Suggested Answers:

Part A:

1. Dividing the cells by 1,058 (the total number of observations), we obtain

	Y	1	2	3	Total
		Voted for			
X		Soong	Lien	Chen	
1	Mainlanders	9.26%	2.74%	2.08%	14.08%
2	Hakha	4.44%	3.12%	3.59%	11.15%
3	Holo	21.17%	17.30%	36.20%	74.76%
Total		34.88%	23.16%	41.87%	100.00%

Note that the numbers don't add up, mainly because of an error in the original table: 224+183+383=790, not 791, so the total should have been 1,057, not 1,058. This is something you should check with real data provided to you.

2. $E(Y \mid X=1)=\sum_{j=1}^{3} j\cdot \Pr(Y=j \mid X=1)=1+\left(0+1\cdot\frac{29}{149}+2\cdot\frac{22}{149}\right)=1\frac{73}{149}\approx 1.49$,

since $\Pr(Y=1 \mid X=1)=\frac{98}{149}$, $\Pr(Y=2 \mid X=1)=\frac{29}{149}$, & $\Pr(Y=3 \mid X=1)=\frac{22}{149}$.

Similarly,

$$E(Y \mid X=2)=\sum_{j=1}^{3} j\cdot \Pr(Y=j \mid X=2)=2+\left(-1\cdot\frac{47}{118}+0+1\cdot\frac{38}{118}\right)=1\frac{109}{118}\approx 1.93$$

$$E(Y \mid X=3)=\sum_{j=1}^{3} j\cdot \Pr(Y=j \mid X=3)=2+\left(-1\cdot\frac{224}{791}+0+1\cdot\frac{383}{791}\right)=2\frac{161}{791}\approx 2.20$$

3. $\Pr(X=1 \mid Y=3)=\frac{22}{443}\approx 4.97\%$, $\Pr(X=2 \mid Y=3)=\frac{38}{443}\approx 8.58\%$,

$\Pr(X=3 \mid Y=3)=\frac{383}{443}\approx 86.456\%$.

4. No, since

$\Pr(X=1)\cdot\Pr(Y=3)=(14.08\%)\cdot(41.87\%)\approx 5.90\%\neq 2.08\%=\Pr(X=1,Y=3)$.(

Note: Any similar calculation would also be valid.)

5. $E(X \mid Y=1)=\sum_{x=1}^{3} x\cdot \Pr(X=x \mid Y=1)=2+\left(-1\cdot\frac{98}{369}+0+1\cdot\frac{224}{369}\right)=2\frac{126}{369}\approx 2.34$

$E(X^2 \mid Y=1)=\sum_{x=1}^{3} x^2\cdot \Pr(X=x \mid Y=1)=4+\left(-3\cdot\frac{98}{369}+0+5\cdot\frac{224}{369}\right)=4\frac{826}{369}\approx 6.24$

Hence, $Var(X \mid Y=1)=E(X^2 \mid Y=1)-\left[E(X \mid Y=1)\right]^2\approx 6.24-2.34^2=0.7644$

$E(X \mid Y=2)=\sum_{x=1}^{3} x\cdot \Pr(X=x \mid Y=2)=2+\left(-1\cdot\frac{29}{245}+0+1\cdot\frac{183}{245}\right)=2\frac{154}{245}\approx 2.63$

$E(X^2 \mid Y=2)=\sum_{x=1}^{3} x^2\cdot \Pr(X=x \mid Y=2)=4+\left(-3\cdot\frac{29}{245}+0+5\cdot\frac{183}{245}\right)=4\frac{828}{245}\approx 7.38$

Hence, $Var(X \mid Y=2)=E(X^2 \mid Y=2)-\left[E(X \mid Y=2)\right]^2\approx 7.38-2.63^2=0.4631$

$E(X \mid Y=3)=\sum_{x=1}^{3} x\cdot \Pr(X=x \mid Y=3)=2+\left(-1\cdot\frac{22}{443}+0+1\cdot\frac{383}{443}\right)=2\frac{361}{443}\approx 2.81$

$E(X^2 \mid Y=3)=\sum_{x=1}^{3} x^2\cdot \Pr(X=x \mid Y=3)=4+\left(-3\cdot\frac{22}{443}+0+5\cdot\frac{383}{443}\right)=4\frac{1849}{443}\approx 8.17$

Hence, $Var(X \mid Y=3)=E(X^2 \mid Y=3)-\left[E(X \mid Y=3)\right]^2\approx 8.17-2.81^2=0.2739$

Thus, candidate Soong has the most diverse supporter base, while candidate Chen has the least diverse supporter base.

