Q1

(a) We wish to create a 95% confidence interval. Given the information for the athletes and non-athletes, and the fact that we $df = 10$ and $t = 2.2622$ for a 95% CI (from t-tables with $df - 1 = 9$), we have that – for CI_1 and CI_2 being for the athletes and non-athletes respectively –

$$\begin{aligned} CI_1 &= \left(\bar{y} - 2.2622\frac{s}{\sqrt{10}}, \bar{y} + 2.2622\frac{s}{\sqrt{10}}\right) \\ &= \left(76.59 - 2.2622\frac{10.85}{\sqrt{10}}, 76.59 + 2.2622\frac{10.85}{\sqrt{10}}\right) \\ &= (68.828, 84.352) \end{aligned}$$

$$\begin{aligned} CI_2 &= \left(\bar{y} - 2.2622\frac{s}{\sqrt{10}}, \bar{y} + 2.2622\frac{s}{\sqrt{10}}\right) \\ &= \left(71.44 - 2.2622\frac{10.51}{\sqrt{10}}, 71.44 + 2.2622\frac{10.51}{\sqrt{10}}\right) \\ &= (63.921, 78.959) \end{aligned}$$

There would appear to be a slight difference in the mean, though, the intervals overlap quite a bit indicating that the mean for both lie in roughly the same interval.

(b) Here, for the standard deviation, we use the χ^2 distribution. Using a table, we find that with $d.f = 10-1 = 9$, the 0.025 and the 0.975 values are $a = 2.700$ and $b = 19.023$ respectively. Thus, we construct the intervals as follows (for CI_1 and CI_2 still representing the athletes and non-athletes respectively)

$$\begin{aligned} CI_1 &= \left(\sqrt{\frac{(n-1)s^2}{b}}, \sqrt{\frac{(n-1)s^2}{a}}\right) \\ &= \left(\sqrt{\frac{9(10.85)^2}{19.023}}, \sqrt{\frac{9(10.85)^2}{2.700}}\right) \\ &= (7.463, 19.809) \end{aligned}$$

$$\begin{aligned} CI_2 &= \left(\sqrt{\frac{(n-1)s^2}{b}}, \sqrt{\frac{(n-1)s^2}{a}}\right) \\ &= \left(\sqrt{\frac{9(10.51)^2}{19.023}}, \sqrt{\frac{9(10.51)^2}{2.700}}\right) \\ &= (7.229, 19.189) \end{aligned}$$

There doesn't appear to be a stark difference in the standard deviations (i.e. they could still be different but the intervals, as they stand at this sample size, still overlap significantly).

(c) Should we increase n to 150, we would notice narrower interval for both the mean and the standard deviation. This is due to the fact that, for the mean, the term we add and subtract in proportional to $1/\sqrt{n}$, thereby making the interval smaller and smaller, and likewise for the standard deviation. With these narrower intervals, we could infer a much more concise conclusion on whether the mean and standard deviation are different because the interval would either overlap less, or more.

Q2

(a) The null and alternative hypothesis are

$$H_0 = \text{proportion of spenders has not increase (i.e. } p = 0.15)$$
$$H_1 = \text{proportion of spenders has increase (i.e. } p > 0.15)$$

Since the alternative hypothesis only concerns itself with an increase in the percentage, this is a one-tailed alternative hypothesis.

(b) The standard deviation, with $\theta = 0.15$ and $n = 2500$, is simply

$$\sigma_{\hat{p}} = \sqrt{\frac{\theta(\theta - 1)}{n}} = \sqrt{\frac{0.15(0.85)}{2500}} \approx 0.007$$

(c) Since n is very large, we can use the normal approximation, and the following test statistic

$$z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}}$$

where $\hat{p} = 0.1804$, and $p_0 = 0.15$. Thus, this simplifies to

$$z_0 = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} = \frac{0.1804 - 0.15}{0.007} \approx 4.343$$

This follows a standard normal distribution under the null hypothesis.

(d) Since we are okay with a false rejection of the null hypothesis with only probability of 0.001, we must compare what we found above in (c) to the z-value corresponding to the 99.9th percentile. From a table (right tail) we note that the critical value is $z_{0.001} = 3.00$.

