In The Name of God
Statistical Inference HW#5

Amin Asadi
810196410

Problem 1

a.

The population parameter of interest :
The proportion of graduates at a mid-sized university who found a job within one year of completing their undergraduate degree.

Point of estimate of this parameters:
$\widehat{p} = \frac{348}{400} = 0.87$

b.

1. Independence:
   - Graduates are randomly sampled from the population
   - The size of sample is 400 which is smaller than 10% of the size of population($400 < \frac{4500}{10} = 450$)
2. Sample size/skew
   - Number of success = 348 > 10
   - Number of failures = 400-348 =52 > 10

Hence the conditions are met.

c.

$$CI = (\widehat{p} - Z_{.975} \times SE\,,\ \widehat{p} + Z_{.975} \times SE)$$

$$SE = \sqrt{\frac{\widehat{p} \times (1-\widehat{p})}{n}} = \sqrt{\frac{0.87 \times -0.13}{400}} \approx 0.017$$

$$CI = (0.87 - 1.96 \times 0.017\,, 0.87 + 1.96 \times 0.017)$$

$$CI = (0.83,\ 0.90)$$

It means that we are 95% confident that 83% to 90% of all the graduates at this university find a job within one year of completing their undergraduate degree.

d.

This means that if we take many samples of this from the population (like the one we have in this question) and build 95% confidence intervals, about 95% of them will capture the true proportion of the population. So, there is a 95% chance for our calculated confidence interval to capture the true proportion of the population.

e.

$$CI = (\widehat{p} - Z_{.995} \times SE\,,\ \widehat{p} + Z_{.995} \times SE)$$

$$SE = \sqrt{\frac{\widehat{p} \times (1-\widehat{p})}{n}} = \sqrt{\frac{0.87 \times 0.13}{400}} \approx 0.017$$

$$CI = (0.87 - 1.96 \times 0.017\,, 0.87 + 1.96 \times 0.017)$$

$$CI = (0.82,\ 0.91)$$

f.

The width of the 95% CI is about 0.07
while the width of the 99% CI is 0.09

So the 99% CI is wider.
99% CI has a higher chance of capturing the true proportion than the 95% CI so it should be wider.

Problem 2

a.

The population parameter of interest :
The proportion of Greeks who consider their lives as suffering.

Point of estimate of this parameters:

$$\widehat{p} = 0.25$$

b.

Independence:

- Individuals are randomly sampled from the population.
- The size of sample is 1000 which is smaller than 10% of the size of population.

Sample size/skew

- Number of success = $25\% \times 1000 = 2500 > 10$
- Number of failures = $75\% \times 1000 = 7500 > 10$

Hence the conditions are met.

c.

$$CI = (\widehat{p} - Z_{.975} \times SE\,,\ \widehat{p} + Z_{.975} \times SE)$$

$$SE = \sqrt{\frac{\widehat{p} \times (1-\widehat{p})}{n}} = \sqrt{\frac{0.25 \times 0.75}{400}} \approx 0.013$$

$$CI = (0.25 - 1.96 \times 0.013\,,0.87 + 1.96 \times 0.013)$$

$$CI = (0.223,\ 0.276)$$

d.

With higher confidence level, the confidence interval will be wider because it will have greater z* and hence the margin around the point estimate will be larger so as to capture the true proportion with higher chance.

e.

With larger sample size, the confidence interval will be narrower because it will have smaller standard error and hence the margin around the point estimate will be smaller.

Problem 3

a. No, it can't since this sample only consists of the students who took the SAT and also wanted to respond to optional surveys and hence it is biased and also there is non-response bias and therefore can't represent the whole population.

b.

$\widehat{p} = 0.55$

$CI = (\widehat{p} - Z_{.975} \times SE\,,\ \widehat{p} + Z_{.975} \times SE)$

$SE = \sqrt{\frac{\widehat{p} \times (1-\widehat{p})}{n}} = \sqrt{\frac{0.55 \times 0.45}{1509}} \approx 0.0128$

$CI = (0.25 - 1.64 \times 0.013\,, 0.87 + 1.64 \times 0.013)$

$CI = (0.528,\ 0.571)$ = (52.8%, 57,1%)

We are 90% confident that 52,8% to 57,1% of the part of the population that this sample represents(students who took SAT and respondend to survey) are fairly certain that they will participate in a study abroad program in college.

c.