6. Recall $E(Y \mid X=1)=1\frac{73}{149}\approx 1.49$.

$E(Y^2 \mid X=1)=\sum_{y=1}^{3} y^2\cdot \Pr(Y=y \mid X=1)=4+\left(-3\cdot\frac{98}{149}+0+5\cdot\frac{22}{149}\right)=4\frac{-184}{149}\approx 2.765$

Hence, $Var(Y \mid X=1)=E(Y^2 \mid X=1)-\left[E(Y \mid X=1)\right]^2\approx 2.765-1.49^2=0.5449$

Recall $E(Y \mid X=2)=1\frac{109}{118}\approx 1.93$.

$E(Y^2 \mid X=2)=\sum_{y=1}^{3} y^2\cdot \Pr(Y=y \mid X=2)=4+\left(-3\cdot\frac{47}{118}+0+5\cdot\frac{38}{118}\right)=4\frac{49}{118}\approx 4.415$

Hence, $Var(Y \mid X=2)=E(Y^2 \mid X=2)-\left[E(Y \mid X=2)\right]^2\approx 4.415-1.93^2=0.6901$

Recall $E(Y \mid X=3)=2\frac{159}{791}\approx 2.20$.

$E(Y^2 \mid X=3)=\sum_{y=1}^{3} y^2\cdot \Pr(Y=y \mid X=3)=4+\left(-3\cdot\frac{224}{791}+0+5\cdot\frac{383}{791}\right)=4\frac{1243}{791}\approx 5.57$

Hence, $Var(Y \mid X=3)=E(Y^2 \mid X=3)-\left[E(Y \mid X=3)\right]^2\approx 5.57-2.20^2=0.73$.

Therefore, Holo have the most diverse choice (though Hakha are a very close second), while the mainlanders have the least diverse choice.

7. (Bonus Question) Possible answers include:
    a. Ethnicity does indeed influence politics.
    b. Some candidates have more homogeneous "fan base" than others.
    c. Some ethnicity groups have more homogeneous preferences about candidates (preferable their own ethnicity group) than others.

Part B:

1. $\hat{p} = \frac{1}{1068}\left(449 + 245 \cdot \frac{449}{823}\right) = \frac{449}{1068}\left(\frac{823+245}{823}\right) = \frac{449}{823} \approx 54.5565\%$
2. $\hat{p} = \frac{449}{823} \approx 54.5565\%$, same as 1.
3. $s.e. = \sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{\frac{449}{823} \cdot \frac{374}{823} \cdot \frac{1}{823}} \approx 1.74\%$.
4. The test statistic is $\frac{\hat{p} - 0.5}{s.e.} = \frac{0.045565}{0.0174} \approx 2.62$, so the *p*-value for the two-sided test H_0: $p = 0.5$ vs. H_1: $p \neq 0.5$ is (consulting the Z-table) 0.0044*2=0.0088, while the *p*-value of the one-sided test H_0: $p = 0.5$ vs. H_1: $p > 0.5$ is 0.0044. The two results are different because the two-sided test includes the possibility that $p<0.5$, while the one-sided test excludes it. (Note that you can "manipulate" the *p*-value by choosing which test to use!)
5. Yes, since H_0 is rejected with very low *p*-value (<1%).
6. (Bonus Question) Possible answers include:
    a. The undecided voters eventually had a different voting pattern compared to those who have already decided at the date of survey. This would be significant if some voters intentionally withheld their choice in the survey.
    b. Selection bias due to a high "declined-to-answer" rate. This would be significant if the group of voters who shy away from responding to UDN's inquiry have different characteristics than the whole population.
    c. Sampling bias due to the time of the survey. This would be significant if the surveys were conducted only in certain period of the day, such as noon (which only certain types of people would pick up their home phones).
    d. Sampling bias due to phone ownership dispersion within the population. This would be significant if telephones were not popular.
7. (Bonus Question) Possible answers include:
    a. Due to the past authoritarian dictatorship, supporters of the incumbents (then opposition party) might be less inclined to share their opinions. This would lead to "undecided" voters voting differently compare to the decided.
    b. If UDN has a bad reputation regarding polls, certain voters might simply hang-up directly when receiving calls from UDN. This would results in a selection bias due to the high "declined-to-answer" rate (699 out of 1767).

