Homework 4

Kai Aragaki

Problem 1

The confidence intervals can be calculated using the following formula:

$$(\bar{X} - \bar{Y}) \pm t \cdot \widehat{SD}(\bar{X} - \bar{Y})$$

$(\bar{X} - \bar{Y})$ can be calculated like so:

```
a <- c(84, 106, 99, 101, 100, 99, 127, 105, 101, 108)
n_a <- length(a)
df_a <- n_a - 1
b <- c(56, 62, 67, 81, 69)
n_b <- length(b)
df_b <- n_b - 1
mean_diff <- mean(a) - mean(b)
```

The pooled estimate of our population σ is essentially a weighted mean of our standard deviations:

```
est_sig_pool <- sqrt((var(a) * df_a + var(b) * df_b) / (df_a + df_b))
```

And the estimated standard error is then calculated as

```
est_se <- est_sig_pool * sqrt(1 / n_a + 1 / n_b)
```

Our t distribution should reflect our number of degrees of freedom (13) as well as our desired span of confidence.

For 95%, we choose

```
t_95 <- qt(0.975, df_a + df_b)
```

So our confidence intervals become

```
ll <- mean_diff - t_95 * est_se
ul <- mean_diff + t_95 * est_se
```

The 95% confidence interval $\mu_a - \mu_b$ is (23.85 , 48.15)

The 95% confidence interval can be calculated using a different cutoffs of our t-distribution:

```
t_99 <- qt(0.995, df_a + df_b)
```

Making our confidence intervals:

```
ll <- mean_diff - t_99 * est_se
ul <- mean_diff + t_99 * est_se
```

The 99% confidence interval $\mu_a - \mu_b$ is (19.06 , 52.94)

The p-value for the hypothesis test that the means are the same can be calculated first by first calculating the T value we observe:

```
t_obs <- (mean_diff / est_se)
```

Then, we can find p by using `pt`

```
prob <- 2*pt(t_obs, 13, lower.tail = FALSE)
```

(We multiply by two because we must remember there are two tails to this test)

The probability is roughly 2.3×10^{-5}

Problem 2

Confidence intervals can be expressed with the following formula:

$$\bar{X} \pm \cdot \widehat{SD}(\bar{X})$$

to find $\widehat{SD}(\bar{X})$, we use $\widehat{SD}(\bar{X}) = SD * \sqrt{(1/n)}$

So

```
ll <- 7.0 - qt(0.95, 24) * 3.0 * sqrt(1/25)
ul <- 7.0 + qt(0.95, 24) * 3.0 * sqrt(1/25)
```

Giving the 90% confidence intervals of (5.97, 8.03)

Problem 3

We must take the differences between before and after values:

```
df <- data.frame(mouse = 1:6,
                 before = c(81, 101, 76, 67, 125, 144),
                 after = c(138, 210, 162, 105, 259, 319))
df$diff <- df$after - df$before
```

The question 'does the treatment have an effect' is an ambiguous one, mostly because we cannot answer this definitively. However, we can answer with some degree of certainty, the degree of which we can choose ourselves. I would like to set my 'false positive' rate to be 5% - the standard, for better or worse. Therefore, $\alpha = 0.05$

We can calculate 95% confidence intervals using largely the same equation as in problem 2:

```
ll <- mean(df$diff) - qt(0.975, nrow(df) - 1) * sd(df$diff) * sqrt(1 / nrow(df))
ul <- mean(df$diff) + qt(0.975, nrow(df) - 1) * sd(df$diff) * sqrt(1 / nrow(df))
```

The 95% confidence intervals are then 47, 153

Recall that this confidence interval is for the DIFFERENCES. The fact that it does not pass through zero leads us to reject H_0.

The probability can be calculated as we did in problem 1

```
t_obs <- mean(df$diff) / (sd(df$diff) * sqrt(1 / nrow(df)))
prob <- 2 * pt(t_obs, nrow(df) - 1, lower.tail = FALSE)
```

The probability is roughly 0.0047

Problem 4

A

H_0 : There is not a significant decrease in white blood cell count between humans infected with *E. canis* and the general population of humans.

H_A : *E. canis* infection in humans is associated with a significant decrease of white blood cell count as compared to the general population.

B

We can calculate a T statistic by doing the following:

```
t_obs <- (7250 - 4767) / (3204 / sqrt(15))
```

And from there, we can calculate probabilities:

```
prob <- pt(t_obs, 14, lower.tail = FALSE)
```

The probability of a sample this extreme given the there was no difference in WBC count between the infected and uninfected population is about 0.0048

C

We reject the null hypothesis that there is not a significant difference in WBC count between infected and uninfected individuals.