S520 Homework 9

Enrique Areyan
March 30, 2012

11.4.#B-1: 2-sample location problem.

(a) The experimental unit is a student.

(b) The experimental units belong to one of two populations:

i. Calculus students who belong to a sorority or fraternity at William & Mary.

ii. Calculus students who do not belong to a sorority or fraternity at William & Mary.

(c) One measurement (score on quiz) is taken on each experimental unit.

(d) Let X_i denote score on the quiz for sorority or fraternity student i.
Let Y_j denote score on the quiz for student j who is not on a sorority or fraternity.
Then, $X_1, X_2, ..., X_{n1} \sim P_1; Y_1, Y_2, ..., Y_{n2} \sim P_2$.
We are interested on drawing inferences about $\Delta = \mu_1 - \mu_2$

(e) $\Delta > 0$ iff $\mu_1 > \mu_2$. Higher score on the test suggest better understanding of the subject matter. Thus, to test the theory in favor of sorority or fraternity students we might want to test $H_0 : \Delta \leq 0$ vs. $H_1 : \Delta > 0$

11.4.#B-2: 2-sample location problem

(a) The experimental unit is an elderly men, defined to be more than 70 years of age.

(b) The experimental units belong to one of two populations:

i. Single, elderly men who own dogs.

ii. Single, elderly men who do not own dogs.

(c) One measurement (score on the Hamilton instrument) is taken on each experimental unit.

(d) Let X_i denote score for dog owner i.
Let Y_j denote score for man j who do not own a dog.
Then, $X_1, X_2, ..., X_{n1} \sim P_1; Y_1, Y_2, ..., Y_{n2} \sim P_2$.
We are interested on drawing inferences about $\Delta = \mu_1 - \mu_2$

(e) $\Delta < 0$ iff $\mu_1 < \mu_2$. Lower scores suggest lack or low levels of depression. Thus, to test the theory in favor of dog owners we might want to test $H_0 : \Delta \geq 0$ vs. $H_1 : \Delta < 0$

11.4.#B-3: 1-sample location problem

(a) The experimental unit is a person.

(b) The experimental units belong to one population, i.e., aerobic students.

(c) Two measurements were taken on each experimental unit:

i. Number of watts expended during protocol S (30-minute ride on the first week)

ii. Number of watts expended during protocol D (30-minute ride on the second week)

(d) Let S_i be the score on protocol S for student i, and let D_i denote score on protocol D for student i.
Then, $X_i = D_i - S_i$ is the random variable of interest. We are interested on drawing inferences about μ

(e) $\mu > 0$ iff $D_i > S_i$. Thus, to test the theory in favor of dynamic streches we might want to test $H_0 : \mu \leq 0$ vs. $H_1 : \mu > 0$

11.4.#B-4: 2-sample location problem

(a) The experimental unit is a tennis ball.

(b) The experimental units belong to one of two populations:

i. Championship balls.

ii. Practice balls.

(c) One measurement (height of the first bounce from 2 meters) is taken on each experimental unit.

(d) Let X_i denote the height of the bounce for practice ball i.
Let Y_j denote the height of the bounce for championship ball j.
Then, $X_1, X_2, ..., X_{n1} \sim P_1; Y_1, Y_2, ..., Y_{n2} \sim P_2$.
We are interested on drawing inferences about $\Delta = \mu_1 - \mu_2$

(e) $\Delta < 0$ iff $\mu_1 < \mu_2$. Thus, to test the theory that practice balls do not wear as well, i.e., lose their bounce more quickly than championship balls, we might want to test $H_0 : \Delta \geq 0$ vs. $H_1 : \Delta < 0$

11.4.#C-1: 2-sample location problem

(a) The experimental unit is a middle-aged man.

(b) The experimental units belong to one of two populations:

i. Type A heavy men.

ii. Type B heavy men.

(c) One measurement (cholesterol level) were taken on each experimental unit.

(d) Let X_i denote the cholesterol level for man i (Type A).
Let Y_j denote the cholesterol level for man j (Type B).
Then, $X_1, X_2, ..., X_{n1} \sim P_1; Y_1, Y_2, ..., Y_{n2} \sim P_2$.
We are interested on drawing inferences about $\Delta = \mu_1 - \mu_2$

(e) $\Delta > 0$ iff $\mu_1 > \mu_2$. Thus, to document that Type A have higher cholesterol than Type B, we might want to test $H_0 : \Delta \leq 0$ vs. $H_1 : \Delta > 0$

11.4.#C-2: Both normal probabilities plots for Type A and B suggest that some values may be inconsistent with the normality assumption, especially the largest values of each set (which can be easily seen in a boxplot). Moreover, for type A, the $IQR(x)/sqrt(var(x)) = 0.9552842$, for type B $IQR(y)/sqrt(var(y)) = 1.282584$. Type B ratio suggest a sample more close to a normal distribution, but also possesses the largest outliers which may be in contradiction with the symmetric and hence, normal assumption. In short, there are too many outliers compared to a typical normal distribution. It seems that it is somewhat unlikely (although not impossible) that the data was drawn from a normal distribution. I would not assume it to be normal distributed.

