Task 2: Independent sample t-test

a. The “model” is a statement of our assumptions about the populations from which our data come. Here the model is that the two samples come from two populations which are normally distributed with a common variance \((\sigma^2)\). The population mean parameters can be called either \(\mu_1\) and \(\mu_2\) or \(\mu\) and \(\mu + \delta\). Without specific assumptions, no conclusions can be drawn. If you said “t-test”, that is a type of analysis, not a model.

b. The mean number of tries for the control group is smaller, so they learned faster (performed better).

c. The IQRs in the boxplots look about the same for the 2 groups. The standard deviations for the two groups are 3.20 and 3.37 (using standard errors here is wrong). Based on the rule of thumb (the bigger one is less than twice the smaller one), we don’t need to worry about unequal spread for the two groups.

d. Based on the boxplots, the distributions look reasonably symmetric. (Histograms would show this better.) The quantile normal plot (a special form of a QQ plot) could be used to go beyond symmetry and specifically check for normality. The absolute value of the skewness for the control group is 0.013, which is smaller than 0.944 (twice its standard error) so we have no evidence of skew here. In fact we have no evidence of skew or kurtosis for either group.

e. The null hypothesis is \(H_0: \mu_1 = \mu_2\) or \(H_0: \delta = 0\). The alternative is either \(H_A: \mu_1 \neq \mu_2\) or \(H_A: \delta \neq 0\). (I prefer the two sided alternative here because I am not completely sure that one treatment must be better than the other.)

f. \(\bar{X}_1 - \bar{X}_2 = (15.08 - 17.38) - 0 = -2.30\)

g. \(SS_1 = 10.254*23 = 235.8, \ SS_2 = 11.375*23 = 261.6\). Here we are “undoing” the equation Variance=SS/df to recover the SS values. \(SS_p = 235.8 + 261.6 = 497.4\)
\(df_p = 23 + 23 = 46\). Now we put the combined (pooled) SS and df back together to get a pooled estimate of the data variance: \(MS_p = 497.4/46 = 10.81\). Variances are hard to think about, so we take the square root to get the data standard deviation (3.29). This tells us that a random subject would typically (68% of the time) have an outcome within 3.29 of the true mean outcome for the population corresponding to that subject’s treatment group.

h. The pooled estimate of the variance of either group mean is 10.81/24 = 0.450. This is smaller than the variance of an individual data value (10.81) because sample means are less variable than individual data values.

i. The standard error of one group mean is \(\sqrt{0.450} = 0.671\). The two individually calculated SE values are 0.654 and 0.688. The pooled estimate is in between because it is composed of data from both groups. The standard error of a group mean is an estimate of the standard deviation of group means of size 24 over repeated sampling. So if we randomly select 24 subjects and find the mean outcome, it will typically be within 0.671 of the true mean outcome for the population.

j. The standard error of the difference of the two sample means is \(\sqrt{2(0.450)} = 0.949\). The variance of a difference (or sum) of two statistics is larger than the variance of either one individually. Because the sample sizes are equal and we
assume that the underlying data variances are equal, we just double the variance of one
mean to get the variance of the difference of the two means.

k. The t-statistic is $-2.30/0.949 = -2.42$. This is a useful statistic because it has
a known theoretical sampling distribution under the null hypothesis: the t-distribution
with $n_1 + n_2 - 2$ df. A t-distribution is quite similar to a normal distribution if the number of
degrees of freedom is at least 20. So values between about −2 and +2 are not uncommon
and therefore are consistent with the null hypothesis. The value of the t-statistic for this
experiment is fairly far from typical t-statistic values under the null hypothesis (p=0.020
indicates that values this far from 0 occur only 2% of the time when the null hypothesis is
true. So we reject the null hypothesis and conclude that the population means for the two
treatments are different from each other.

**Task 3: Link from independent t-test to ANOVA**

The square of the t-statistic is $(−2.42)^2 = 5.86$ which equals the F statistic in the
ANOVA except for some rounding error. The p-values are also equal. We
really don’t need to learn a separate t-test. We can use ANOVA for 2 or
more groups and get the same answers. Why this is true is rather complicated
mathematically.

The Levene’s Test is a separate test of the hypothesis of equal variances for
the two treatment groups. It gives an F value which is compared to the appropriate F
distribution to get the p-value of 0.668. Apparently an F statistic of 0.187 is not too
uncommon when the variances of the two groups are equal because more extreme values
occur 66.8% of the time when the variances are equal. Now that we know that we have
no good reason to reject the null hypothesis that the two variances are equal (as we
assume for our model), we feel more confident that the p-value for the t-test is
meaningful.

**Task 4 ANOVA**

a.

b.

c. To get group statistical comparisons, look at stem and leaf plots and side-
by-side boxplots.
It looks like tutor is the worst treatment (highest median minutes). The equal spread assumption looks OK. Although we can’t check normality with boxplots, at least the distributions appear symmetric.

d. The ANOVA shows $p<0.0005$, so we reject the null hypothesis that all 3 population treatment means are equal. The large F statistic (much larger that 1) was compared to the theoretical $F_{2,18}$ distribution (under the null hypothesis) and found to be very unusual, resulting in that $p$-value. The denominator of F is the within groups estimate of the model $\sigma^2$ value, 242.333, which is unaffected by the group means. The numerator is the between groups MS value which appears highly inflated (much larger than 242) causing a large F and suggesting that the group population means are different.

**ANOVA**

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>5978.000</td>
<td>2</td>
<td>2989.000</td>
<td>12.334</td>
<td>.000</td>
</tr>
<tr>
<td>Within Groups</td>
<td>4362.000</td>
<td>18</td>
<td>242.333</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>10340.000</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

e. The grand mean column has only one value, the mean of all of the (outcome) data. The group mean column has 3 different values, which represent the appropriate group mean for each of the subjects. The deviation from the grand mean tells how far each value is from the overall mean (ignoring group). The deviation from the group means tells how far each subject is from the mean of all of the subjects in the same group. The
The group deviation column has only three numbers and tells how far each group mean is from the grand mean. The deviation columns have positive and negative numbers because some numbers must be higher than the mean and some lower for any mean. The corresponding square columns are all positive because the square of a negative number is positive.

g. The sum of the group mean column is the same as the sum of the grand mean column because whether you equalize the outcome for everyone together or separately within the groups, the total outcome is the same.

h. E.g., for “within”, MS=SS/df=4362/18=242.3. F=2989/242.3=12.3.

i. MS_{within} is a measure of the variance within groups (\(\sigma^2\)) and is not affected by unequal population means because it only uses deviations from the group means to estimate the underlying variance. When the null hypothesis is true, MS_{between} is another estimate of (\(\sigma^2\)) which is independent of MS_{within}. Therefore, when the null hypothesis is true, the F ratio, MS_{between}/MS_{within} tends to be around 1. Because the group means are similar only if the null hypothesis is true, and MS_{between} is based on deviations of the group means from the grand mean, MS_{between} is inflated when the null hypothesis is false. In that case the F value is large, as in this problem. When F is large, (and hence the p-value is small), we can conclude that, if the assumptions are (approximately) correct, the null hypothesis is extremely unlikely. So here we have good evidence that at least one of the treatments affects the outcome differently from the others. (We suspect that tutor is bad and that standard and practice are about equal.)