STAT 22000: Final Part 1

Zixi Li

Problem 1

(a)

Let Y be the number of time he wins in the 100 spins. Now Y is Bin(n = 100,p = 18/38). So the expected value and SD are

$$np = 100 \times (18/38) = 47.368$$

$$SD = \sqrt{np(1-p)} = \sqrt{100(18/38)(20/38)} = 4.993$$

(b)

By normal approximation of binomial, we know

$$Y \sim \text{Bin}(n = 100, p = 18/38) \approx N(\mu = np = 47.368, \sigma = \sqrt{np(1-p)} = 4.993)$$

With continuity correction, the end point 55 is changed to 54.5 since $\{Y \geq 55\}$ and $\{Y \geq 54.5\}$ contains the same set of integers.

$$P(Y \geq 54.5) = P\left(Z \geq \frac{54.5 - 47.368}{4.993}\right) \approx P(Z \geq 1.428) \approx 0.0766$$

```
pnorm(1.428,lower.tail = F)
```

```
## [1] 0.0766459
```

(c)

Event A is more likely to happen, it's more likely to win at least half of the times in 20 spins. Because the more spins are made, the more the number of red deviated from 18/38 of the number of spins, and hence 25 in 50 is more far from center than 10 in 20.

We can also use the following R codes to check the answer:

```r
1-pbinom(9, size=20, p=18/38)
```

```
## [1] 0.493719
```

```r
1-pbinom(24, size=50, p=18/38)
```

```
## [1] 0.407858
```

Problem 2

(a)

We can because the household size must be integers which are ≥ 1, and we already know about 28.4% of U.S. households are of size 1, such that the probability that a randomly chosen household in the U.S. is of size 2 or more is $1 - 28.4\% = 71.6\%$

(b)

We can't accurately calculate because the population distribution is skewed and has outliers, and the sample size 50 is not large enough to guaranteen the sampling distribution of the mean household to be normal.

But to roughly calculate the mean size:

$$P(\bar{X} > 2.8) = P\left(Z > \frac{2.8 - 2.52}{1.4}\right) = P(Z > 0.2) = 0.42$$

```r
pnorm(0.2,lower.tail = F)
```

```
## [1] 0.42074
```

(c)

Histogram A is (iii) histogram of the sizes of 500 randomly sampled households. Because the household values are all integers and the distribution is skewed.

Histogram B is (i) sampling distribution of the mean size of 16 randomly selected households. Because the distribution is approximately normal and the standard error is relatively larger than histogram C.

Histogram C is (ii) sampling distribution of the mean size of 100 randomly selected households. Because the distribution is approximately normal and the standard error is relatively smaller than histogram B.

Problem 3

(a)

This is an example of paired data. The test statistic can be calculated as

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{10}{10.7438/\sqrt{8}} = 2.633$$

with $8 - 1 = 7$ degrees of freedom.

(b)

The upper-sided p-value can be found using the following R codes:

```
pt(2.633, df=7, lower.tail = F)
```

```
## [1] 0.0168825
```

We know that $p = 0.0169 < \alpha = 0.05$, so we conclude that H_0 is rejected, and the corneal thickness of healthy eye is larger than the corneal thickness of diseased eye.

Problem 4

(a)

$H_0 : p_r = p_n$

$H_A : p_r > p_n$

where p_r is the proportion of irradiated garlic bulbs that are marketable after 240 days, p_n is the proportion of nontreated garlic bulbs that are marketable after 240 days.

(b)

Among the irradiated sample, the proportion that are marketable is $\hat{p}_r = 153/180 = 0.85$, among the untreated sample, the proportion is $\hat{p}_n = 119/180 = 0.661$

Under H_0, the estimate of the common p is the pooled sample proportion,

$$\hat{p} = \frac{153 + 119}{180 + 180} = 0.756$$

The standard error of the difference under H_0 is then

$$\text{SE} = \sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_r}+\frac{1}{n_n}\right)} = \sqrt{0.756\times(1-0.756)\left(\frac{1}{180}+\frac{1}{180}\right)} \approx 0.0453$$

The test statistic is

$$z = \frac{\widehat{p}_r - \widehat{p}_n}{\text{SE}} \approx \frac{0.85-0.661}{0.0453} \approx 4.172$$

(c)

The upper-sided p-value can be calculated using the following R codes:

```
pnorm(4.172,lower.tail = F)
```

```
## [1] 0.0000150969
```

We know that $p = 0.0000150969 < \alpha = 0.01$, so we conclude that H_0 is rejected, and ionizing radiation is beneficial as far as marketability is concerned.

