Analysis of Scientific Data – Summer Semester, 2020

Short Answers

Scenario 1 - Hypertension

1. The standard error of the difference in the means is $\sqrt{\frac{5.08^2}{11}+\frac{1.97^2}{13}} = \mathbf{1.626213\ mmHg}$.
2. The t statistic is $\frac{(\bar{x}_1-\bar{x}_2)-0}{\text{se}(\bar{x}_1-\bar{x}_2)} = \frac{(145.0-142.3)-0}{1.6262} = \mathbf{1.6603}$.
3. The minimum degrees of freedom are $\min(11-1, 13-1) = \mathbf{10}$.
4. The p value is `1-pt(1.6603,df=10)` = **0.06392**, giving **weak** evidence to suggest that the mean systolic blood pressure is lower with the new drug.
5. The margin of error is $t^*\text{se}(\bar{x}) = 2.228139 \times 1.626213 = \mathbf{3.623427\ mmHg}$. The value of t^* comes from `qt(.975,df=10)`.
6. The pooled variance is $s_p^2 = \frac{(11-1)5.08^2+(13-1)1.97^2}{(11-1)+(13-1)} = 13.84704$, $s_p = \mathbf{3.721161\ mmHg}$.
7. The t statistic is $\frac{(\bar{x}_1-\bar{x}_2)-0}{3.721161\sqrt{\frac{1}{11}+\frac{1}{13}}} = \frac{(145.0-142.3)-0}{1.52446} = \mathbf{1.771119}$.
8. The p value is `1-pt(1.771119,df=22)` = **0.04519788**, giving **moderate** evidence to suggest that the mean systolic blood pressure is lower with the new drug.
9. Although this is stronger evidence, the sample standard deviations are quite different, so you would recommend that they use the result from Question 4 since **the assumption of common standard deviations does not seem appropriate**.
10. The expected value is $E(W) = \frac{n_1(n_1+n_2+1)}{2} = 13 \times \frac{25}{2} = \mathbf{162.5}$. Note that we defined n_1 as the sample size whose ranks are summing (that is n_1 = sample size of new drug group). Then n_2 is the sample size of the other group (placebo group here)
11. The standard deviation is $\text{sd}(W) = \sqrt{\frac{n_1n_2(n_1+n_2+1)}{12}} = \sqrt{297.92} = \mathbf{17.26}$.
12. The p value is

$$P(W \le 136) \approx P\left(Z \le \frac{136-162.5}{17.26}\right) = P(Z \le -1.535) = 0.062,$$

using `pnorm(-1.535)`. Thus, there is **weak** evidence to suggest that the systolic blood pressure distribution tends to be lower with the new drug.

13. For evidence at the 1% level, we want $P(Z \le z^*) = 0.01$. `qnorm(.01)` gives $z^* = -2.326$. Thus $W = 162.5 - 2.326 \times 17.26 = 122.35$, so $W = \mathbf{122}$ would give evidence at the 1% level.

Scenario 2 – Blood Alcohol Content

1. `aggregate(BAC ~ Sex, alcohol, mean)` gives the mean **0.071 g/dL** for females.
2. `aggregate(BAC ~ Group, alcohol, mean)` gives the mean difference 0.084 – 0.050 = **0.034 g/dL** between High and Low .
3. `summary(aov(BAC ~ Group, alcohol))` gives the ANOVA table, showing that the total sum of squares is 0.005067 + 0.009123 = **0.01419**.
4. From the ANOVA table, the degrees of freedom are **2,18**.
5. From the ANOVA table, p value is **0.019**, giving **moderate** evidence to suggest that there is a difference in mean blood alcohol concentration between the different dosage groups
6. The R^2 value is $\frac{0.005067}{0.01419} = \mathbf{0.36}$.
7. With 3 groups, there are 3 pairwise comparisons to make. This can be calculated using `choose(3,2)` in R.
8. `TukeyHSD(aov(BAC ~ Group, alcohol))` gives the interval as **(0.0036,0.0650) g/dL**.
9. Since the 95% interval does not contain 0, there is **evidence at the 5% level** of a difference in mean blood alcohol concentration between **High and Low dosage groups**.
10. The null hypothesis is $\boldsymbol{H_0: \rho = 0}$. To test for a negative association, we have $\boldsymbol{H_1: \rho < 0}$.
11. `cor(alcohol$BAC,alcohol$Mass)` gives $\boldsymbol{r = -0.215}$. This a **negative** relationship, but relatively **weak**.
12. Since $H_1: \rho < 0$, the p value is `pt(-0.957,df=19)` = **0.1753**.
13. The multiple regression model is `lm(BAC ~ Mass + Drinks, alcohol)`, giving a coefficient of **0.0128** g/dL/drinks for the `Drinks` variable.
14. Estimated BAC is $0.0417802 - 0.0002372 \times 70 + 0.0128376 \times 5.2 = \mathbf{0.092}$ g/dL.
15. Using `summary(lm(BAC ~ Mass + Drinks, alcohol))`, the p value for `Mass` is 0.62583, giving **no evidence** to suggest that there is an association between blood alcohol concentration and body mass, after taking into account the number of drinks ($p > 0.1$).
16. Margin of error is $t^* \text{ se}(b_1) = 2.1009 \times 0.0004782 = \mathbf{0.00100}$ g/dL/kg, where t^* comes from `qt(.975,df=18)`.
17. `plot(lm(BAC ~ Mass + Drinks, alcohol), which=2)` shows no systematic deviations from the line, so the **normality of residuals assumption is appropriate**.

