P1 Gambling strategy

a. Let $p$ denote the true winning odds of the strategy recommended by the newsletter. The question basically asks to test whether $p$ is greater than 0.5. The null and alternative hypotheses are

$$H_0 : p = 0.5, \quad H_a : p > 0.5.$$ 

The estimated winning probability is $\hat{p} = \frac{30}{50} = 0.6$ and the test statistic is

$$z^* = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} = \frac{0.6 - 0.5}{\sqrt{0.5 \times 0.5/50}} = 1.414.$$ 

The rejection region is $RR = (Z > Z_{a}) = (Z > Z_{0.05}) = (Z > 1.645)$ or the p-value is $P(Z > z^*) = P(Z > 1.414) = 1 - P(Z < 1.414) = 1 - 0.921 = 0.079$. Since $z^*$ falls within the rejection region (or p-value < 0.05), we fail to reject null hypothesis. There is no sufficient evidence to conclude that the newsletter’s strategy has a significantly higher winning odds than random selection given the available data at a 0.05 of significance.

b. Type II error of this test is that one fails to conclude that the newsletter’s winning strategy is significantly better than random selection (i.e., $p > 0.5$) when it is actually the case.

General relationship between the probabilities of type I and II errors: the two probabilities can not both be controlled at their minimum values for a fixed sample size and if we decrease one probability, the other will increase. However, we can decrease the two probabilities at the same time by increasing the sample size.

c. The power of the test is 0.396 when the true winning probability is 0.6. Interpretation: the probability of making a correct decision of rejecting the null hypothesis $p = .5$ and concluding the alternative $p > .5$ for the z-test with sample size $n = 50$ and $\alpha = 0.05$ is 0.396. If you conduct such z tests many times, about 39.6% times you are making a correct decision of rejecting null hypothesis.
d. When increasing the number of games from 50 to 100, the power of the test will increase. When we decrease the level of significance from 0.05 to 0.01, the power will decrease (it is more difficult to reject null hypothesis).
P2 SAT scores

a. Let $\mu_1$ and $\mu_2$ denote the mean verbal SAT scores for high school students who intend to major in engineering and in language (or literature), respectively. We want to test

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 - \mu_2 \neq 0$$

and we can use a two-sample t test since the two samples of verbal scores are independent and the sample sizes are small. Assume the two populations have the same variances. The pooled variance is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(15 - 1) \cdot 42^2 + (15 - 1) \cdot 45^2}{15 + 15 - 2} = 1894.5.$$ 

The test statistic is

$$t^* = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{s_p^2 (\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{446 - 534}{\sqrt{1894.5 \cdot (1/15 + 1/15)}} = -5.54.$$ 

The rejection region is $RR = (|t| > t_{a/2, n_1 + n_2 - 2}) = (|t| > t_{0.025,28}) = (|t| > 2.048)$. Since $t^* \in RR$, we reject the null hypothesis and conclude that the two groups indeed have different mean verbal SAT scores.

b. We can use a 2-sample F test for checking the equal variance assumption. For this test, we assume: (i) two independent random samples of verbal SAT scores and (ii) normality for each sample of verbal SAT scores.

c. We need to a paired t-test since the verbal and math scores are from the same persons and they are correlated. Recall that the test statistic is

$$t = \frac{\bar{x}_d - 0}{s_d / \sqrt{n_d}},$$

where $\bar{x}_d$, $s_d$, and $n_d$ are the sample mean, standard deviation, and sample size of the differences of the paired observations, respectively. We can obtain $\bar{x}_d$ and $n_d$ but cannot get $s_d$ using the summary statistics provided in the table. So the answer is No. We cannot use the summary results from the table to conduct the paired t-test because we cannot obtain the sample standard deviation of the differences, which requires to use the raw data.
P3 Fire damage

a. The simple linear regression model seems to provide a good fit for this data. The scatterplot shows a clear linear pattern between the response (damage) and the predictor (distance). The estimated $R^2$ is 0.9235, which means that 92.35% of the variation of the response variable is explained by the linear relationship between the response variable (damage) and the predictor (distance). The model diagnosis plots indicate a good model fit as specifically explained in question (b).

b. The diagnostic plots suggest that the assumptions required for simple linear regression hold. Specifically, the plot of residual vs. fitted values shows a random pattern of residuals centering around 0 and a roughly equal variance. The Q-Q plot suggests no serious violation of normality assumption of the random errors. The plot of residuals vs. order does not show a serious violation of independence. There is another assumption in simple linear regression: the random error distribution is independent of the predictor. Since the plot of residuals vs. predictor (distance) is not given, we do not know if this assumption holds.

c. The estimated slope is $\hat{\beta}_1 = 4.92$. Interpretation: for every mile further away from its nearest fire station, the damage is expected to increase by 4.92 thousands of dollars.

d. The estimated intercept is $\hat{\beta}_0 = 10.28$. Usually the slope can be interpreted as the expected response when $x = 0$. This interpretation is not practical in this case as $x$ taking 0 means the house is in fire station.

e. The estimated correlation is $\hat{\rho} = \sqrt{0.9235} = 0.961$. Note that $\rho = 0$ implies no linear correlation and the slope parameter $\beta_1 = 0$ in the simple linear regression model. For this reason, testing $H_0 : \rho = 0$ vs. $H_a : \rho \neq 0$ is equivalent to testing $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ here. The p-value of this test is 0 from the output. This suggests that one needs to reject null hypothesis and claim that $\rho$ is not equal to 0.
Particle processing

a. We expect mean production to increase by 45 parcels when worker 3 is present, as compared to absent, controlling presence/absence of other workers.

