STAT 500 Homework 3

Vu Thi-Hong-Ha; NetID: 851924086

September 11, 2020

1 Question 1

(a)
(i) Denote μ_H to be the mean cholesterol of people who have heart attack, and μ_N to be the mean cholesterol of people who do not have heart attack.
H_0: $\mu_H - \mu_N = 0$
H_A: $\mu_H - \mu_N > 0$
(ii)
$\bar{Y_1} - \bar{Y_2} = 265.4 - 193.1 = 72.3$
$S_1^2 = 43.645^2 = 1904.886$; $S_2^2 = 21.623^2 = 467.5541$. The estimate of the pooled standard deviation is $S_p = \sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} = \sqrt{\dfrac{15 \times 1904.886 + 19 \times 467.5541}{16+20-2}} = 33.19143$
(iii)
$$T = \frac{\bar{Y_1} - \bar{Y_2}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{72.3}{33.19143 \times \sqrt{\frac{1}{16} + \frac{1}{20}}} = 6.494356$$
(iv)
Degree of freedom: $n_1 + n_2 - 2 = 34$
P-value $= P(t_{34} > T) < 0.0001$
(v)
At 5% significance level, we reject the null that the mean cholesterol levels for people who have and do not have heart attack are equal. We conclude that the mean cholesterol for people who have had a heart attack is greater than that for people who have not had a heart attack.
(b)
A 95% confidence interval is $((\bar{Y_1} - \bar{Y_2}) \pm t_{n_1+n_2-2,1-\alpha/2} S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}) \approx (72.3 \pm 2.042 \times 33.19143 \times \sqrt{\dfrac{1}{16} + \dfrac{1}{20}}) =$ $(49.56694; 95.03306)$. The 95% confidence interval does not contain 0, so at 5% significance level, we reject the null that the two group has equal means.

2 Question 2

(a)

n_1	n_2	$Var(\bar{Y_1} - \bar{Y_2})$
1	49	$1.020408\sigma^2$
5	45	$0.2222222\sigma^2$
10	40	$0.125\sigma^2$
25	25	$0.08\sigma^2$
30	20	$0.0833\sigma^2$
40	10	$0.125\sigma^2$
45	5	$0.2222222\sigma^2$
49	1	$1.020408\sigma^2$

(b) We obtained the least variance when $n_1 = n_2$, so we should allocate 25 subjects to each treatment.

3 Question 3

$n_1 = n_2 = n = 6$
$\hat{\sigma} = s = 4$
(a)

$Var(\bar{Y_1} - \bar{Y_2}) \approx \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$. The standard error is $\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$

Since we assume equal variance across the two treatments, we can pool the variance. Then $S_p^2 = \frac{5S_1^2 + 5S_2^2}{10} = 16$. Hence $S_1^2 + S_2^2 = 32$.

The standard error of the difference between the two means is $\sqrt{\frac{32}{6}} = 2.309401$.

(b)
$n_1 + n_2 - 2 = 6 + 6 - 2 = 10$.
$t_{10,0.975} = 2.228$
(c)
$\alpha = 0.05$, $\delta = 5$.

$$-t_{10,1-\beta} \approx \frac{t_{10,0.975} \times 4 \times \sqrt{\frac{2}{6}} - 5}{4 \times \sqrt{\frac{2}{6}}} = 0.06293649 \approx -t_{10,0.4755}$$

The power is 47.55%.

(d)
$\alpha = 0.05$, then $z_{0.975} = 1.96$. Width of a 95% CI for the mean difference ≤ 8.
Consider $width = 8$. Then $8 = 2t_{2(n-1),0.975}S_p\sqrt{2/n}$.
Choose $n_0 = 8(\frac{1.96 \times 4}{8})^2 = 7.6832$
$t_{2(n_0-1),0.975} = t_{13.3664,0.975} = 2.16$
Iteration 1: $n_1 = 8(\frac{2.16 \times 4}{8})^2 = 9.3312$
$t_{2(n_1-1),0.975} = t_{16.6624,0.975} = 2.11$
Iteration 2: $n_2 = 8(\frac{2.11 \times 4}{8})^2 = 8.9042$
$t_{2(n_2-1),0.975} = t_{15.8084,0.975} = 2.12$
Iteration 3: $n_3 = 8(\frac{2.12 \times 4}{8})^2 = 8.9888$
$t_{2(n_3-1),0.975} = t_{15.9776,0.975} = 2.12$
Iteration 4: $n_4 = 8(\frac{2.12 \times 4}{8})^2 = 8.9888$
We choose n to be 9.
Furthermore, consider when we decrease the width of a confidence interval, then the confidence level will decrease. However, we want to keep the confidence level fixed, so we need to minimize the sampling errors. The sampling errors depend on the sample variance and the sample size. Normally, it is difficult to minimize the variance in the sample (due to natural reasons for example), so one common way to fix the confidence level but decrease the confidence interval width is to increase the sample size. Therefore, if we want to achieve confidence intervals with width smaller than 8, we should increase sample sizes (greater than 9).
(e)
$\beta = 0.9; \alpha = 0.05; \delta = 5; S_p = 4$.

