Solution for Midterm of STAT 3004

1

(a)

Let p be the proportion of deaths from lung cancer in this plant, and denote p_0 as the proportion in the general population. Then we want to test

$$H_0 : p = p_0 \quad H_1 : p \neq p_0$$

(b)

Two-sided test, since only check whether there is a difference.

(c)

Since $np_0(1-p_0) = 20 \times 0.12 \times 0.88 < 5$, then we should use the exact method. Since $\hat{p} = 0.25 > p_0 = 0.12$, the textbook uses the formula

$$p\text{-value} = 2 \times \Pr(X \geq x) = \min\left[2\sum_{k=x}^{n}\binom{n}{k}p_0^k(1-p_0)^{n-k}, 1\right]$$

then the code can be written as

```
> min(2*sum(choose(20, 5:20) * 0.12^(5:20) * (1-
0.12)^(15:0)),1)
[1] 0.1654388
```

Or with the probability mass function (pmf) of Binomial distribution,

$$\Pr(X \geq x) = 1 - \Pr(X \leq x - 1) = 1 - F(x - 1)\,,$$

where $F(x) = \Pr(X \leq x)$, that is

```
> min(2*(1-pbinom(5-1, 20, 0.12)), 1)
[1] 0.1654388
```

Also, it can be written as

$$\Pr(X \geq x) = \Pr(X > x + 1) = G(x + 1)\,,$$

where $G(x) = 1 - F(x)$.

```
> min(2*pbinom(5-1, 20, 0.12, lower.tail=FALSE), 1)
[1] 0.1654388
```

Furthermore, we can directly call the function built in R software,

```
> binom.test(5, 20, p=0.12, alternative="two.sided")


    Exact binomial test


data:  5 and 20
number of successes = 5, number of trials = 20, p-
value = 0.08272
alternative hypothesis: true probability of success is
not equal to 0.12
95 percent confidence interval:
 0.08657147 0.49104587
sample estimates:
probability of success
                  0.25
```

But the formula of p-value is different from the textbook,

$$p\text{-value} = \Pr(X \geq x) + \Pr(X \leq y)\,,$$

where

$$y = \underset{0 \leq m \leq np_0}{\operatorname{argmax}} \left(\Pr(X \leq m) \leq \Pr(X \geq x)\right)\,.$$

2

(a)

Let X be the BMI, then the test statistic is

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

and hence the 95% CI can be constructed from

$$\Pr(|T| < t_{n-1,0.975}) = 0.95\,,$$

that is,

$$\mu \in \left(\bar{X} - t_{n-1,0.975}\frac{s}{\sqrt{n}}, \bar{X} + t_{n-1,0.975}\frac{s}{\sqrt{n}}\right) = (24.29007, 25.70993)\,.$$

(b)

Consider the hypothesis testing

$$H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0\,,$$

where $\mu_0 = 24$, then

$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = 2.820657 > 2.002465 = t_{57,0.975}\,,$$

so we would reject the null hypothesis and conclude that the BMI of the considered group is NOT equal to 24.0.

(c)

Reject, since 24.0 does not lie in the 95% CI.

3

(a)

Let σ_1^2, σ_2^2 be the variance of the two populations respectively, then we test

$$H_0 : \sigma_1^2 = \sigma_2^2 \qquad H_1 : \sigma_1^2 \neq \sigma_2^2 \,.$$

The test statistic is

$$F = \frac{\sigma_1^2}{\sigma_2^2} \sim F_{n_1-1, n_2-1} \,.$$

Since

$$\frac{s_1^2}{s_2^2} = 0.06659729 < F_{36,18,0.025} = 0.465444 \,,$$

we would reject H_0 and argue that $\sigma_1^2 \neq \sigma_2^2$.

(b)

Let μ_1, μ_2 be the mean of the two populations respectively, then we test

$$H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 \neq \mu_2 \,.$$

We should use two sample t test with unequal variance based on (a), and the test statistic is

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} \sim \mathcal{T}(d) \,,$$

where d can be computed using Satterthwaite's approximation,

$$d = \frac{\left(\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2\right)^2}{\left(\hat{\sigma}_1^2/n_1\right)^2/(n_1-1) + \left(\hat{\sigma}_2^2/n_2\right)^2/(n_2-1)} = 19.24095\,.$$

Since

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = -2.152988 < -2.091251 = t_{d,0.025}$$

then we would reject H_0 and argue that these two populations do NOT have the same mean age.