Part C:

1. Since 1ft = 0.3048m, we may scale the sample average and standard deviation by 1/0.0929 since $1ft^2 = (0.3048)^2 m^2 \doteqdot 0.0929\ m^2$. In fact, an average (with standard deviation in parentheses) of 644,864 (581,618) in square meters would be roughly 6,941258 (6,260,689) in square feet. That is what you should tell your American friends.
2. Since $1.96 \times \left(\frac{581{,}618}{\sqrt{32}} \right) = 201{,}520$, the 95% confidence interval for the mean campus size of public universities is 644,864 ± 201,520 = (443,344, 846,384).
3. $s.e. = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{581{,}618^2}{30} + \frac{238{,}078^2}{32}} \approx 114{,}225$. Also, from the Z table we see that Z= -1.645 would yield Pr (Z < -1.645) = 0.05. Hence, the 90% confidence interval for the difference in mean campus size is (644,864-283,557) ± (1.645*114, 225) = 361,307 ± 187,900 ≈ (173,407, 549,207).
4. Yes. Since the test statistic is $\frac{361{,}307}{114{,}225} \approx 3.16$, we have a *p*-value < 0.004 (beyond the lower bound of the Z table provided).
5. (Bonus Question) Possible answers include:
    a. Public schools, founded by the government, have an easier time obtaining large pieces of land when founding the school.
    b. Public schools receive more regular subsidies from the Ministry of Education, and hence, have more resources to maintain larger campuses.

Part D:

1. $\| Y - \hat{Y} \|^2 = \sum_{j-1}^{n} (Y_j - a - bX_j - cZ_j)^2$

2. $\frac{\partial}{\partial a} \| Y - \hat{Y} \|^2 = \sum_{j-1}^{n} -2 \cdot (Y_j - a - bX_j - cZ_j) = 0$ implies $na = \sum_{j=1}^{n} Y_j - b\sum_{j=1}^{n} X_j - c\sum_{j=1}^{n} Z_j$, or, $a = \overline{Y} - b\overline{X} - c\overline{Z}$.

3. $$\| Y - \hat{Y} \|^2 = \sum_{j-1}^{n} (Y_j - a - bX_j - cZ_j)^2 = \sum_{j-1}^{n} \left[Y_j - (\overline{Y} - b\overline{X} - c\overline{Z}) - bX_j - cZ_j \right]^2$$
$$= \sum_{j-1}^{n} \left[(Y_j - \overline{Y}) - b(X_j - \overline{X}) - c(Z_j - \overline{Z}) \right]^2 = \sum_{j-1}^{n} (y_j - bx_j - cz_j)^2$$

4. $$\left. \begin{aligned} \frac{\partial}{\partial b} \| Y - \hat{Y} \|^2 &= \sum_{j=1}^{n} -2x_j \cdot (y_j - bx_j - cz_j) = 0 \\ \frac{\partial}{\partial c} \| Y - \hat{Y} \|^2 &= \sum_{j=1}^{n} -2z_j \cdot (y_j - bx_j - cz_j) = 0 \end{aligned} \right\} \text{ imply } \begin{cases} \sum_{j=1}^{n} x_j y_j = b\sum_{j=1}^{n} x_j^2 + c\sum_{j=1}^{n} x_j z_j \\ \sum_{j=1}^{n} z_j y_j = b\sum_{j=1}^{n} x_j z_j + c\sum_{j=1}^{n} z_j^2 \end{cases}$$

