1. Suppose you have a continuous distribution with

\[ f(x) = \begin{cases} 
-\frac{x}{2} & -1 \leq x < 0 \\
\frac{x}{2} & 0 \leq x \leq 1 \\
0 & \text{otherwise}
\end{cases} \]

a. Sketch this distribution.

b. What is the mean of this distribution.

That this distribution is perfectly symmetrical (see above), the mean and the median would be at the same location, right in the center, the mean must be zero. But, for those who prefer math:

\[ E(X) = \mu_X = \int_{-\infty}^{\infty} x f(x) dx = \int_{-1}^{0} -x^2 dx + \int_{0}^{1} x^2 dx = \left[-\frac{x^3}{3}\right]_{x=-1}^{0} + \left[\frac{x^3}{3}\right]_{x=0}^{1} = \left(0 - \left(-\frac{1}{3}\right)^3\right) + \left(\frac{1}{3} - 0\right) = 0 \]

c. Find the standard deviation of this distribution.

First we find the variance, \( \sigma^2 \):

\[ \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \int_{-\infty}^{0} x^2 f(x) dx + \int_{0}^{1} x^2 f(x) dx = \int_{-1}^{0} -x^3 dx + \int_{0}^{1} x^3 dx = \left[-\frac{x^4}{4}\right]_{x=-1}^{0} + \left[\frac{x^4}{4}\right]_{x=0}^{1} = \left(0 - \left(-\frac{1}{4}\right)^4\right) + \left(\frac{1}{4} - 0\right) = \frac{1}{2} \]

Thus we find that the standard deviation \( \sigma = \frac{1}{\sqrt{2}} = 0.7071 \)

d. Find \( P(X > 0.5) \).

\[ P(X > 0.5) = \int_{0.5}^{\infty} f(x) dx = \int_{0.5}^{1} x^2 dx = \left[\frac{x^3}{3}\right]_{x=0.5}^{1} = \left(\frac{1}{3} - \left(\frac{1}{2}\right)^3\right) = \frac{1}{2} - \frac{1}{8} = \frac{3}{8} = 0.375 \]

e. Find the 0.5th percentile of this distribution (i.e. determine \( h \) such that \( \int_{-\infty}^{h} f(x) dx = 0.005 \)).

Clearly the 0.5th percentile will be negative.

\[ \int_{-\infty}^{h} f(x) dx = \int_{-1}^{h} -x^2 dx = \left[-\frac{x^3}{3}\right]_{x=-1}^{h} = -\frac{h^3}{3} + \frac{1}{3} = -\frac{h^2}{2} - \frac{1}{2} = \frac{1-h^2}{2} \]

Setting this equal to .005 and solving for \( h \):

\[ \frac{1}{200} = \frac{1-h^2}{2} \quad \text{and so} \quad \frac{1}{100} = 1-h^2 \quad \text{which implies that} \quad h^2 = 1 - \frac{1}{100} = .99 \quad \text{so we know} \quad h = \pm \sqrt{0.99} = \pm 0.9950. \]

Because we know that our value is negative, our 0.5th percentile is -.9950.
2. Suppose that you have 100 datapoints with summary statistics:

<table>
<thead>
<tr>
<th>mean</th>
<th>sd</th>
<th>Min</th>
<th>Q1</th>
<th>Med</th>
<th>Q3</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.023</td>
<td>0.735</td>
<td>-0.995</td>
<td>-0.717</td>
<td>0.023</td>
<td>.721</td>
<td>.997</td>
</tr>
</tbody>
</table>

a. Sketch a boxplot of such data.

![Boxplot](image)

b. Is this dataset skewed to the left, symmetrical or skewed to the right? Explain your answer.

These data appear approximately symmetrical. The quartiles are approximately equidistant from the median. Same for the min and max. Furthermore, the mean and median are exactly the same. Yes … all signs of symmetry.

c. Explain, conceptually, why you know that this dataset has no outliers.

If outliers are defined as datapoints more than 1.5 IQRs below Q₁ or above Q₃ there are none.

d. If there were another datapoint added to the dataset, a zero, explain conceptually or from an equation how the standard deviation would change.

