Problem 1

(a) Step 1: Observe the histogram and give and value of each bar, then get the density function.

From -20 to -15, the height is 0.001, and the proportion is $5 \times 0.001 = 0.005$. From -15 to -10, the height is 0.004, and the proportion is $5 \times 0.004 = 0.02$. From -10 to -5, the height is 0.025, and the proportion is $5 \times 0.025 = 0.125$. From -5 to 0, the height is 0.07, and the proportion is $5 \times 0.07 = 0.35$. From 0 to 5, the height is 0.065, and the proportion is $5 \times 0.065 = 0.325$. From 5 to 10, the height is 0.03, and the proportion is $5 \times 0.03 = 0.15$. From 10 to 15, the height is 0.005, and the proportion is $5 \times 0.005 = 0.025$.

And you can find the sum of proportion is 1. Thus, we have:

$$
\begin{aligned}
&P(-20 \le x < -15) = 0.005,\\
&P(-15 \le x < -10) = 0.02,\\
&P(-10 \le x < -5) = 0.125,\\
&P(-5 \le x < 0) = 0.35,\\
&P(0 \le x < 5) = 0.325,\\
&P(5 \le x < 10) = 0.15,\\
&P(10 \le x < 15) = 0.025,
\end{aligned}
$$

Moreover, as the histogram, it's obviously density function in each internal is constant. Thus, for example, in interval $[-20, 15)$, we can get:

$\int_{-20}^{-15} f(x)dx = P(-20 \le x < -15) = 0.005$

Which implies:

For $x \in [-20, -15), f(x) = \frac{P(-20 \le x < -15)}{-15-(-20)} = 0.001$.

Actually, in this case, in each interval, the value of density function $f(x)$ equal to the height of bar.

Step 2: Calculate Median: From the probability function given above, we can find $P(x < 0) = 0.5$ and $P(x > 0) = 0.5$, thus the median is 0.

Step 3: Calculate Mean: Use the Formula: $EX = \int_{-20}^{15} xf(x)dx$, we can have:

$$
\begin{aligned}
EX &= \int_{-20}^{15} xf(x)dx \\
&= \int_{-20}^{-15} xf(x)dx + \int_{-15}^{-10} xf(x)dx + \int_{-10}^{-5} xf(x)dx + \int_{-5}^{0} xf(x)dx + \\
&\quad \int_{0}^{5} xf(x)dx + \int_{5}^{10} xf(x)dx + \int_{10}^{15} xf(x)dx
\end{aligned}
$$

Here, we need to calculate each part. Consider the interval $[a, b)$, let $f(x) = \frac{P(a \leq x < b)}{b-a}$, then we know:

$$\int_a^b xf(x)dx = \frac{P(a \leq x < b)}{b-a} \int_a^b xdx = \frac{a+b}{2} \cdot P(a \leq x < b)$$

Thus, go back to expectation, we have:

$$\begin{aligned} EX =& 0.005 \times (-17.5) + 0.02 \times (-12.5) + 0.125 \times (-7.5) + 0.35 \times (-2.5) \\ &+ 0.325 \times 2.5 + 0.15 \times 7.5 + 0.025 \times 12.5 \\ =& 0.1 \end{aligned}$$

Step 4: Standard deviation: Calculate variance first, and its square root is standard deviation.

As the formula $Var(X) = EX^2 - (EX)^2$ and expectation got above, we only need to find EX^2. Use the same method as expectation:

$$\begin{aligned} EX^2 =& \int_{-20}^{15} x^2 f(x)dx \\ =& \int_{-20}^{-15} x^2 f(x)dx + \int_{-15}^{-10} x^2 f(x)dx + \int_{-10}^{-5} x^2 f(x)dx + \int_{-5}^{0} x^2 f(x)dx + \\ & \int_0^5 x^2 f(x)dx + \int_5^{10} x^2 f(x)dx + \int_{10}^{15} x^2 f(x)dx \end{aligned}$$

And still consider the interval $[a, b)$, let $f(x) = \frac{P(a \leq x < b)}{b-a}$:

$$\int_a^b x^2 f(x)dx = \frac{P(a \leq x < b)}{b-a} \int_a^b x^2 dx = \frac{a^2+ab+b^2}{3} \cdot P(a \leq x < b)$$

Go back to EX^2, we have:

$$\begin{aligned} EX^2 =& 0.005 \times [(-20)^2 + (-20)(-15) + (-15)^2] + 0.02 \times [(-15)^2 + (-15)(-10) \\ &+ (-10)^2] + \cdots + 0.025 \times [10^2 + 10\dot{1}5 + 15^2] \\ =& 91 \end{aligned}$$

Thus standard deviation is $\sqrt{Var(X)} = \sqrt{EX^2 - (EX)^2} \simeq 9.54$.

Step 5: 20% percentile data: By observing the histogram and the probability function we get, we can conclude the u, which makes $P(x < u) = 20\%$, is between -5 and 0, so any answer from -5 to 0 is acceptable.

(b) if this is a frequency histogram, then make the y axis divide by sample size 1000 $\times$ bin, then we could get the density histogram which has the same shape of frequency histogram.

