Because a diet rich in saturated fats raises the cholesterol level, it is plausible that dogs owned as pets have higher cholesterol levels than dogs owned by a veterinary research clinic. A clinic compared the cholesterol levels (in mg per deciliter) of healthy dogs it owned with that of healthy dogs brought to the clinic to be neutered. Here are the results:

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>( \bar{x} )</th>
<th>s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pets</td>
<td>26</td>
<td>193</td>
<td>68</td>
</tr>
<tr>
<td>Clinic</td>
<td>23</td>
<td>174</td>
<td>44</td>
</tr>
</tbody>
</table>

(a) What are the populations of interest? Give at least one source of potential bias in the samples.

Although it isn’t completely clear from the statement of the question, the populations of interest appear to be healthy clinic dogs and healthy pet dogs. Since the sample of pets only involves dogs brought in to be neutered, this sample is probably biased, since the type of dog owner who neuters a pet may feed the pet a slightly different diet than the type of dog owner who does not (for example, neutering costs money, so neutered dogs may come from slightly more wealthy households than average).

(b) Is there strong evidence that pets have a higher mean cholesterol level than clinic dogs? State the null and alternative hypotheses and carry out an appropriate test. Give the \( p \)-value and state your conclusion.

\[
H_0 : \mu_{\text{pet}} - \mu_{\text{clinic}} = 0
\]
\[
H_a : \mu_{\text{pet}} - \mu_{\text{clinic}} > 0
\]
The test statistic will be a t-statistic. Since there are two independent samples whose values cannot be matched in an obvious way, we use the 2-sample t test. The t statistic is therefore
\[
 t = \frac{\bar{x}_{\text{pet}} - \bar{x}_{\text{clinic}}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}. 
\]
We just plug in the values from the table to find that \( t = 1.17 \) on 22 degrees of freedom. We use 22 degrees of freedom because it’s the smaller of \( n_1 - 1 \) and \( n_2 - 1 \). From the table, we see that the 1-sided p-value for our statistic is between 0.10 and 0.15. Thus, we do not reject the null; there is not strong evidence that pets have a higher mean cholesterol level than clinic dogs.

(c) Give a 95% confidence interval for the difference in mean cholesterol levels between pets and clinic dogs.

The formula we use for a CI for difference of two means is
\[
\bar{x}_{\text{pet}} - \bar{x}_{\text{clinic}} \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.
\]
Everything we need is given in the table except \( t^* \), which is 2.074 using 22 df. Plugging in all the values gives a confidence interval of 19 ± 33.6. Note that this interval does contain 0, which shouldn’t be surprising given our answer to (b).

(d) Give a 95% confidence interval for the mean cholesterol level in pets.
Since we don’t know \( \sigma \), the formula for a population mean CI is

\[
\bar{x} \pm t^* \left( \frac{s}{\sqrt{n}} \right).
\]

In this case, we use 25 df since there are 26 pets in the sample. This gives \( t^* = 2.060 \). Plugging in values gives the 95% CI \( 193 \pm 27.5 \).

(e) What assumptions must be satisfied to justify the procedures you used in (a), (b), and (c)? Assuming that the cholesterol measurements have no outliers and are not strongly skewed, what is the chief threat to the validity of the results of the study?

We are assuming that the sampling distribution of \( \bar{x} \) is normal in each of our two samples. We also that the samples are independent of each other and that each is a SRS. Assuming that we don’t have outliers or strong nonnormality, therefore, the biggest threat to the validity of the study is violation of the SRS assumption; in other words, bias is the biggest threat.