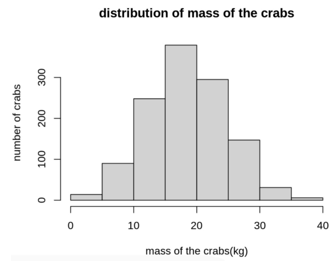
(1) First we will examine the masses of individual crabs:

a. Check whether crab mass is approximately normally distributed (16 points)

The crab mass is approximately normally distributed for four reasons. First, the mean and the median is approximately equal to each other. Also, the skewness is equal to 0 as well. Furthermore, according to histogram, the dataset is normally distributed. Finally, qq-plot, which present in a straight line, also show it is normally distributed.



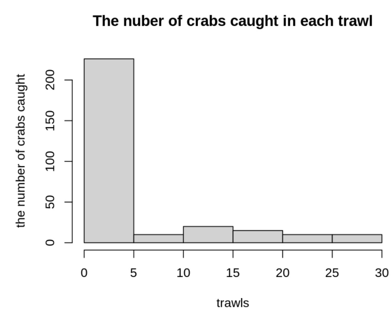
b. Calculate a range of masses within which you would expect 95% of crabs to fall (4 points)
The range of masses within which I would expect 95% of crabs to fall is from 6.099 kg to 30.857 kg.

c. Calculate the probability of catching a large crab weighing 25 kg or more (4 points)
The probability of catching a large crab weighing 25 kg or more is 14.6%.

(2) Now we will examine the number of crabs caught in each trawl. This poses two problems: first, crabs tend to be found in groups, so many trawls result in 0 crabs caught; second, the number caught is technically a discrete variable.

a. This data is zero-inflated, meaning it has a relatively simple distribution, except for the fact that it is way more zeros than is “normal” for that distribution. Make a plot of the data distribution to confirm this (4 points).

The plot confirms that the data is zero-inflated, because there are many more zeros than there are in normal distribution.



b. Using Boolean operators, make a new variable nonzero.catch that captures the total number of crabs caught by a trawler when the trawler catches at least one crab. (Note that this is not always the correct way to deal with zero-inflation; 4 points)

c.Normally we might wish to use the Poisson distribution to model this variable, since it is a discrete count of “events” that happen without a strict upper limit. Calculate the sample mean and sample variance of this variablenonzero.catch, and explain why the Poisson distribution might not be appropriate (4 points).

The Poisson distribution might not be appropriate since the mean and the variance of the dataset are not equal.
Variance: 67.81982
Mean: 16.13333

d. For discrete variable such as this, it is common to use the normal distribution as a way to simplify calculations, even though the data is really discrete. Make a Q-Q plot to check whether the validity of this assumption (Note: in a real analysis you would also use the other metrics as well; 4 points)

The QQ plot shows that the assumption is not valid. In the QQ-plot of normal distribution, the dots(observations) should lie on approximately a straight line. However, for the dataset of nonzero-catch, the dots are not on a straight line.

ILP Problems

(1) Suppose every hour a staph infection grows by a factor of Rt, where log(Rt) is Normal(0.3,0.8). Calculate the average growth factor per hour, E[Rt], and interpret what this means (5 points).

Because of right-ward skew, the mean value should be calculate as exp(mean+variance/2). If the log(Rt) is Normal(0.3,0.8), then E[Rt] is equal to exp((mean+variance)/2), which is exp((0.3+0.8^2)/2). E[Rt] equals 1.599994.

(2) Calculate the standard deviation of Rt. Compare it to E[Rt], and interpret what this means given that log-normal random variables can only have positive values (5 points).

Because of right-ward skew, the variance of lognormal distribution should be calculated as [exp(variance)-1]exp(2mean + variance). If the log(Rt) is Normal(0.3,0.8), then the variance of E[Rt] is [exp(0.8^2)-1]exp(2(0.3) + 0.8^2) is 3.097891. The standard deviation is the square root of variance, which is 1.76008.

(3) **Use the log-normal CDF in R to calculate the probability that, over the course of 1 hour, the staph infection actually decreases in size** (5 points)
The probability that over the course of 1 hour, the staph infection actually decreases in size is 18%.
plnorm(1, mean, sd)