(d)

The necessary conditions required to perform the test above is that $n_1\hat{p}$, $n_1(1-\hat{p})$, $n_2\hat{p}$, $n_2(1-\hat{p})$ all ≥ 10.

$$n_1\hat{p} = 136, \quad n_1(1-\hat{p}) = 44, \quad n_2\hat{p} = 136, \quad n_2(1-\hat{p}) = 44$$

We know that all the conditions are satisfied.

(e)

The 99% CI for $p_r - p_n$ is

$$\begin{aligned}
&(\widehat{p}_r - \widehat{p}_n) \pm 2.576\sqrt{\frac{\widehat{p}_r\,(1-\widehat{p}_r)}{n_r}+\frac{\widehat{p}_n\,(1-\widehat{p}_n)}{n_n}} \\
=&0.85-0.661 \pm 2.576\sqrt{\frac{0.85(1-0.85)}{180}+\frac{0.661(1-0.661)}{180}} \\
=&(0.075, 0.303)
\end{aligned}$$

(f)

The 99% CI for p_r is

$$\widehat{p}_r \pm 2.576\sqrt{\frac{\widehat{p}_r\,(1-\widehat{p}_r)}{n_r}} = 0.85 \pm 2.576\sqrt{\frac{0.85(1-0.85)}{180}} = (0.781, 0.919)$$

Problem 5

(a)

y = jaw width (inches), x = shark length (feet).

$$\begin{aligned} \text{slope } &= r\frac{s_y}{s_x} \approx 0.875\frac{2.81}{2.55} \approx 0.964 \\ \text{intercept } &= \bar{y} - (\text{ slope })\bar{x} = 15.7 - (0.964)(15.59) \approx 0.671 \end{aligned}$$

The equation of the least square regression line is

Predicted jaw width in inches $= 0.671 + 0.964$ (shark length in feet)

(b)

Slope: for each extra feet in shark length, a shark is expected to have 0.964 inches wider jaw on average.

(c)

The correct statement is (iv) The sample variance of the residuals is $1 - r^2$ of the sample variance of jaw widths.

(d)

With $n - 2 = 44 - 2 = 42$ degrees of freedom, the critical value is $t^* = 2.02$. The 95% CI for the slope is

$$\text{estimate } \pm t^*\text{SE} = 0.964 \pm 2.02 \times 0.08228 \approx (0.798, 1.130)$$

```
qt(0.05/2,df=44-2,lower.tail = F)
```

```
## [1] 2.01808
```

(e)

The correct answer is (i) 0.875

(f)

y = shark length (feet), x = jaw width (inches).

$$\begin{aligned}\text{slope } &= r\frac{s_y}{s_x} \approx 0.875\frac{2.55}{2.81} \approx 0.794\\ \text{intercept } &= \bar{y} - (\text{ slope })\bar{x} = 15.59 - (0.794)(15.7) \approx 3.124\end{aligned}$$

The equation of the least square regression line is

Predicted shark length in feet $= 3.124 + 0.794$ (jaw width in inches)

Predicted shark length of a shark with 20 inches jaw is $3.124 + 0.794 \times 20 = 19.004$ feet

Problem 6

False. The conditions required for using large-sample CI is that $n\hat{p}$ and $n(1-\hat{p})$ need to be both ≥ 10. Here $n(1-\hat{p}) = 100(1-0.97) = 3 < 10$.

Problem 7

False. The two sample proportions, $\hat{p}_B$ and $\hat{p}_F$ are calculated based on the same sample. They were not independent, but correlated. Some students are familiar with the map of Europe and they can identify both countries, but some students aren't. One can not use a two-sample CI here.