The standard deviation would decrease. The new datapoint will increase n, but the mean won’t change much at all so the sum of the squared deviations, $\sum (x_i - \bar{x})^2$, will not change much at all.

e. Would a normal quantile plot show that these data could have been from a normal distribution or that they would certainly not have been from a normal distribution? Explain your reasoning.

Normal would not match up well to these data. Normal would suggest that about 5% of the distribution be outside the ± 1 SD range. Here the SD is 0.7 so we should have about 5% of the values outside the ± 1.4 range. Here we have 100% of the values inside the ± 1 range. [Note: our mean, median, min, max and standard deviation are nearly exactly what one would expect based on question 1’s distribution … this is just a sample of 100 datapoints from that distribution.]
3. Suppose the fuel efficiency of 2016 Ford F-150 trucks driven on the highway by real people in regular conditions is normally distributed with mean 24.3 and standard deviation 1.1 miles per gallon, respectively.
   
a. What percentage of Ford F-150 trucks would have miles per gallon below 26?

   \[ Z = \frac{26 - 24.3}{1.1} = 1.55 \]

   Thus (from the normal table) we say that 93.94% of trucks would have miles per gallon below 26.

   b. Why, then, does Ford say that that the F-150 gets “up to 26 miles per gallon” as their highway fuel efficiency number?

   The best answer here would discuss how “up to 26 miles per gallon” makes sense because some 6% of the trucks achieve this mpg level.

   c. What is the 99th percentile of highway fuel efficiency for the F-150?

   First, \( Z = 2.33 \). Thus the 99th percentile is \( x = \mu + 2.33\sigma = 24.3 + 2.33 \times 1.1 = 26.86 \)

   d. If you took a random sample of 200 F-150 owners, what is the chance that exactly one of them has highway miles per gallon more than the 99th percentile?

   So, if only 1% of the truck owners have a mpg more than the 99th percentile, the number out of 200 sampled truck owners who would have such an mpg is a binomial variable with \( n=200 \) and \( p=.01 \).

   \[
   P(1) = \frac{200!}{1!(200-1)!} \cdot .01^1 (1-.01)^{200-1} = 100 \times .01 \times .99^{199} = .2707
   \]

   One could also use the Poisson approximation to the Binomial (because we have a large \( n \) and small \( p \), so \( \lambda = np = 200 \times .01 = 2 \) and \( P(1) = \frac{e^{-2} \times 2^1}{1!} = 2e^{-2} = .2707 \).
4. If the distribution of the number of trees per acre in a semi-wooded area is Poisson with mean 27:
   a. What proportion of acres would have exactly 20 trees?

   \[ P(27) = \frac{e^{-27} 27^{20}}{20!} = 0.0327 \]

   b. Approximately what proportion of acres would have less than 20 trees?

   Here we use the normal approximation to the Poisson (legit because \( \lambda > 25 \)) and find that ...

   \[ P(X < 20) \approx P \left( Z < \frac{(20 - 5) - 27}{\sqrt{27}} \right) = \Phi(-1.44) = 0.0749. \]

   c. If we were to go to 20 such semi-wooded areas, what is the chance that exactly two of them would have less than 20 trees?

   This seems Binomial to me ... with \( n=20 \) and \( p=0.0749 \). Then

   \[ P(2) = \frac{20!}{2!(20-2)!} \times 0.0749^2 \times (1 - 0.0749)^{20-2} = 0.2625 \]

   d. Approximately what proportion of acres would have 22 to 30 trees?

   We can again use the Normal approximation:

   \[ P(22 \leq X \leq 30) \approx P \left( \frac{(22 - 0.5) - 27}{\sqrt{27}} < Z < \frac{(30 + 0.5) - 27}{\sqrt{27}} \right) = \Phi(0.67) - \Phi(-1.06) = 0.7486 - 0.1446 = 0.6040 \]

   e. If you were to go to two such semi-wooded areas, the number of trees in total would follow a Poisson distribution with mean \( 2\lambda \). What would the standard deviation of the number of trees per two acres be?

   This would suggest that we have Poisson with mean \( 2 \times 27 = 54 \). The SD of such a distribution is \( \sqrt{54} = 7.35 \).