Exercise 1.14: The accompanying data set consists of observations on shower-flow rate for a sample of $n = 129$ houses in Perth, Australia.

a. Construct a stem-and-leaf display of the data.
Solution:

```
> stem(x)

The decimal point is at the |

 2 | 23
 3 | 2344567789
 4 | 01356889
 5 | 00001114455666789
 6 | 0000122223344456667789999
 7 | 00012233455555668
 8 | 02233448
 9 | 012233335666788
10 | 2344455688
11 | 2335999
12 | 37
13 | 8
14 | 36
15 | 0035
16 |
17 |
18 | 9
```

b. What is a typical, or representative flow rate?

Solution:

```
   > fivenum(x)
Minimum   Lower Quartile   Median   Upper Quartile   Maximum
    2.2            5.6         7.0            9.6        18.9

   > mean(x)
     [1] 7.707752

   > sd(x)
     [1] 3.076844
```

I would say that a representative flow rate would be closer to the median at 7.0 since there is an outlier in 18.9 that is more than 3 standard deviations away from the mean.

c. Does the display appear to be highly concentrated or spread out?

Solution: Generally speaking this display appears to be highly concentrated with a small slight skew towards the right.

d Does the distribution of values appear to be reasonably symmetric?If not, how would you describe the departure from symmetry?

Solution: The data is does not appear to be symmetrical, since it seems to have a slight positive skew.

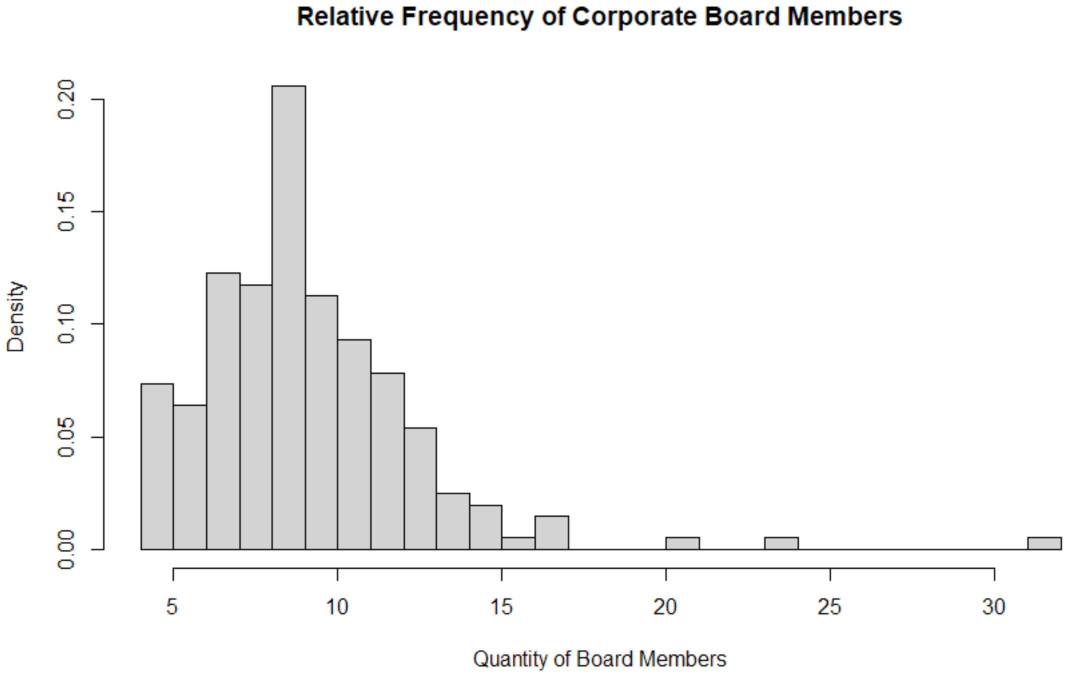
e Would you describe any observation as being far from the rest of the data (an outlier)?

Solution: I would describe the data point $x = 18.9$ as an outlier, since it is almost 4 standard deviations from the mean.

Exercise 1.18: Every corporation has a governing board og directors. The number of individuals on a board varies from one corporation to another. One of the authors of the article provided the accompanying data on the number of directors on each board in a random sample of 204 corporations.

a. Construct a histogram of the data based on relative frequencies and comment on any interesting features?