$$n = \frac{(t_{2(n-1),1-\alpha/2} + t_{2(n-1),1-\beta})^2(2S_p^2)}{\delta^2}$$

$z_{1-\alpha/2} = 1.96; z_{1-\beta} = 1.28$

$$n_0 = \frac{(1.96 + 1.28)^2 \times 2 \times 4^2}{5^2} = 13.43693$$

$t_{2(n_0-1),0.975} = 2.05954; t_{2(n_0-1),0.9} = 1.31635$

Iteration 1: $n_1 = \frac{(2.05954 + 1.31635)^2 \times 2 \times 4^2}{5^2} = 14.58769$

$t_{2(n_1-1),0.975} = 2.05183; t_{2(n_1-1),0.9} = 1.3137$

Iteration 2: $n_2 = \frac{(2.05183 + 1.3137)^2 \times 2 \times 4^2}{5^2} = 14.49829$

$t_{2(n_2-1),0.975} = 2.05183; t_{2(n_2-1),0.9} = 1.3137$

Iteration 3: $n_2 = \dfrac{(2.05183 + 1.3137)^2 \times 2 \times 4^2}{5^2} = 14.49829$
We choose n to be 15.
(f)
(i)
When power increases, sample size increases, so we will need more samples than in (e).
(ii)
When δ decreases, n_0 increases. Hence, $2(n_0 - 1)$ increases, and $t_{2(n_0-1),0.975}$ and $t_{2(n_0-1),0.90}$ decrease. However, the amount of decrease in $t_{2(n_0-1),0.975}$ and $t_{2(n_0-1),0.90}$ is still smaller than the decrease in δ when we change δ from 5 to 3, so n_1 will increase. With the same trend over iterations, the sample size should increase, and we will need more samples than in (e).
(iii)
When the std deviation increases, n_0 increases, so degree of freedom will increase. Then $t_{2(n_0-1),0.975}$ and $t_{2(n_0-1),0.90}$ decrease, but with a smaller rate than the increasing rate of standard deviation. Hence, over the iterations, sample size will increase. We will need more samples than in (e).
When α decreases, $z_{1-\alpha}$ increases, n_0 increases, degree of freedom increases. Even then, $t_{2(n_0-1),0.995}$ still increases. Therefore, over iterations, the sample size will increase. We need more samples.

4 Question 4

(a) Let Y_{normal} be the sample of normal rats, and Y_{LSD} the sample of treatment group.
H_0: $\mu_{normal} - \mu_{LSD} = 0$
H_A: $\mu_{normal} - \mu_{LSD} \neq 0$
$\bar{Y}_{normal} = 450$; $\sigma = 60$; $\delta = 25$, $n_{normal} = n_{LSD}$
The number of samples for each group is 105.

```
proc power;
  twosamplemeans
     alpha = 0.05
     meandiff = 25
     stddev= 60
     npergroup = .
     power = 0.85;

  run;
```

The SAS System

The POWER Procedure
Two-Sample t Test for Mean Difference

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Alpha	0.05
Mean Difference	25
Standard Deviation	60
Nominal Power	0.85
Number of Sides	2
Null Difference	0

Computed N per Group	
Actual Power	N per Group
0.852	105

(b) Let x be the cost of a normal rat, and y be the cost of a rat under treatment.
If we change the method and keep the sample size of each group to be n, the new standard error of the difference of means is $\dfrac{S_p\sqrt{\frac{2}{n}}}{2} = \dfrac{S_p}{\sqrt{2n}}$. The total cost of changing is $2xn + 2yn$.
If we do not change the methods, with the same cost $2xn + 2yn$, the new sample size of each group is $2n$. Then, we standard error of the difference of means is $S_p\sqrt{\dfrac{1}{2n} + \dfrac{1}{2n}} = \dfrac{S_p}{\sqrt{n}} > \dfrac{S_p}{\sqrt{2n}}$.
Therefore, this change is cost effective in terms of variance.

5 Question 5

$H_0 : \mu_1 - \mu_2 = 0$
$H_a : \mu_1 - \mu_2 > 0$
Given $\alpha, \delta, 1-\beta, n_1 = n_2 = n$:

$1 - \beta = \text{P}(\text{reject } H_0 \mid \mu_1 - \mu_2 = \delta) = \text{P}(\frac{(\bar{Y_1} - \bar{Y_2}) - 0}{S_p\sqrt{\frac{1}{n} + \frac{1}{n}}} > t_{2(n-1),1-\alpha} \mid \mu_1 - \mu_2 = \delta)$

$= \text{P}((\bar{Y_1} - \bar{Y_2}) > t_{2(n-1),1-\alpha} S_p \sqrt{\dfrac{1}{n} + \dfrac{1}{n}} \mid \mu_1 - \mu_2 = \delta)$

$= \text{P}(\dfrac{(\bar{Y_1} - \bar{Y_2}) - \delta}{S_p\sqrt{\frac{2}{n}}} > \dfrac{t_{2(n-1),1-\alpha} S_p \sqrt{\frac{2}{n}} - \delta}{S_p\sqrt{\frac{2}{n}}} \mid \mu_1 - \mu_2 = \delta)$

This implies: $-t_{2(n-1),1-\beta} \approx \dfrac{t_{2(n-1),1-\alpha} S_p \sqrt{\frac{2}{n}} - \delta}{S_p\sqrt{\frac{2}{n}}}$

Then, $-t_{2(n-1),1-\beta} S_p \sqrt{\frac{2}{n}} + \delta \approx t_{2(n-1),1-\alpha} S_p \sqrt{\frac{2}{n}}$
$n \approx \frac{(t_{2(n-1),1-\alpha} + t_{2(n-1),\beta})^2 2 S_p^2}{\delta^2}$