
$$\sum_{i=1}^{N} (X_i - \bar{X})^2$$



NOTE: you do *not* need to have memorised this equation for the exam. The important thing is to understand what it's doing.

Idea #3: how far each point is from the centre of mass (standard deviation)

Let's see it in terms of the equation



Dividing by the number of data points gives the average distance

$$\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2$$

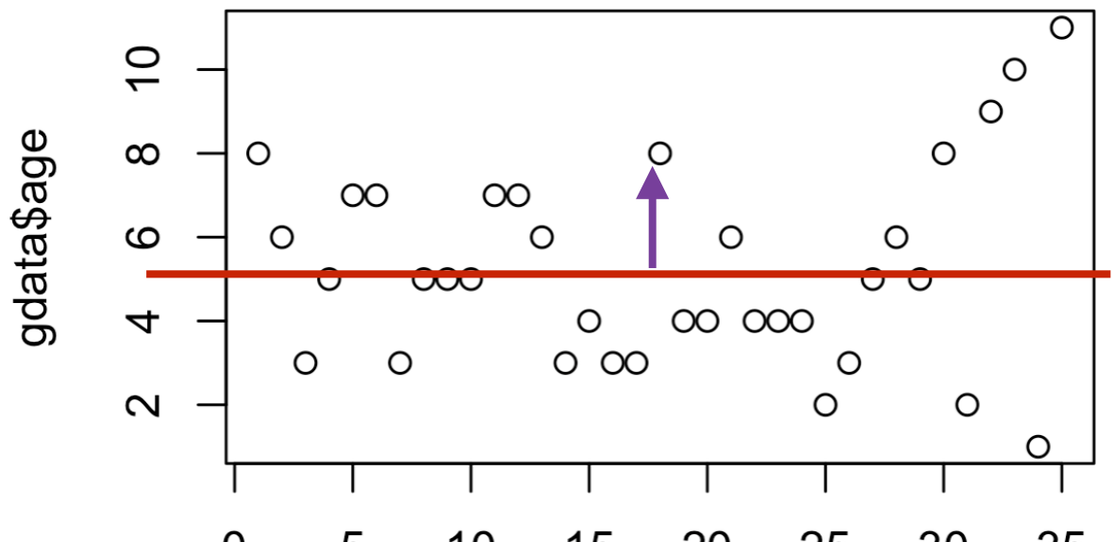
NOTE: you do *not* need to have memorised this equation for the exam. The important thing is to understand what it's doing.

Idea #3: how far each point is from the centre of mass (standard deviation)

Let's see it in terms of the equation

But then remember we made them all bigger by squaring, so we have to reverse that by taking the square root!

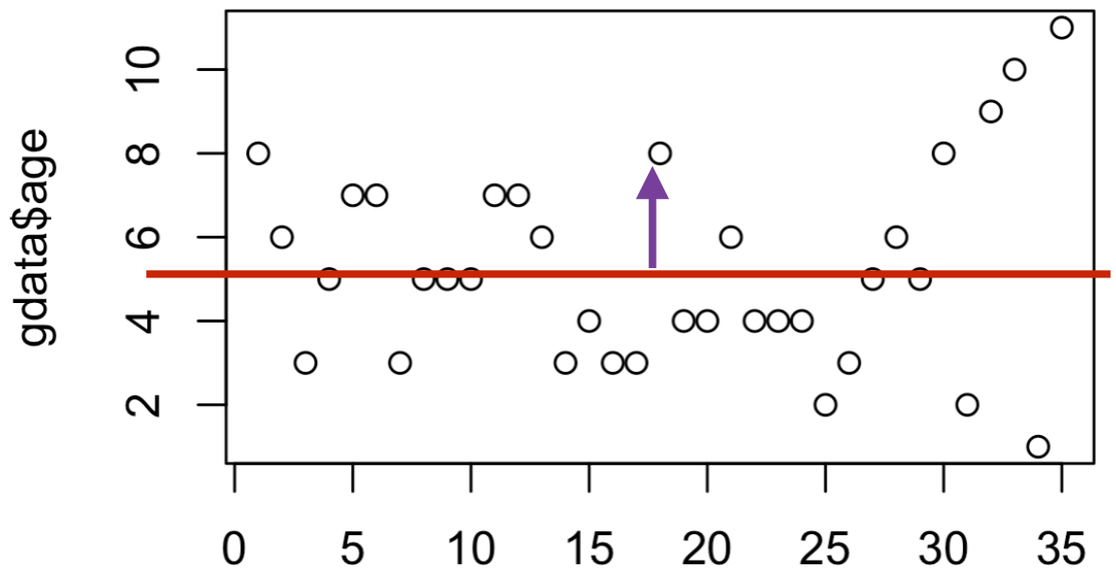
$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}$$