Solution:



The distribution appears to be roughly unimodal and right skewed. The data set contains outliers at (21,24,32) as there are gaps in the histogram. Spread appears to be between 4 to 17 with a majority of corporations having 9 members.

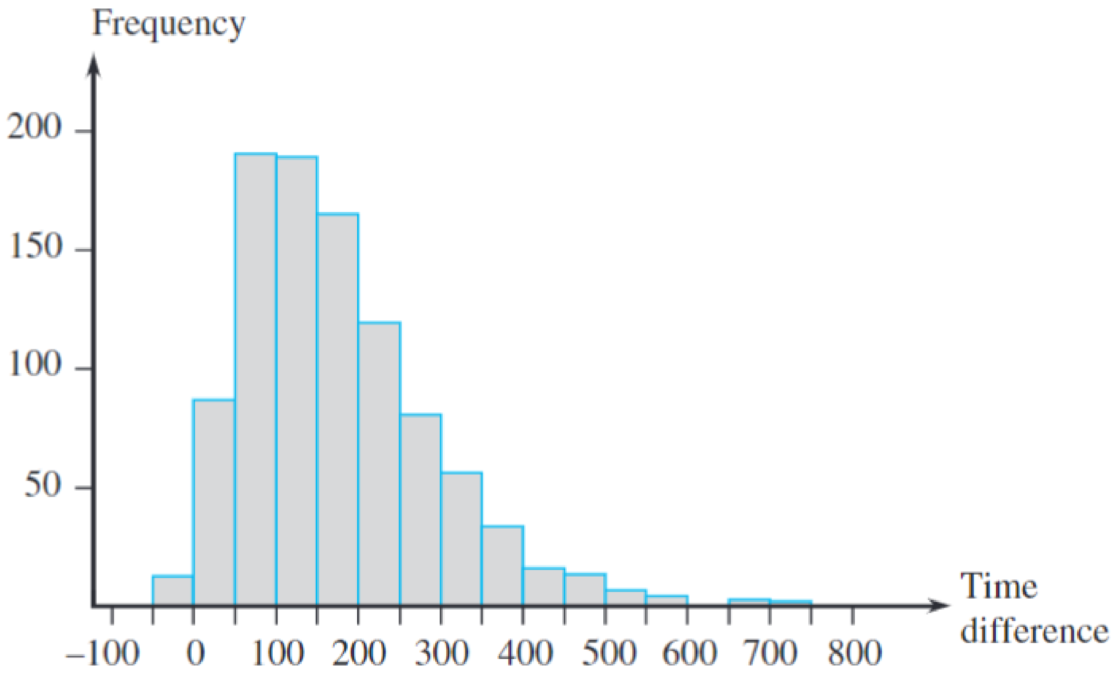
c. What proportion of these corporations have at most 10 directors.

Solution:

```
> View(ex01.18)
> x = X.Frequenc.
> sum(x)
  [1] 204
> sum = 0;
% 7 = Index of the value of 10
> for (i in 1:7){
  +   sum<- sum+x[i]
  + }
> sum
  [1] 142
> 142/204
  [1] 0.6960784
```

Thus the proportion of companies with at most 10 board members is $\frac{142}{204} \approx .6961$ or 69.61%.

Exercise 1.22: How does the speed of a runner vary over the course of a marathon? Consider determining both the time to run the first 5km and the run between the 35-km and 40-km points, and then subtracting the former time from the latter time. A positive value of this difference corresponds to a runner slowing down toward the end of the race. The accompanying histogram is based on times of runners who participated in several different Japanese marathons. What are some interesting features of this histogram? What is a typical difference value? Roughly what proportion of runners ran the late distance more quickly than the early distance?



Solution: Again this histogram appears to be roughly unimodal with a noticeable right skew. Since a positive value of the difference means that the runner is slowing down at the end of the race the histogram shows us that a large majority of runners are in fact slowing down. Furthermore since the histogram is positively skewed we know that some runners seem to slow down much more than the median runner, making the average greater and creating a positive skew. Given that the histogram is unimodal we can hazard a guess that the representative value would be around 175 seconds. Roughly estimating the proportion of runners ran the late distance more quickly than the early distance I would say they account for less than 1% of the population.

Exercise 1.38: Blood pressure values are often reported to the nearest 5 mmHg. Suppose the actual blood pressure values for nine randomly selected individuals are,

```
'Blood_Pres (mmHg)'
118.599998474121
127.400001525879
138.399993896484
130
113.699996948242
122
108.300003051758
131.5
133.199996948242
```

a. What is the median (and mean) of the reported blood pressure values?