b. The intercept corresponds to all $x$’s = 0, so all workers absent, which means only the foreman is working. We expect the foreman to produce about 104 parcels when working by himself / herself.

c. $H_0 : \beta_2 = 0$ vs. $H_a : \beta_2 > 0$ ("productive" would correspond to a one-sided hypothesis). Test statistic $t = -0.28$ does not support $H_a$; the one-sided P-value is thus $1 - (1 - .7858)/2 = 1 - .1071 = 0.8929$. There is not strong evidence in this data that worker 2 is productive. [Teachable moment: be careful about, e.g., firing worker 3 based on this model does not include interactions, for example. Perhaps worker 3 does not get along with worker 5, so the coefficient of $x_3 x_5$ would be negative and significant if this term was in the model. Hence worker 3 may be very productive if worker 5 is not present. At any rate - it’s only a model. Take a closer look at the reality of the situation when you see something like this.]

d. This corresponds to $x_3 = 1$ and all other $x$’s = 0, so predicted $y = b_0 + b_3(1) = 104.23 + 44.65 = 148.88$. The approximate 95% margin of error for this prediction is $2(\text{RootMSE}) = 2(3.29743) = 6.6$.

e. Since worker 1 and worker 5 were always present or absent at the same time, including both $x_5$ and $x_1$ in the model would create a perfect collinearity. The coefficient 101.2 for “$x_1$” thus represents the joint efforts of both workers in this model.

f. No - the largest VIF is 3.28. we do not consider collinearity to be strong unless some VIFs are greater than 10.
P5 Fly repellents

a. Completely Randomized design (CRD).

b. Observations from each treatment have a long-run average that depends on the treatment, but the a common variance between all treatments. All observations are mutually independent and ideally Normally distributed. Based on these boxplots, the assumption of equal variances seems shaky (seems like possibly smaller variability for treatment 1 - the control - and/or treatment 4); I’d do some tests for homoscedasticity (e.g. Levine’s and/or Bartlett’s test). Also, normality is technically false given that the y’s are whole numbers between 0 and 20, but that assumption is the least important for comparing long-run means between the treatments.

c. This $F = 13.55$ test statistic has 6 and $49 - 7 = 42$ df. The 99th percentile of the $F$ distribution with 6 and 40 df is 3.29 (see the last table). Hence, the P-value < 0.01. There is very strong evidence that the long-run average number of repelled flies is not the same for these 7 treatments.

d. (i) The ANOVA F-test (when it rejects) does not allow us to say which treatment long-run-means differ from which others; when it does not reject it doesn’t allow us to say that they are all equal, or even nearly equal; (ii) Dunnett’s method (using $= 0.05$ for example) will allow us to compare the control’s long-run-mean with each of the six other treatment means controlling the family-wise error (i.e. the probability of at least one false declaration of difference in any of the 6 tests), but we cannot use it to compare (e.g.) the treatment 4 long-run-mean with the treatment 2 long-run-mean; (iii) Tukey’s method allows us to compare each treatment’s long-run-mean with each other (a total of $7 \times 6/2 = 21$ tests of confidence intervals) controlling the family-wise error. The cost for these extra comparisons will be reduced power (or wider conf. intervals) for the comparisons that Dunnett’s performs.

e. This is a classic post-hoc comparison - a comparison suggested by the data. The Scheffé multiple comparison method is recommended in this case for family-wise error rate protection.
P6 Salary equity

a. The unequal-slopes ANCOVA assumes a separate linear regression model (with common error variance) for each treatment. The relationship of untransformed $Y$ to $X$ appears to be nonlinear, especially for the males. The relationship is much more linear in the transformed response. This makes sense because salaries tend to grow exponentially over time (lately, with a rather small exponent).

b. If slopes were unequal it would mean (in lay terms) that the size of the ”gender gap” is different depending on years of experience. For example, it would be possible that there would be a sizeable difference in adjusted mean (log) salary for male vs female employees who have 20 years’ experience, but no difference at all for employees who have 5 years’ experience.

c. Note that the estimated adjusted mean for females is 4.927 and for males is 4.903; so in this (made-up!) data, the women actually make a little more than the men on average, adjusting for experience. The difference is statistically significant ($p = 0.0438$). So, if one does a two-sided test we’d have to say that ”there is strong evidence that the long-run mean (log) salary for men differs from that of women, adjusting for years of experience” which is unacceptably vague; if one does a one-sided test ”by the book”, the Ha chosen in advance would probably be $H_a : W < M$, and the data do not support this Ha, so the P-value would be insignificant. If one uses a confidence interval, the difference in long-run mean adjusted log-salary, in the order F-M, is estimated at 0.024035, and a 95\% interval is (0.000684, 0.047387). If we backtransform this by raising base 10, it gives an estimate $10^{0.024035} = 1.057$ for the ratio (long-run mean adjusted salary, F)/(long-run mean adjusted salary, M), which would say that we expect females to have higher long-run mean salary by about 5.7 percent adjusting for years of experience. The confidence interval backtransforms to (0.000158, 1.115): with 95\% confidence the women have at worst only a smidgen higher long-run mean adjusted salary, but their advantage could be as high as 11.5\%. Again, just made-up data though!