Another way for Problem 1. To find mean and standard deviation (SD) from the histogram, here is another way to give an approximation. We can see there are $0.005 \cdot 1000$ points between -20 and -15 and thus may consider all these 5 points concentrated at the middle point -17.5. In other words, we may approximate the data as taking the value -17.5 with probability 0.005, similarly, -12.5 with probability 0.02, -7.5 with probability 0.125, -2.5 with probability 0.35, 2.5 with probability 0.325, 7.5 with probability 0.15 and 12.5 with probability 0.025. Hence,

$$\begin{aligned}\text{Mean} &= (-17.5)\cdot 0.005 + (-12.5)\cdot 0.02 + (-7.5)\cdot 0.125 + (-2.5)\cdot 0.35 + \\ &\quad 2.5\cdot 0.325 + 7.5\cdot 0.15 + 12.5\cdot 0.025 \\ &= 0.1,\end{aligned}$$

and

$$\begin{aligned}\text{SD} &= [(-17.5-0.1)^2\cdot 0.005 + (-12.5-0.1)^2\cdot 0.02 + (-7.5-0.1)^2\cdot 0.125 + \\ &\quad (-2.5-0.1)^2\cdot 0.35 + (2.5-0.1)^2\cdot 0.325 + (7.5-0.1)^2\cdot 0.15 + \\ &\quad (12.5-0.1)^2\cdot 0.025]^{1/2} \\ &\approx 9.54.\end{aligned}$$

`Problem 2`

The mean of the second data set is $\frac{1}{n}\sum_i y_i = \frac{1}{n}\sum_i ax_i + b = a\frac{1}{n}\sum_i x_i + \frac{1}{n}\sum_i b = a\bar{x} + \frac{1}{n}nb = a\bar{x} + b$.

The variance of the second data set is $\frac{1}{n-1}\sum_i (y_i - \bar{y})^2 = \frac{1}{n-1}\sum_i (ax_i + b - (a\bar{x}+b))^2 = \frac{1}{n-1}\sum_i (ax_i - a\bar{x})^2 = a^2\frac{1}{n-1}\sum_i (x_i - \bar{x})^2 = a^2 s_1^2$.

`Problem 3`: First label the pictures: in the first line from left to right as 1, 2, 3 and in the second line from left to right as 4, 5, 6. Then let us decide the shape of the data whether is symmetric, right-skewed or left-skewed. Our rule is that if the median is in the middle of the first quantile and third quantile, then the data is symmetric, if the median inclines to the first quantile, then it is right-skewed and similarly, if the median to the third quantile, then it is left-skewed. So we have in figure 4 and 6 data is symmetric, in figure 1 and 5 data is left-skewed, and in figure 2 and 3 data is right-skewed. Next let us consider volatility of the data. Now we focus on the outliers. Comparing figure 4 and 6, we can see figure 4 has more outliers in the left which makes the data much volatile and thus figure 4 is in large volatile and figure 5 in small volatile. Then comparing figure 1 and 5, since we have known that they are both left-skewed, then if there are more outliers in the left, it means data has large volatility. Therefore, figure 1 is in large volatile while figure 5 in small volatile. Similarly, we have figure 2 is in small volatile while figure 3 is in large volatile.

In conclusion, data in figure 1 is from left-skewed and large volatile distribution, in figure 2 from right-skewed and small volatile, in figure 3 from right-skewed and large volatile, in figure 4 from symmetric and large volatile, in figure 5 from left-skewed and small volatile, in figure 6 from symmetric and small volatile.

`Problem 4`

(a) $P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{3} + \frac{2}{3} - \frac{1}{9} = \frac{8}{9}$.

(b) No, $P(\text{only } A) \neq P(A \cup B \cup C) - P(B) - P(C)$. The event "only A" is the complement of $B \cup C$, so it'd be $P(A \cup B \cup C) - P(B \cup C) = P(A \cup B \cup C) - (P(B) + P(C) - P(B \cap C) = P(A \cup B \cup C) - P(B) - P(C) + P(B \cap C)$.

(c) We begin with $P(A \cup B \cup C)$, but that's wrong because it overcounts the double intersections $A \cap B, A \cap C, B \cap C$ (just like in the case with two sets). So we remove them, giving us $P(A \cup B \cup C) - P(A \cap B) - P(A \cap C) - P(B \cap C)$. But the triple intersection $A \cap B \cap C$ is in each of the double intersections, so it's been removed and isn't being counted (the triple intersection is a subset of A, B, and C, as well as of $A \cap B, A \cap C, B \cap C$) . Hence we add back in the triple intersection, giving us $P(A \cup B \cup C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$ as the correct answer.