Hence, we have $\begin{cases} b = \dfrac{\left(\sum_{j=1}^{n} x_j y_j\right)\left(\sum_{j=1}^{n} z_j^2\right) - \left(\sum_{j=1}^{n} z_j y_j\right)\left(\sum_{j=1}^{n} x_j z_j\right)}{\left(\sum_{j=1}^{n} x_j^2\right)\left(\sum_{j=1}^{n} z_j^2\right) - \left(\sum_{j=1}^{n} x_j z_j\right)^2} \\ c = \dfrac{\left(\sum_{j=1}^{n} x_j y_j\right)\left(\sum_{j=1}^{n} x_j z_j\right) - \left(\sum_{j=1}^{n} z_j y_j\right)\left(\sum_{j=1}^{n} z_j^2\right)}{\left(\sum_{j=1}^{n} x_j z_j\right)^2 - \left(\sum_{j=1}^{n} x_j^2\right)\left(\sum_{j=1}^{n} z_j^2\right)} \end{cases}$

(Don't you think this is very complicated? That is why we need linear algebra and matrix formulations!!)

5. $\beta = (W'W)^{-1}W'Y$ implies $W'W\beta = W'Y$ where $W'W = \begin{pmatrix} (X^0)' \\ (X^1)' \\ (X^2)' \end{pmatrix}\begin{pmatrix} X^0 & X^1 & X^2 \end{pmatrix}$.

Since $W'Y = \begin{pmatrix} (I_n)' \\ (X^1)' \\ (X^2)' \end{pmatrix} Y = \begin{pmatrix} \sum_{j=1}^{n} Y_j \\ (X^1)'Y \\ (X^2)'Y \\ \end{pmatrix}$, we have

$$W'W\beta = \begin{pmatrix} n & \sum_{j=1}^{n} X_j & \sum_{j=1}^{n} Z_j \\ \sum_{j=1}^{n} X_j & \sum_{j=1}^{n} X_j^2 & \sum_{j=1}^{n} X_j Z_j \\ \sum_{j=1}^{n} Z_j & \sum_{j=1}^{n} X_j Z_j & \sum_{j=1}^{n} Z_j^2 \end{pmatrix}\begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} an + b\sum_{j=1}^{n} X_j + c\sum_{j=1}^{n} Z_j \\ a\sum_{j=1}^{n} X_j + b\sum_{j=1}^{n} X_j^2 + c\sum_{j=1}^{n} X_j Z_j \\ a\sum_{j=1}^{n} Z_j + b\sum_{j=1}^{n} X_j Z_j + c\sum_{j=1}^{n} Z_j^2 \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{n} Y_j \\ (X^1)'Y \\ (X^2)'Y \\ \end{pmatrix}$$

The first row immediately yields the formula in 2., or $na = \sum_{j=1}^{n} Y_j - b\sum_{j=1}^{n} X_j - c\sum_{j=1}^{n} Z_j$.

Since $x = X - \overline{X}, y = Y - \overline{Y}, z = Z - \overline{Z}$, plugging $a = \overline{Y} - b\overline{X} - c\overline{Z}$ into the second row, we obtain $b\left[\sum_{j=1}^{n} X_j^2 - n\overline{X}^2\right] + c\left[\sum_{j=1}^{n} X_j Z_j - n\overline{X}\cdot\overline{Z}\right] = \sum_{j=1}^{n} X_j Y_j - n\overline{X}\cdot\overline{Y}$, which is exactly the first formula in 4, namely, $\sum_{j=1}^{n} x_j y_j = b\sum_{j=1}^{n} x_j^2 + c\sum_{j=1}^{n} x_j z_j$. Similarly, the third row would yield $b\left[\sum_{j=1}^{n} X_j Z_j - n\overline{X}\cdot\overline{Z}\right] + c\left[\sum_{j=1}^{n} Z_j^2 - n\overline{Z}^2\right] = \sum_{j=1}^{n} Y_j Z_j - n\overline{Y}\cdot\overline{Z}$, which is exactly the second formula in 4, namely, $\sum_{j=1}^{n} z_j y_j = b\sum_{j=1}^{n} x_j z_j + c\sum_{j=1}^{n} z_j^2$.