NOTE: you do *not* need to have memorised this equation for the exam. The important thing is to understand what it's doing.

Idea #3: how far each point is from the centre of mass (standard deviation)

Let's see it in terms of the equation



This is the standard deviation!

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

Tedious detail: In practice, the "average" is calculated by dividing by N-1 and **not** by N. I'll explain why in the theory lectures

NOTE: you do *not* need to have memorised this equation for the exam. The important thing is to understand what it's doing.

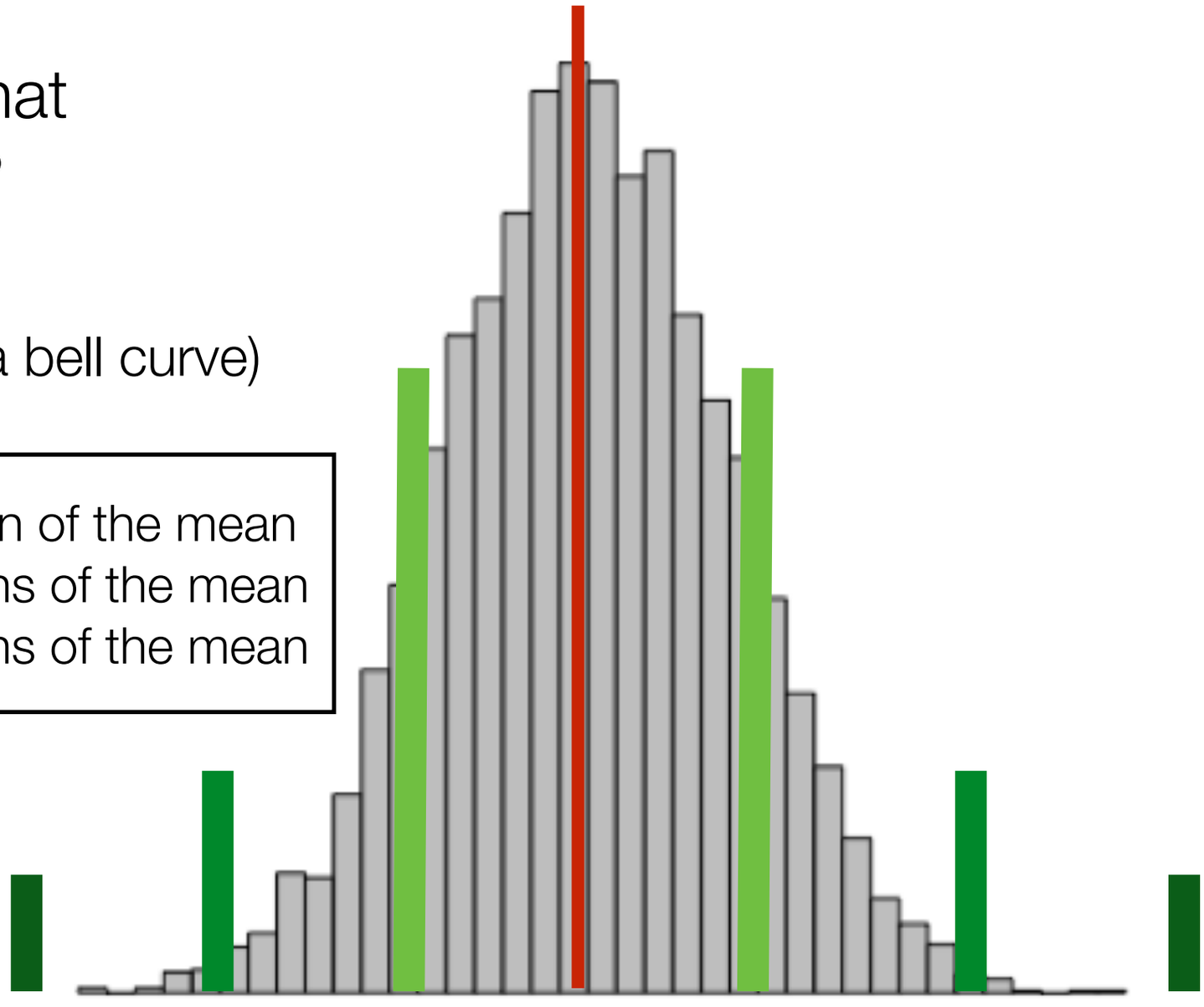
Idea #3: how far each point is from the centre of mass (standard deviation)