Scenario 3 – Tomato Quality

1. `prop.table(table(toms$Insects))` gives the proportion 0.4167 for 'Yes'.
2. `table(toms$Insects)` and `table(toms$Irrigation)` give the marginal counts. The expected value is $\frac{15\times12}{36} = \mathbf{5.0}$.
3. `chisq.test(table(toms$Irrigation,toms$Insects))` gives the statistic 6.17.
4. The p value from the output is 0.046, giving moderate evidence to suggest that there is an association between insect attraction and irrigation group.
5. `t.test(Mass ~ Insects, toms)` shows that R is going to subtract No – Yes in its calculation of the difference. Since we are trying to show the mean for the 'Yes' group is lower, we would expect the No – Yes difference to be positive. Thus, the one-sided p value comes from `t.test(Mass ~ Insects, toms, alternative="greater")`, a value of **0.643**, no evidence to suggest that mean fruit mass is lower for pots with insects attracted to them. (In fact, the mean was higher for the pots with insects.)
6. The Welch t-test only assumes **Normal variability**.
7. This is a **randomised complete block design**.
8. `aggregate(Mass ~ Biochar, toms, mean)` gives the difference as $1362.1 - 1181.7 = \mathbf{180.4}$ g.
9. `summary(aov(VitC ~ Biochar*Irrigation, toms))` gives the p value **0.0636**.
10. The p value for the interaction effect is 0.00215, giving **strong evidence** to suggest that there is an interaction between biochar level and irrigation regimes on mean vitamin C content.
11. From the ANOVA table, the degrees of freedom for the interaction test are **2,30**.
12. The R^2 value is $1 - \frac{436.0}{404.0+87.9+220.6+436.0} = \mathbf{0.62}$, so 62% of variability in Vitamin C content is explained by the combination of biochar level and irrigation regime.
13. The estimate from an ANOVA model is the sample mean for the combination of factors. Here `aggregate(VitC ~ Biochar*Irrigation, toms, mean)` gives **17.20** mg/100g as the mean for 5% biochar and PRD irrigation.

Scenario 4 – Soft Drinks

1. We have $\boldsymbol{H_0: p = 0.5}$. If we are trying to show that people prefer brand A, then we test $\boldsymbol{H_1: p > 0.5}$.
2. `sum(dbinom(13:18,18,0.5))` is 0.0481, so observing $\boldsymbol{x = 13}$ participants preferring brand A would be evidence against H_0 at the 5% level.
3. `sum(dbinom(13:18,18,0.76))` gives **0.7512**, the probability of observing at least the critical value of 13 if the proportion is actually 76%.
4. Question 3 gives the power of the test, so the probability of a Type II error is $1 - 0.7512 = \boldsymbol{0.2488}$. You could also calculate it directly as `sum(dbinom(0:12,18,0.76))`.
5. The sample size must be more than 18, since the power was only 75% there. For a sample size of 28, `sum(dbinom(19:28,28,0.5))` shows that the critical value is $x = 19$, so the power is `sum(dbinom(19:28,28,0.76))` = 0.8881, still not 90%. For a sample size of 30, `sum(dbinom(20:30,30,0.5))` shows that the critical value is $x = 20$, so the power is `sum(dbinom(20:30,30,0.76))` = 0.9167, so **30** is the required sample size for a power of 90%, if the true proportion is 76%.
6. The required sample size is $n = \left(\frac{1.96}{0.02}\right)^2 \times 0.5 \times 0.5 = \boldsymbol{2401}$.
7. **Increasing the margin of error** would reduce the required sample size.