Solution:

```
>    View(ex01.38)
> x <- ex01.38$X.Blood_Pres
> fivenum(x)
Minimum    Lower Quartile    Median    Upper Quartile    Maximum
 108.3          118.6          127.4          131.5          138.4
> mean(x)
   [1] 124.7889
```

From the *r* computation we can see that the *Median* = 127.4 and the *Mean* = 124.788.

b. Suppose the blood pressure of the second individual is 127.6 rather than 127.4. How does this affect the median of the reported values? What does this say about the sensitivity of the median to rounding or grouping in the data?

[Solution:]

```
> x[2] = 127.6
> fivenum(x)
Minimum   Lower Quartile   Median   Upper Quartile   Maximum
 108.3        118.6         127.6        131.5        138.4
> mean(x)
  [1] 124.8111
```

We can see that our rounding experiment a greater effect on the median than the mean with a δ of .2 as apposed to a δ of .03. Had our rounding been enough to change the order of our data points I could see it greatly affecting the median, however in this case it did not. Note that the median is not sensitive to the grouping of data since by definition it is simply the middle value. For example suppose a bimodal data set where the median occurs at the trough.

Exercise 1.42:

a. If a constant c is added to each x_i in a sample, yielding $y_i = x_i + c$, how do the sample mean and median of the y_is relate to the mean and median of the x_is? Verify you conjectures.

Proof: Suppose a sample set y and x such that $y_i = x_i + c$. By the definition of sample mean we know,

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

Using our definition of y_i and some algebra we get,

$$\begin{aligned}\overline{y} &= \frac{\sum_{i=1}^{n} x_i + c}{n},\\ &= \frac{(\sum_{i=1}^{n} x_i) + (cn)}{n},\\ &= \overline{x} + c.\end{aligned}$$

Thus we have shown that the given sample results in a sample mean of $\overline{y} = \overline{x} + c$. Furthermore, regardless of the parity of n, the middle value, and average of the two middle values still gets shifted by c therefore we get that $\hat{y} = \hat{x} + c$.

□

b. If each x_i is multiplied by a constant c, yielding $y_i = cx_i$, answer the questions of part a. Again verify you conjectures.

Proof: Similarly suppose a sample set y and x such that $y_i = cx_i$. By the definition of sample mean we know,

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

Using our definition of y_i and some algebra we get,

$$\begin{aligned} \overline{y} &= \frac{\sum_{i=1}^{n} cx_i}{n}, \\ &= \frac{(c\sum_{i=1}^{n} x_i)}{n}, \\ &= c\overline{x}. \end{aligned}$$

We have shown that the given sample results in a sample mean of $\overline{y} = c\overline{x}$. And again, regardless of the parity of n, the middle value, and average of the two middle values still gets shifted by c therefore we get that $\hat{y} = c\hat{x}$.

□

Exercise 1.44: Poly (3-hydroxybutyrate) (PHB), a semicrystalline polymer that is fully biodegradable and bio-compatible, is obtained from renewable resources. From a sustainability perspective, PHB offers many attractive properties though it is more expensive to produce than standard plastics. The accompanying data on melting point (°C) for each 12 specimens of the polymer using a differential scanning calorimeter.

a. Compute the sample range.

Solution:

```
>     View(ex01.44)
> x <- ex01.44$X.Melting_Point.
> max(x)-min(x)
  [1] 2.3
```

Therefore the sample range is 2.3 $°C$.

b. Compute the sample variance s^2 from the definition. [Hint: First subtract 180 from each observation.]

Solution: Recall the definition of s^2,

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

From the r computation,

```
> x <- ex01.44$X.Melting_Point.
> m = mean(x)
> y = x-m
> sum(y^2)/11
  [1] 0.5244697
```

we get that the sample variance, $s^2 = 0.5244697$.

c. Compute the sample standard deviation s .

Solution: By the definition of sample standard deviation we know that it is simply the positive square root of the sample variance. Thus $s = 0.7242028$

Exercise 1.50: In 1997 a woman sued a computer keyboard manufacturer, charging that her repetitive stress injuries were caused by the keyboard. The injury awarded about 3.5 million for pain and suffering, but the court then set aside that award as being unreasonable compensation. In making this determination, the court identifies a ”normative” group of 27 similar cases and specifies a reasonable award as one within two standard deviations of the mean of the awards in the 27 cases. What is the maximum possible amount that could be awarded under the two-standard-deviation rule?