So that's the equation… but what does it mean?!?

If your data are "normal" (i.e., like a bell curve)

68% is within 1 standard deviation of the mean
95% is within 2 standard deviations of the mean
99% is within 3 standard deviations of the mean

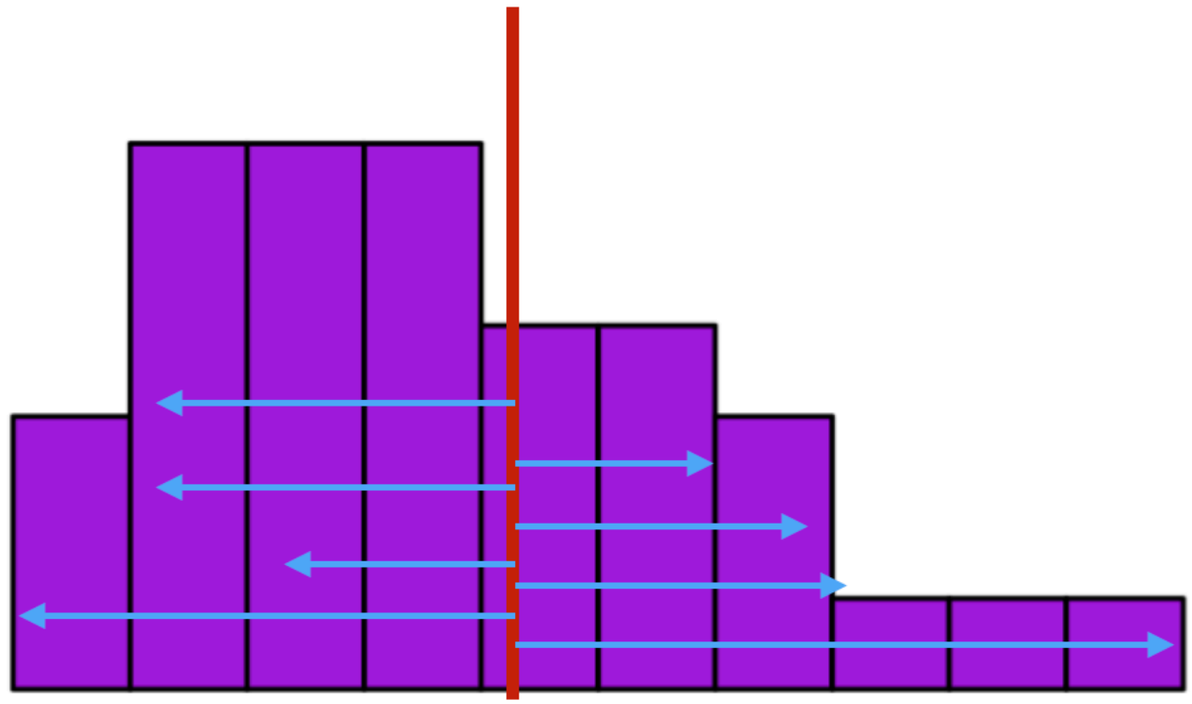
Even if they're not exactly normal, this is a reasonable guide (unless it is way off)



Idea #3: how far each point is from the centre of mass (standard deviation)