❖ As stated in part a, this sample is not representative of the whole population, therefore it wouldn't be appropriate to generalize the result to the whole population.

❖ With that being said, if this sample would be representative of the whole population we could conclude that the majority of high school seniors are fairly certain that they will participate in a study abroad program (because H_0 is that p=50% and H_A is that p > 50% and since 50% doesn't fall into the CI → H_0 is rejected in favor of the alternative.)

Problem 4

a.

H_0 : There is no difference in proportions of yawns between the treatment group and control group

$(p_{control} = p_{treatment})$

H_A : proportions of yawns in the treatment group and is greater than that of control group

$(p_{control} < p_{treatment})$.

b.

Yawn rate in treatment group = $\frac{10}{34} \approx 0.29$

Yawn rate in control group = $\frac{4}{16} = 0.25$

→ observed difference = 0.29 - 0.25 = 0.04

c.

"Note: As stated in the announcement we assume that conditions for the test are met."

Because we are calculating p-value for difference of proportions, we used pooled

$$\widehat{p}_{pool} = \frac{total\ success}{total\ n} = \frac{10+4}{14+36} = \frac{14}{50} = 0.28$$

$$SE_{pooled} = \sqrt{\widehat{p}_{pool}(1-\widehat{p}_{pool})(\frac{1}{n_1}+\frac{1}{n_2})}$$

$$= \sqrt{0.28 \times 0.72 \times (\frac{1}{16}+\frac{1}{34})} \approx 0.14$$

$$Z = \frac{point\ estimate - nul\ value}{SE} = \frac{0.04}{0.14} \approx 0.294$$

Because we have a one tail test:

$$\Rightarrow\ pvalue = P(Z > 0.294) \approx 0.61$$

We fail to reject H_0 since at a 5% significance level, there is no significant difference between the proportion of people yawning in the control and the treatment group and we conclude that yawning is not contagious.

Problem 5

Waiting Time	0-15	16-30	31-45	46+	Total
Expected %	50%	30%	10%	10%	100%
Expected #	75	45	15	15	150
Observed #	75	55	15	5	150

a.

H_0 : The distribution of wait-times this year is the same as the wait-times last year(it is as expected).

H_A : The distribution of wait-times this year is different from the wait-times last year(it is as expected).

b.

Test Conditions:

1. Independence:
    - The workers are randomly sampled
    - 150 < 10 % of total number of workers
    - Each worker contributes to one cell in the table

2. Sample size:
    - Each particular scenario has at least 5 expected cases

So the conditions are met for the chi-square test.

$$\chi^2 = \frac{(75-75)^2}{75} + \frac{(55-45)^2}{45} + \frac{(14-14)^2}{14} + \frac{(5-15)^2}{15} \approx 8.88$$

$$df = k - 1 = 4 - 1 = 3$$

Using the table we see that p-value is less than 0.05 so we reject H_0 in favor of the alternative hypothesis.

Also the exact p-value is about 0.030 as shown below:

```
> pchisq(8.888889, 3, lower.tail = F)
[1] 0.03080523
>
```

Therefore again we reject H_0 in favor of the alternative hypothesis.

c.

```
> observed <- c(75, 55, 15, 5)
> chisq.test(observed, p=c(0.5, 0.3, 0.1, 0.1))

        Chi-squared test for given probabilities

data:  observed
X-squared = 8.8889, df = 3, p-value = 0.03081
```

As we can see the χ^2 and p-value is exactly the same as part b. So in both parts we reject H_0 in favor of H_A.