Solution:

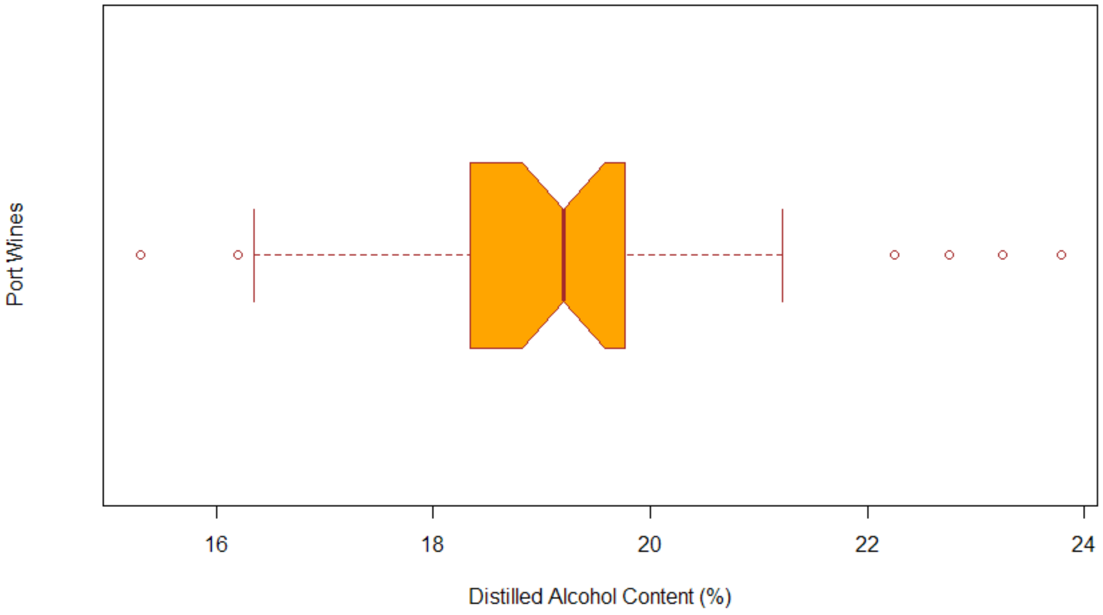
```
> x <- ex01.50$X.awards.
> m = mean(x)
> ss = sd(x)
> m+2*ss
  [1] 1961.158
```

The maximum amount that could be awarded under the 2-standard-deviation rule is $1,961,158. Any more would not be within 2 standard deviations.

Exercise 1.56: The following data on distilled alcohol content for a sample of 35 port wines. Each value is an average of two duplicate measurements. Use methods from this chapter, including a boxplot that show outliers, to describe and summarize the data.

Solution: Analyzing the data with r,

```
> x <- ex01.56$X.alcohol
> fivenum(x)
Minimum   Lower Quartile   Median   Upper Quartile   Maximum
15.30        18.34          19.20        19.76         23.78
> mean(x)
  [1] 19.25743
> sd(x)
  [1] 1.831662
> var(x)
  [1] 3.354984
> IQR(x)
  [1] 1.42
> boxplot(x, horizontal = TRUE, ylab = 'Port Wines',
  +         xlab = 'Distilled Alcohol Content (%)', col = "orange",
  +         border = "brown", notch = TRUE)
```



Firstly we can see that there is a relatively small δ between the upper/lower forth and the median, the relatively small IQR and the δ between the median and the mean, despite the large number of positive outliers demonstrated in the boxplot suggests that most of the data points are grouped between 18 to 20 percent. Furthermore this means that the representative value could be either the mean or median.

Exercise 2.4: Each of a sample of four home mortgages is classifies as fixed rate (F) or variable rate (V).

a. What are the 16 outcomes of s

Solution:

$$s = \begin{bmatrix} FFFF & VFFF & FVFV & VVFV \\ FFFV & FFVV & VFFV & VFVV \\ FFVF & FVVF & VFVF & FVVV \\ FVFF & VVFF & VVVF & VVVV \end{bmatrix}$$

b. Which outcomes are in the event that exactly three of the selected mortgages are fixed?

Solution:

$$\begin{bmatrix} VFFF \\ FFFV \\ FFVF \\ FVFF \end{bmatrix}$$

c. Which outcomes are in the event that all four mortgages are of the same type?