`Problem 5: Larsen and Marx 2.7.12` We are going to arrange the letters in the phrase *A ROLLING STONE GTHERS NO MOSS* randomly and ask what is the chance that not all of the S's will be adjacent. There are 4 S's out of 26 total letters in the phrase. Let us assume that each letter is a distinct object, or in other words, we are going to label the first, second, third, and fourth S so that we can distinguish between them, and the same with the other letters (**Note: I am assuming this is what the question means when it says that the letters are arranged at random. There is more than one way to interpret this**). Since all of the objects are distinct, there are 26! possible permutations of the letters, all equally likely. We should count the number of ways that we can have a permutation that has all 4 S's in a row. This will give us the probability that all of the S's will be adjacent, which we can then subtract from 1 to get the probability that the question asks for. So how many ways can we permute the letters so that all 4 S's are adjecent?

First we should pick where the block of S's will be in our sequence of letters. Since the block has length 4, it must start somewhere between the first and twenty-third letter in the sequence, so there are 23 ways to pick where the block of S's will be.

Once we pick where the block of S's will be, there are $4! = 24$ ways to permute the 4 S's within the block of 4.

Once we do that, we must decide how many ways there are to pick the rest of the sequence. There are 22! ways to do this.

So the probability that there are 4 S's in a row is given by

$$\frac{23 \cdot 4! \cdot 22!}{26!},$$

which means that the probability that there are not 4 S's in a row is given by

$$1 - \frac{23 \cdot 4! \cdot 22!}{26!}.$$

`Problem 6: Larsen and Marx 2.7.15` First we wil figure out the probability that the first two people in the group have the same birthday, and nobody else in the group does, then we will multiply this number by $\binom{k}{2}$. We can solve the problem this way because each event is disjoint from one another, i.e. if Person 1 and Person 2 share a birthday and nobody else does, then it is impossible that Person 3 and Person 4 share a birthday and nobody else does.

We assume that birthdays are independent events that are evenly distributed, each day having a $\frac{1}{365}$ chance of being a person's birthday. Then the probability that Person 1 and Person 2 will have the same birthday is $\frac{1}{365}$ (one way to see this is to notice that it doesn't matter what Person 1's birthday is, just that Person 2's matches it). Now we must find the probability that they are the only pair that have this birthday. This means that the other $k-2$ people must take $k-2$ distinct bithdays out of the remaining 364 days. We can figure out this probability by considering each person individually in a sequence, and then multiplying the probabilities. Assume we know Person 1 and 2's birthday, and we want no other pair to share a birthday. If we reveal Person 3's birthday, then the chance that he does not violate this condition is $\frac{364}{365}$ (he cannot have the same birthday as Person 1 and 2). Now we reveal Person 4's birthday and the chance that we do not violate the condition is $\frac{363}{365}$ (he cannot have Person 1 and 2's birthday, nor Person 3's birthday). Continuing, we see that the probability that Person 1 and 2 share the same birthday and nobody else does is given by

$$\frac{1}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{365-(k-2)}{365}.$$

Thus the probability that exactly one pair of people have the same birthday is given by

$$\binom{k}{2}\frac{1}{365}\cdot\frac{364}{365}\cdot\frac{363}{365}\cdot\cdots\frac{365-(k-2)}{365}.$$

Note if there are more than 366 people the condition is impossible to satisfy, and our formula gives that the probability is 0.

`Problem 7: Larsen and Marx 2.7.19` Once Tim has drawn his hand and discarded, there are 47 cards remaining and he will choose 2 of them. Thus there are $\binom{47}{2}$ possibilities for what he can draw. Now we must count the number of ways for him to draw a straight flush or a flush. Since Tim has a 6, 8, and 9 of hearts, there are two ways to draw into a straight flush (drawing the 5 and 7 of hearts, or drawing the 7 and 10 of hearts). Thus the chance that he draws into a straight flush is given by

$$\frac{2}{\binom{47}{2}}.$$

Now, to draw into a flush he must draw two hearts. There are 10 hearts remaining in the deck, so there are $\binom{10}{2}$ ways to draw two hearts. Thus his chance of drawing into any flush is

$$\frac{\binom{10}{2}}{\binom{47}{2}}.$$

If we consider drawing into a straight flush to not count as drawing into a flush, then his chance of drawing into a flush is given by

$$\frac{\binom{10}{2}-2}{\binom{47}{2}}.$$

`Problem 8: OpenIntro 2.12`

(a) $P(0) = 1 - P(1) - P(2) - P(\geq 3) = 1 - .25 - .15 - .28 = .32$

(b) $P(\leq 1) = P(0) + P(1) = 1 - P(2) - P(\geq 3) = 1 - .15 - .28 = .57$

(c) $P(\geq 1) = 1 - P(0) = 1 - .32 = .68$

(d) $P(\text{neither miss any school}) = P(\text{first misses no school}) \times P(\text{second misses no school}) = P(0) \times P(0) = .32^2$. Note that we assumed that the two events "first misses no school" and "second misses no school" were independent!

(e) $P(\text{both miss at least one}) = P(\text{first miss at least one}) \times P(\text{second miss at least one}) = P(\geq 1) \times P(\geq 1) = .68^2$. Again, we assumed that the two events were independent.

(f) We assumed whether a child missing school or not was independent of his/her sibling missing school or not. This is probably not true, since children living together can get sick because of each other. However, our answer could still be somewhat close to the correct answer.