4

(a)

advantages:

- the tests with rank require no or very limited assumptions to be made about the format of the data
- the ranks can alleviate the issue induced by outlying observations
- the ranks would be meaningful in the analysis of ordinal data, while treating them as continuous measurements are inappropriate.

disadvantages:

- the ranks would discard information captured by the continuous measurements
- the tests with rank focus on hypothesis testing instead of estimation of effects
- tied values would reduce the number of points to be analyzed, and hence might be problematic

(b)

Since the sample size $n = 14 < 20$, then we need to use the exact version. The differences are as follows,

SUBJECT	1	2	3	4	5	6	7	8	9	10	11	12	13	14
d_i	16	-7	-2	38	12	2	23	-14	6	-13	-3	36	8	40

and there are $C = 9$ plus signs. The p-value is

$$p\text{-value} = 2\sum_{k=C}^{n}\binom{n}{k}\frac{1}{2^n} = 0.4239502\,,$$

so we cannot reject H_0.

(c)

First rank the data by absolute value of the difference,

SUBJECT	1	2	3	4	5	6	7	8	9	10	11	12	13	14
d_i	16	-7	-2	38	12	2	23	-14	6	-13	-3	36	8	40
order	10	5	1	13	7	2	11	9	4	8	3	12	6	14
rank	10	5	1.5	13	7	1.5	11	9	4	8	3	12	6	14

Since the number of pairs with nonzero d_i is $14 < 16$, then we cannot use the normal approximation. The rank sum for positive differences

$$R_1 = 10 + 13 + 7 + 1.5 + 11 + 4 + 12 + 6 + 14 = 78.5$$

From Table 10 in the textbook's Appendix, the critical values for $\alpha = 0.05, n = 14$ are $(21, 84)$, and since R_1 lies in this range, then we cannot reject H_0.

Alternatively, the critical values can also be obtained by

```
> qsignrank(0.025,14)
[1] 22
> qsignrank(0.975,14)
[1] 83
```

that is, $(22, 83)$.

5

(a)

Let p_1, p_2 be the proportions of subjects who withdrew in these two groups respectively. Then the sample proportions are

$$\hat{p}_1 = \frac{27}{314} = 0.08598726\,, \quad \hat{p}_2 = \frac{20}{308} = 0.06493506\,.$$

(b)

We can test for association by computing the test statistic,

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

The expected E_{ij} can be calculated as

	YES	**NO**
Calcitriol	$314 \times \frac{47}{622} = 23.72669$	$314 \times \frac{575}{622} = 290.2733$
Calcium	$308 \times \frac{47}{622} = 23.27331$	$308 \times \frac{575}{622} = 284.7267$

Note that none of the E_{ij} are < 5, we can validly use the chi-square test, and $X^2 \sim \chi_1^2$ under H_0.

Since

$$X_{obs}^2 = 0.9865047 < 3.841459 = \chi_{1,0.95}^2 \,,$$

then we cannot reject H_0, and conclude that there is NO enough evidence to say that there is association between these two groups.

Alternatively, with Yates' continuity correction,

$$X_{Yates}^2 = \sum_{i,j} \frac{(|O_{ij} - E_{ij}| - 1/2)^2}{E_{ij}}$$

then

$$X_{\text{Yates},obs}^2 = 0.7081442 < 3.841459 = \chi_{1,0.95}^2 \,.$$

(c)

No need to use Fisher's exact test since none of the expected value < 5. If required, the procedures are described in the equation 10.10 of the textbook.

- rearrange the rows and columns of the observed table so the smaller row total is in the first row and the smaller column total is in the first column,

	YES	**NO**
Calcium	20	288
Calcitriol	27	287

- start with the table with 0 in the $(1,1)$ cell. The other cells in this table are then determined from the row and column margins. This gives the 0-table with $\Pr(K=0)$.
- construct the next table by increasing the $(1,1)$ cell by 1, this is the 1-table with $\Pr(K=1)$.
- continue increasing cell $(1,1)$, until one of the other cells reaches 0.

The p-value is

$$p = 2\min\{\Pr(K \le 20), \Pr(K \ge 20), 0.5\}\,.$$

Or directly from R,

```
table = matrix(c(20, 27, 288, 287), 2)
p_lower=fisher.test(table,alternative = "l")$p.value
p_upper=fisher.test(table,alternative = "g")$p.value
2*min(p_lower,p_upper,0.5) #0.4003627

# or
fisher.test(table)$p.value #0.3639729
```