Solution:

$$\begin{bmatrix} FFFF \\ VVVV \end{bmatrix}$$

d. Which outcomes are in the event that at most one of the four is a variable-rate mortgage?

Solution:

$$\begin{bmatrix} FFFF \\ FFFV \\ FFVF \\ FVFF \\ VFFF \end{bmatrix}$$

e. What is the union of the events in parts (c) and (d), and what is the intersection of these two events?

Solution:

$$c \cup d = \begin{bmatrix} VFFF \\ FFFV \\ FFVF \\ FVFF \end{bmatrix}$$

$$c \cap d = \begin{bmatrix} FFFF \end{bmatrix}$$

f. What is the union of the events in parts (c) and (b), and what is the intersection of these two events?

Solution:

$$c \cup b = \begin{bmatrix} VVVV \\ FFFF \\ FFFV \\ FFVF \\ FVFF \\ VFFF \end{bmatrix}$$

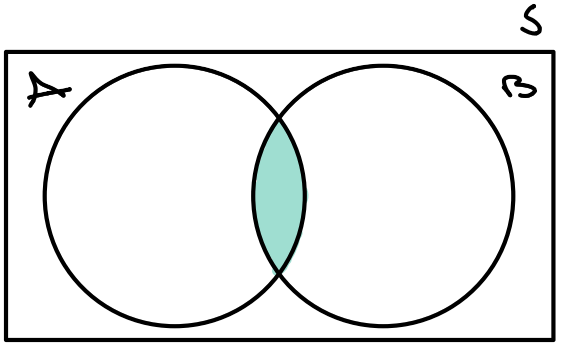
$$c \cap b = \varnothing$$

Exercise 2.9: Use Venn Diagrams to verify the following two relationships for any events A and B,

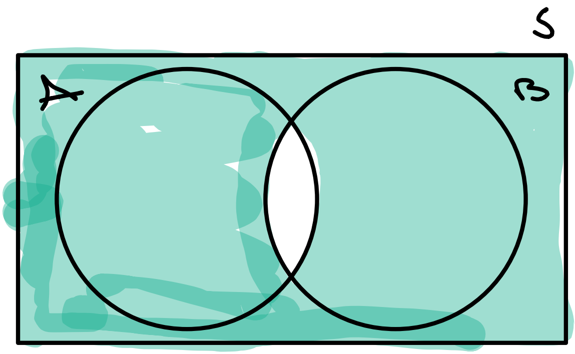
a. $(A \cap B)' = A' \cup B'$

b. $(A \cup B)' = A' \cap B'$

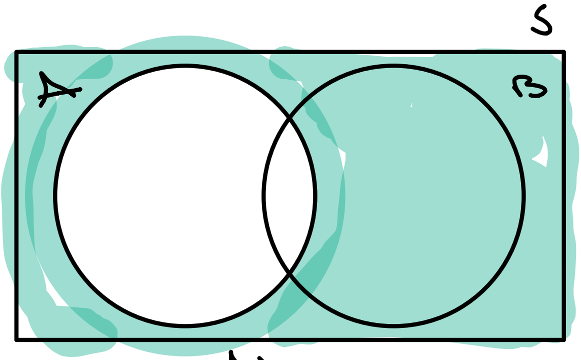
$$(A \cap B)' = A' \cup B'$$



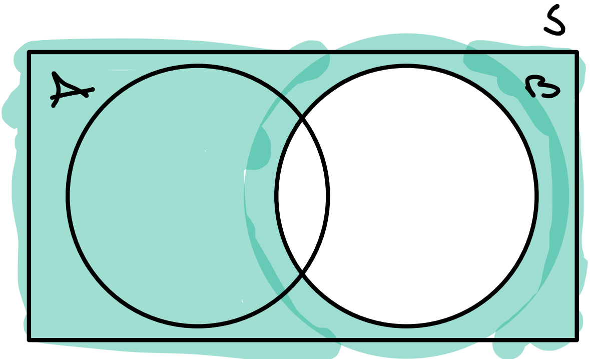
$A \cap B$



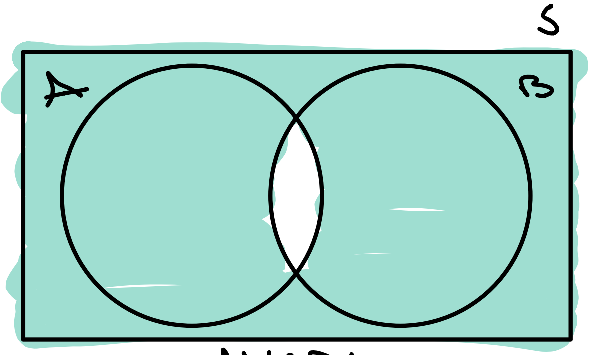
$(A \cap B)'$



A'