11.4.#C-3: (a). We want to calculate the following:

$$T_W = \frac{\hat{\Delta} - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

From the data we know the following:

$$\hat{\Delta} = 34.75;\ \Delta_0 = 0;\ S_1^2 = 1342.366;\ S_2^2 = 2336.747;\ n_1 = n_2 = 20. \text{ Thus,}$$

$$t_W = \frac{34.75 - 0}{\sqrt{1342.366^2/20 + 2336.747^2/20}} = 2.5621$$

Now we need to calculate the degrees of freedom for the Welch's approximate t-test:

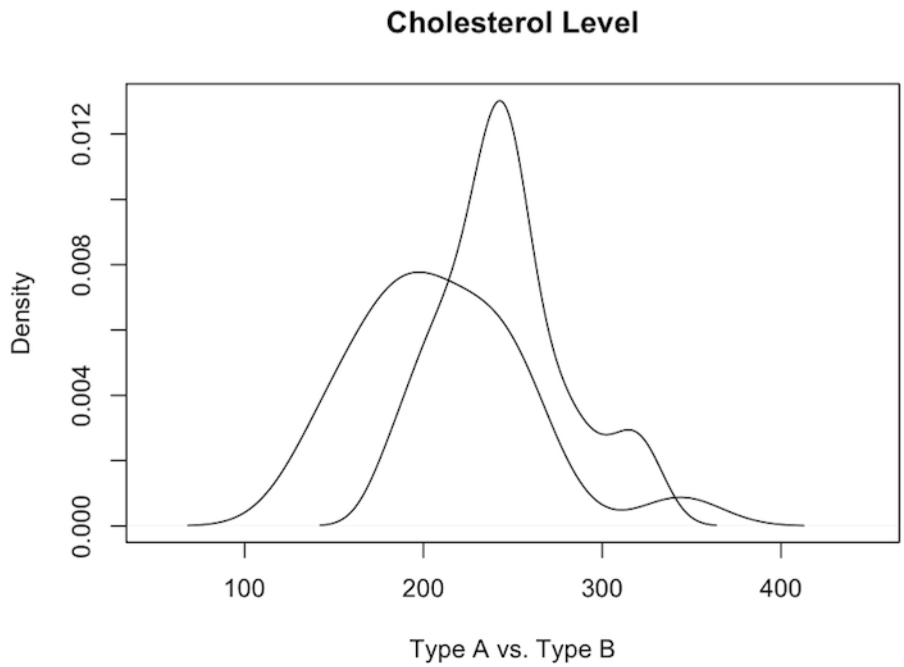
$$\hat{\nu} = \frac{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})^2}{\frac{(S_1^2/n1^2)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}} = 35.4131$$

If we let $\alpha = 0.05$, then $\mathbf{p} = P_{\Delta_0} = P(T_W \geq t_w) = 1 - pt(2.5621, 35.4131) = 0.0074 < 0.05 = \alpha \Longrightarrow$ reject H_0

(b). We want a 90% confidence interval for Δ, thus, let $qt = qt(0.95, 35.4131) = 1.6890$. Then,

$$\hat{\Delta} \pm q_t \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} = 34.75 \pm 1.6890 * 13.5630 = (11.8421, 57.6579)$$

11.4.#C-4:



From the graph of Kernel densities above it is clear that P_1 and P_2 do not belong to a family shift. These two distributions do not have comparable variances. The variance for X (Type A) is $s_1^2 = 1342.365789$, and for Y (Type B) is $s_2^2 = 2336.747368$, moreover: $s_1^2/s_2^2 = 0.5745$. There is less variability in X compared with Y, which can be seen graphically (X's graph is narrower than Y's).

11.4.#C-5: (a) Replace $\Delta = \mu_1 - \mu_2$ with $\Delta = \theta_1 - \theta_2$. Now, let $\vec{x}$ contain the date for Type A, and $\vec{y}$ the data for Type B. The data contains ties, thus:

$W2.p.ties(x, y, 0, 100000) = 0.0118$ to test the one sided hypothesis: $\mathbf{p} = 0.0118/2 \approx 0.0059 < 0.05 = \alpha \implies$ reject H_0

(b) $W2.ci(x, y, .1, 100000) =$
k Lower Upper Coverage
138 12 60 0.90843
139 13 60 0.90244
140 13 60 0.89949
141 13 60 0.89084
142 14 59 0.88648.

Thus, with $k = 140$ we obtain an interval of almost 90%. The interval is: (13,60).