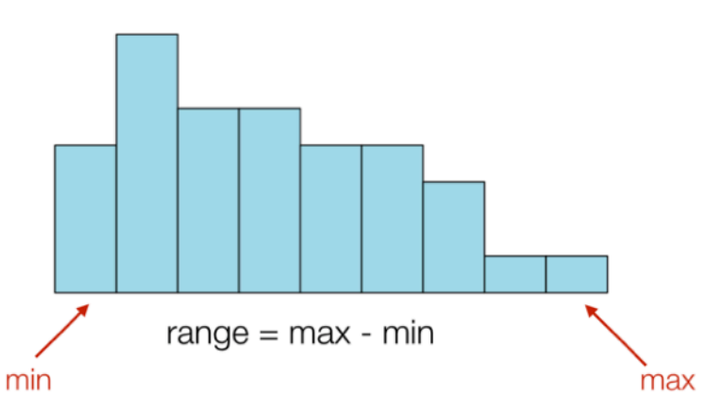
Doing it in R

```
> sd(gdata$age)
[1] 2.340006
```



Measures of spread

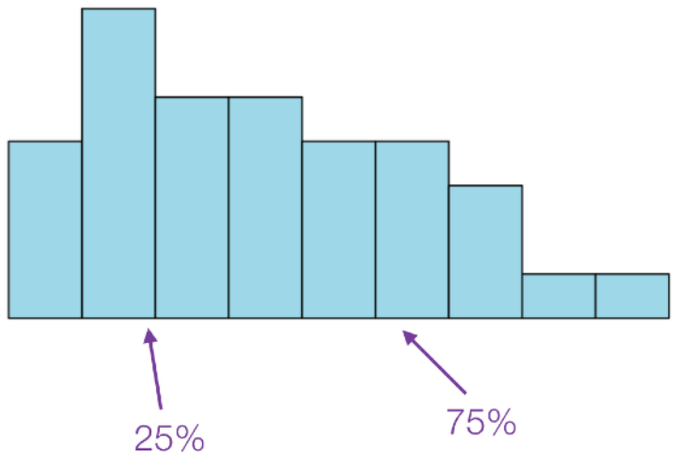
range



max - min

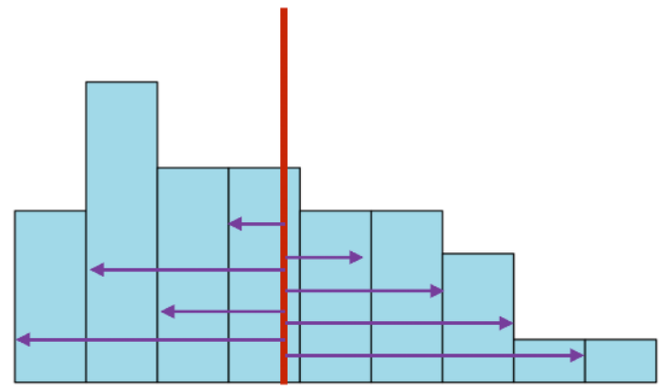
```
range(data)
```

interquartile range




```
IQR(data)
quantile(data,c(0.25,0.75))
```

standard deviation



```
sd(data)
```

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

Is there a faster way
to do this?!?

Viewing the dataset…

An easy way is to use the function `head()`, which shows the first few rows of the dataset

```{r showdatachunk}
head(gdata)
```

name <chr>	species <chr>	year <dbl>	food <chr>	carrot <dbl>	cake <dbl>	m… <dbl>	age <dbl>
bunny	bunny	2013	carrot	10	10	1	8
gladly	bear	2015	honey	7	10	2	6
flopsy	bunny	2018	lettuce	10	9	1	3
doggie	dog	2016	chicken	8	10	1	5
lfb	bunny	2014	lettuce	10	9	2	7
cuddly paws	bunny	2014	*NA*	*NA*	*NA*	*NA*	7

6 rows | 1-8 of 9 columns

Viewing the dataset…

I also like the `glimpse()` function, which summarises things in a different way…

