① We learnt about sample variance (and sample standard deviation) on Wednesday. How to interpret sample variance? What is it used for?

Answer: While sample mean illustrates central tendency (where the centre is), sample variance shows you how spread out the data are.

For example, the numbers 1,2,3 have a sample mean of 2, and a sample standard deviation of 1 (work it out yourself to confirm).

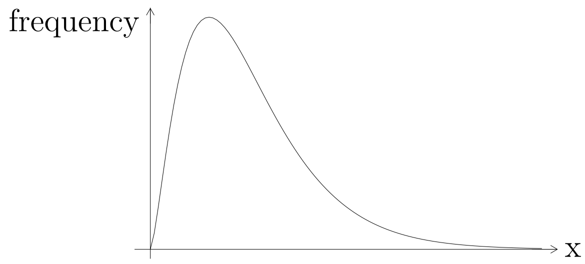
On the other hand, the numbers 0,2,4 also has a sample mean of 2, but the sample standard deviation is larger than 1 since the data are more "spread out" (the sample standard deviation is 2).

(Test: What do you think is the sample standard deviation of the numbers 7,7,7,7 ?)

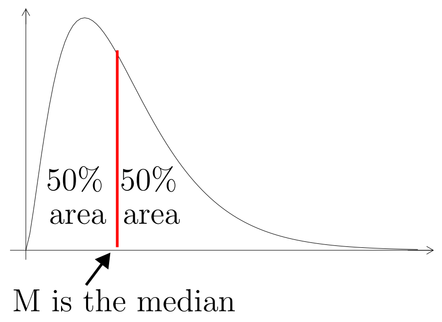
Answer: It's 0. There is no variation in the numbers.

② A student asked "how do we get the median for a skewed distribution?"

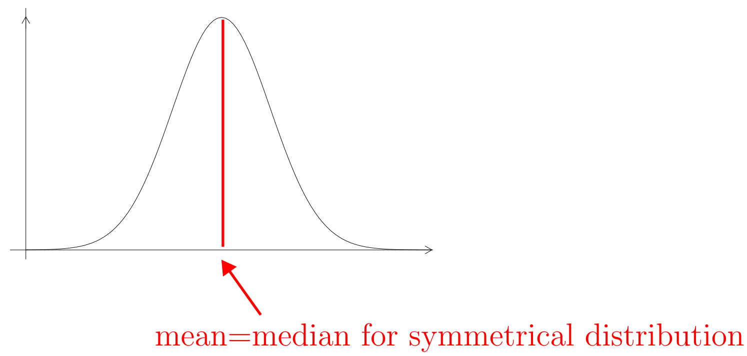
Example, distribution looks like this:



Answer: Locate the point M on the x axis such that half the area is on the right of M and half the area is on the left of M. That is



③ If a distribution is symmetrical, note that the mean and median coincide.



④ We learnt quartiles (Q1, Q3) last lecture. Note that there are several algorithms for calculation of quartiles (we learnt just one). R software uses a different algorithm/formula by default, so don't be surprised if R software gives you a different Q1,Q3 value from what you expect.

[For students who are interested, in R,

quantile(data, type=1) ← R will use method taught in class

quantile(data, type=7) ← R's default (different from our method)]

⑤ We learnt mean and median in the previous lecture. What is the mode of a distribution?

Ans: It is the observation that occurs the most number of times.

Example, 3, 4, 5, 5, 5, 6, 7, 9, 10, 11

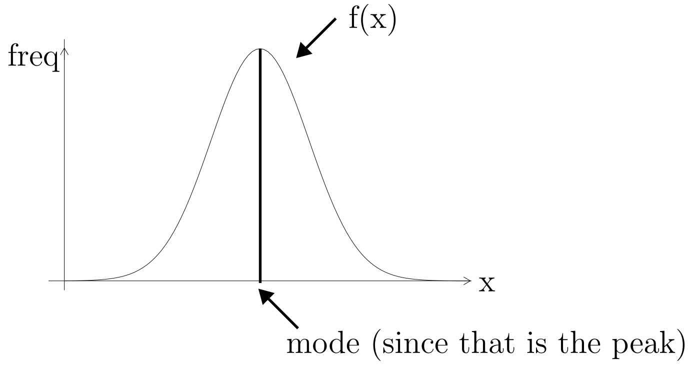
median=5.5

mean=$\frac{\sum_{i=1}^{10} x_i}{n} = \frac{65}{10} = 6.5$

mode=5 (appears most frequently)

More on mode.

Graphically, if the distribution is



Idea: If the above curve is f(x), how to find the mode?

Set $f'(x) = 0$ [that is, $\frac{dy}{dx} = 0$] and solve for x. [Recall in calculus: finding maximum point]

⑥ In previous lecture, it was mentioned that "median is resistant". What does that mean? Resistant to what?

Ans: Median is resistant to outliers. But what is an outlier?

This leads us to the beginning of today's lecture.

What are outliers

Outliers are observations that lie outside the overall pattern of a distribution.

That is quite vague, and statistical software often use the following rule to identify which observations are outliers.

The $1.5 \times$ IQR rule for outliers

Statistical software often use the $1.5 \times$ IQR rule. We call an observation a suspected outlier if it falls more than $1.5 \times$ IQR above Q3 (third quartile) or below Q1 (first quartile).

[Recall IQR=Q3-Q1]

Example: Here are 7 observations:

1 3 5 6 7 11 25

	Our method to compute Q1, Q3	R's method to compute Q1,Q3
Mean	8.286	8.286
Q1	3	4
Q3	11	9
IQR	11-3=8	9-4=5
$1.5\times$ IQR	$1.5 \times 8 = 12$	$1.5 \times 5 = 7.5$

Any values below Q1-($1.5\times$ IQR) is an outlier.

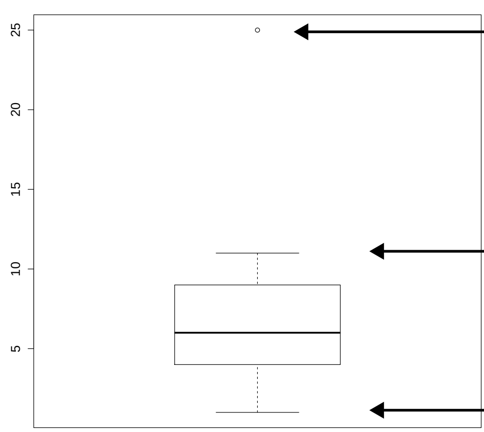
Similarly, any value above Q3+($1.5\times$ IQR) is an outlier.

In this example, the observation 25 is above Q3+($1.5\times$ IQR), so it is an outlier.

We use R to find some statistics of the data and to draw a **modified boxplot** (with outliers).

```
> data <- c(1,3,5,6,7,11,25)

>summary(data)

   Min.  1st Qu.  Median  Mean  3rd Qu.  Max.

   1.000  4.000  6.000  8.286  9.000  25.000

> boxplot(data)
```

When we have outliers, most software will produce a modified boxplot:



shows us this is an outlier

there is an outlier above Q3+1.5IQR, hence this does not represent the maximum value. Rather, it represents the maximum observation that is not an outlier (11).

there is no outlier below Q1-1.5IQR, so this represents the minimum value as before (min=1)

Remind students to register their iClickers before start of class next week. We are using iClickers starting next Wednesday.