(e) Since the calculate test statistic is greater than the critical statistic, we reject the null hypothesis, and thus, there is sufficient evidence to say that the spending rate has increase since the update.

Q3

(a) No, the distribution is nothing like a normal; very heavy tail and completely skewed.

(b) Yes, it's reasonable by CLT. Since our sample size is quite high, we know that by CLT, the sample mean approximately follows a normal distribution.

(c) The Null and Alternative hypothesis are, respectively,

$$\begin{aligned} H_0 &= \text{average amount spent per user (i.e. } \mu = \$3.00) \\ H_1 &= \text{average amount spent per user (i.e. } \mu \neq \$3.00) \end{aligned}$$

Since the alternative hypothesis is two sided (as the amount can either change to be higher, or lower), it is two-tailed.

(d) The estimated standard deviation of the estimator of the sample mean is simply

$$\sigma_{\overline{y}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{s}}{\sqrt{2500}} = \frac{\sqrt{70.78}}{50} \approx 0.168$$

(e) The relevant test statistic formula follows a t-distribution with $d.f. - 1 = 2499$ degrees of freedom.

$$t = \frac{\overline{y} - \mu_0}{\sigma_{\overline{y}}}$$

where $\overline{y}$ is the sample mean, and the two other quantities are stated / derived above. Plugging in these numbers, we get the observed value of the test statistic is

$$t = \frac{\overline{y} - \mu_0}{\sigma_{\overline{y}}} = \frac{3.54 - 3.00}{0.168} \approx 3.21$$

(f) If we are willing to accept a false rejection of the null with only probability 0.001, we look for a z-value for $0.001/2 = 0.0005$ (since this is two-tailed), which is $z = \pm 3.29$

(g) Since we have that the observed test statistic is less than the critical t-statistic, we do not reject the null hypothesis; there is not enough evidence to conclude that the spending has changed.

Q4

(a) From the table we see that the associated p-value is (approximately) $p = 1 - 0.968 = 0.032$. Using the R code gives us 0.03175577

(b) From the table we find that (since this is now a two tailed test), that the p value is $p = 2(1 - 0.968) = 0.064$, and the code gives us exactly that, 0.06351155

(c) We must first find the z score, the sample proportion and SD are

$$\hat{\hat{p}} = \frac{50}{70} \approx 0.714$$
$$\sigma_{\hat{p}} = \sqrt{\frac{(0.6)(1-0.6)}{70}} \approx 0.0590$$

Thus the associated z score is

$$z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} = \frac{0.714 - 0.6}{0.0590} \approx 1.9322$$

With this z-score, we have a p-value of $1 - (0.5 + 0.4732) = 0.0268$; which is awfully close to what the R code gives us, 0.02588749

(d) Following the exact same steps as before, but now since this is a two tailed test, we'd have from the above z-score, a p-value of $2(1 - (0.5 + 0.4732)) = 0.0536$. This is similar to what the R code would give us which is 0.05177498

(e) The likely hood ratio statistic is given by

$$\Lambda(\hat{\theta}) = -2 \log \left[\left(\frac{\theta_0}{\hat{\theta}} \right)^y \left(\frac{1 - \theta_0}{1 - \hat{\theta}} \right)^{n-y} \right]$$

Using the fact that (for a binomial RV), $\hat{\theta} = y/n = 50/70$, and $\theta_0 = 0.6$, we have that

$$\Lambda(\hat{\theta}) = -2 \log \left[\left(\frac{0.6}{50/70} \right)^{50} \left(\frac{1 - 0.6}{1 - 50/70} \right)^{70-50} \right] \approx 3.976$$

Looking at the table, we have that the p-value lies in the interval $[0.050, 0.025]$ (as a sanity check, one can use R to find that the p-value does in fact lie in the interval, as it is 0.04614)

(f) Note that, under the null hypothesis, the chi-squared test statistic gives us

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(19)(130)}{100} \approx 24.7$$

Looking at the given tables, we find that the p-value lies in the interval $[0.100, 0.250]$ (one can write the code for this and find that the actual p-value is 0.170)