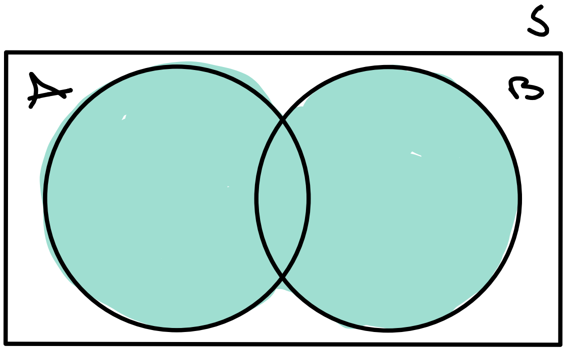


B'

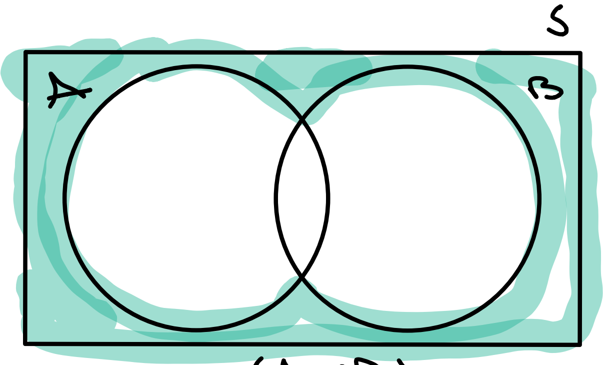


$A' \cup B'$

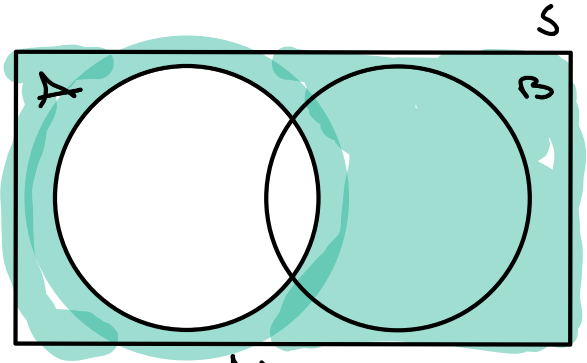
$(A \cup B)' = A' \cap B'$



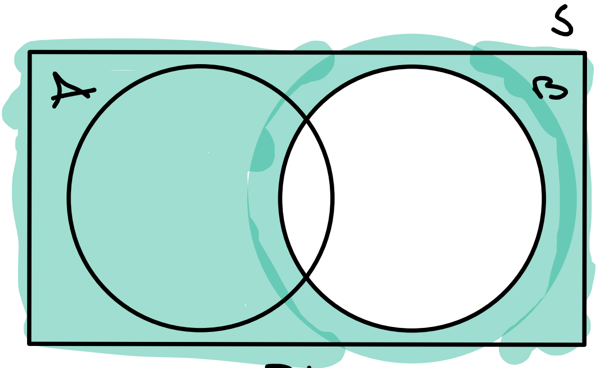
$A \cup B$



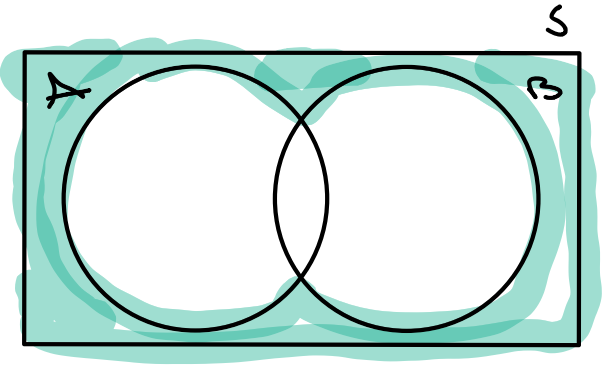
$(A \cup B)'$



A'



B'



$A' \cap B'$