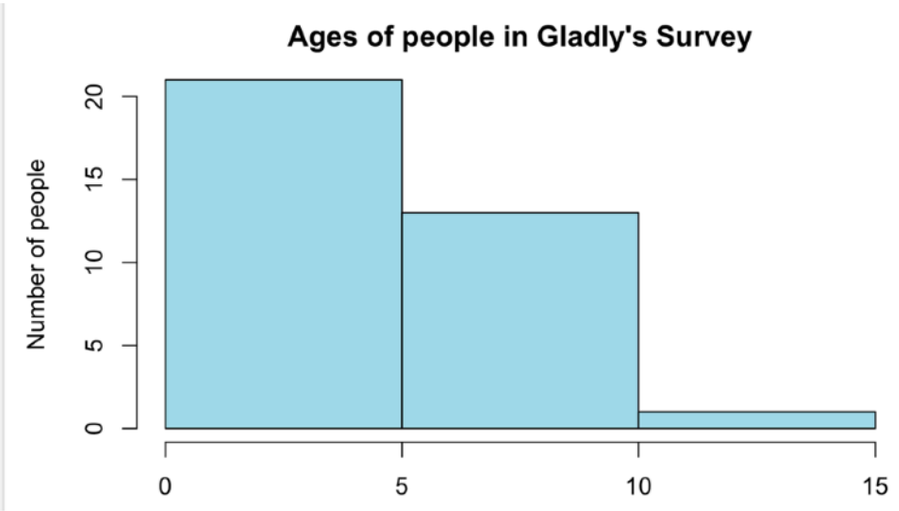
```
```{r showdatachunk}
glimpse(gdata)
```
```

```
Rows: 35
Columns: 9
$ name    <chr> "bunny", "gladly", "flopsy", "doggie", "lfb", "c…
$ species <chr> "bunny", "bear", "bunny", "dog", "bunny", "bunny…
$ year    <dbl> 2013, 2015, 2018, 2016, 2014, 2014, 2018, 2016, …
$ food    <chr> "carrot", "honey", "lettuce", "chicken", "lettuc…
$ carrot  <dbl> 10, 7, 10, 8, 10, NA, 9, 9, 10, 8, 6, 9, 7, 6, 7…
$ cake    <dbl> 10, 10, 9, 10, 9, NA, 10, 7, 8, 7, 10, 9, 8, 8, …
$ mud     <dbl> 1, 2, 1, 1, 2, NA, 2, 1, 1, 2, 1, 2, 2, 1, 1, 1,…
$ age     <dbl> 8, 6, 3, 5, 7, 7, 3, 5, 5, 5, 7, 7, 6, 3, 4, 3, …
$ gender  <chr> "female", "male", "nb", "male", "female", "femal…
```

But what if you want summary statistics?

The `summary()` function is nice: gives means, quantiles, etc

```
```{r showdatachunk}
summary(gdata)
```
```

```
     name              species              year          food          
 Length:35          Length:35          Min.   :2010   Length:35         
 Class :character   Class :character   1st Qu.:2014   Class :character  
 Mode  :character   Mode  :character   Median :2016   Mode  :character  
                                       Mean   :2016                     
                                       3rd Qu.:2018                     
                                       Max.   :2020                     

     carrot            cake             mud              age        
 Min.   : 5.000   Min.   : 6.000   Min.   :1.000   Min.   : 1.000  
 1st Qu.: 7.000   1st Qu.: 8.000   1st Qu.:1.000   1st Qu.: 3.500  
 Median : 8.000   Median : 9.000   Median :1.000   Median : 5.000  
 Mean   : 7.912   Mean   : 8.529   Mean   :1.471   Mean   : 5.229  
 3rd Qu.: 9.000   3rd Qu.: 9.750   3rd Qu.:2.000   3rd Qu.: 7.000  
 Max.   :10.000   Max.   :10.000   Max.   :3.000   Max.   :11.000  
 NA's   :1        NA's   :1        NA's   :1                       
    gender         
 Length:35          
 Class :character   
 Mode  :character   
```

Exercises

1. Change your histogram of ages to look like the one on the right. (Hint: you might need to google around to find names of colours to use).



2. Without using `summary()`, calculate the mean, standard deviation, and median of the variables `carrot`, `cake`, and `mud`. (Hint: you'll need to use the `na.rm` argument in your functions). Where you can, check your answers against the ones shown by `summary()`.

3. Calculate the 10th and 90th percentile for `age`.

4. How would you interpret the responses to the questions about eating mud, carrots, and cake?