Problem 6

Note: As stated in the question, we assume the conditions for the test are met.

a.

$$\text{Expected Count} = \frac{(row\ total) \times (column\ total)}{table\ total}$$

$$E_{Toyger,\,Yellow} = \frac{60 \times 72}{180} = 24$$

$$E_{Toyger,\,Purple} = \frac{60 \times 69}{180} = 23$$

$$E_{Toyger,\,Black} = \frac{60 \times 39}{180} = 13$$

b.

H_0 : Breed of cat and the it's prefered color of the toy are independent (prefered color of the toy doesn't vary by breed of cat).
H_A : Breed of cat and the it's prefered color of toy are dependent (prefered color of the toy varies by breed of cat).

c.

Because total of each row is 60 (as the same as Tyoger Breed) so:

- the expected values for all cells in column 1 =

  $\frac{60 \times 72}{180} = 24$

- the expected values for all cells in column 2 =

  $\frac{60 \times 69}{180} = 23$

- the expected values for all cells in column 3 =

  $\frac{60 \times 39}{180} = 13$

$$\chi^2 = \frac{(16-24)^2}{24} + \frac{(27-24)^2}{24} + \frac{(29-24)^2}{24} + \frac{(30-23)^2}{23} + \frac{(23-23)^2}{23} + \frac{(1}{}$$

$$+ \frac{(14-13)^2}{13} + \frac{(10-13)^2}{13} + \frac{(15-13)^2}{13} \approx 9.42$$

$$df = (R - 1) \times (C - 1) = 2 \times 2 = 4$$

P-value = pchisq(9.421126, 4, lower.tail = F) = 0.513
So because p-value is a little greater than $\alpha = 0.05$we fail to reject H_0.

Problem 7

a.

Since each individual has 24 options to choose from, so we expect participants to be correct $\frac{1}{24}$(about 4.16%) of the time.

Null Hypothesis(H_0): $p = \frac{1}{24} \approx 4.16\%$

Null Hypothesis(H_0): $p > \frac{1}{24}$

b.

	Correct	Wrong	Total
Expected %	4.16 %	95.83%	100%
Expected #	12	288	300
Observed #	125	175	300

Checking conditions for test:

1. Independence:
    a. random sampling has been done
    b. Sample size $< 10\%$ of population size
    c. Each case only contributes to one cell of table
2. Each scenario has at least 5 expected cases

```
2  # P7
3  observed <- c(125, 175)
4  p <- c(1/24, 23/24)
5  chisq.test(observed, p=p)
6
```

```
> chisq.test(observed, p=p)

        Chi-squared test for given probabilities

data:  observed
X-squared = 1056.5, df = 1, p-value < 2.2e-16
```

As we can see in the result:

$p - value \approx 0$

So H_0 is rejected in favor of H_A which means that the observed values for correct and wrong answers haven't occurred by chance and people have some ability to recognize the smell of their coffee.

Problem 8

Checking conditions for test:

3. Independence:

a. random sampling has been done
b. Sample size $< 10\%$ of population size
c. Each case only contributes to one cell of table

4. Each scenario has at least 5 expected cases

H_0 : a student's smoking habit and exercise level are **independent**.
H_A : a student's smoking habit and exercise level are **dependent**.

```
1  library(MASS)
2
3  t <- table(survey$Exer, survey$Smoke)
4
5  chisq <- chisq.test(t)
6
7  print(chisq$expected)
8  print(chisq$observed)
9  print(chisq)
10
11
```

```
print(chisq$expected)

        Heavy    Never    Occas    Regul
Freq 5.360169 92.09746 9.258475 8.283898
None 1.072034 18.41949 1.851695 1.656780
Some 4.567797 78.48305 7.889831 7.059322
```

```
print(chisq$observed)

     Heavy Never Occas Regul
Freq     7    87    12     9
None     1    18     3     1
Some     3    84     4     7
```

```
> print(chisq)

        Pearson's Chi-squared test

data:  t
X-squared = 5.4885, df = 6, p-value = 0.4828
```

As we can see in the test results, the p-value = 0.48

0.48 >> 0.05 → Because p-value is by far greater than significance level(0.05) we fail to reject H_0.

This means that a student's smoking habit and exercise level **are not dependent (we don't have enough evidence to claim their dependence).**