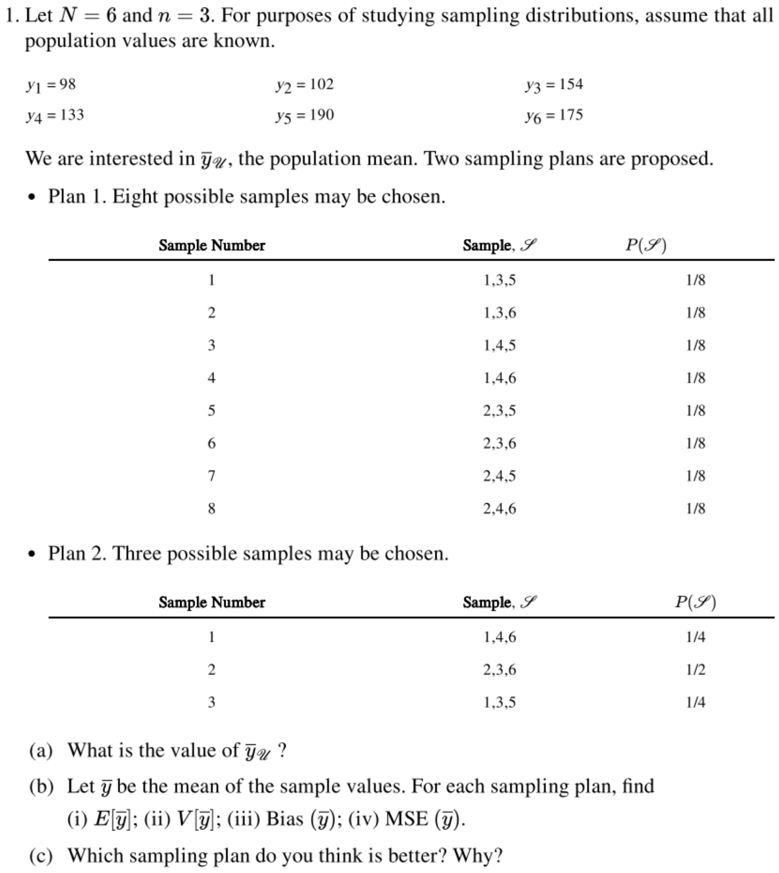
MA500HW2

Chang Lu

Feb 03 2025

1 Question 1



1. Let $N = 6$ and $n = 3$. For purposes of studying sampling distributions, assume that all population values are known.

$y_1 = 98$	$y_2 = 102$	$y_3 = 154$
$y_4 = 133$	$y_5 = 190$	$y_6 = 175$

We are interested in $\overline{y}_{\mathscr{U}}$, the population mean. Two sampling plans are proposed.

- Plan 1. Eight possible samples may be chosen.

Sample Number	Sample, $\mathscr{S}$	$P(\mathscr{S})$
1	1,3,5	1/8
2	1,3,6	1/8
3	1,4,5	1/8
4	1,4,6	1/8
5	2,3,5	1/8
6	2,3,6	1/8
7	2,4,5	1/8
8	2,4,6	1/8

- Plan 2. Three possible samples may be chosen.

Sample Number	Sample, $\mathscr{S}$	$P(\mathscr{S})$
1	1,4,6	1/4
2	2,3,6	1/2
3	1,3,5	1/4

(a) What is the value of $\overline{y}_{\mathscr{U}}$?

(b) Let $\overline{y}$ be the mean of the sample values. For each sampling plan, find (i) $E[\overline{y}]$; (ii) $V[\overline{y}]$; (iii) Bias $(\overline{y})$; (iv) MSE $(\overline{y})$.

(c) Which sampling plan do you think is better? Why?

Figure 1: This is the question 1

1.1 Population Mean

$$\bar{y}_U = \frac{1}{6}(98 + 102 + 154 + 133 + 190 + 175) = 142.0$$

1.2 Plan 1

1. **Expected Value ($E[\bar{y}]$):**

$$E[\bar{y}] = \frac{1}{8}(147.33+145.0+128.33+126.67+148.67+146.33+141.0+138.67) = 142.0$$

2. **Variance ($V[\bar{y}]$):**

$$V[\bar{y}] = 18.94$$

3. **Bias (Bias($\bar{y}$)):**

$$\text{Bias}(\bar{y}) = E[\bar{y}] - \bar{y}_U = 0$$

4. **Mean Squared Error (MSE($\bar{y}$)):**

$$\text{MSE}(\bar{y}) = V[\bar{y}] + \text{Bias}^2 = 18.94$$

1.3 Plan 2

1. **Expected Value ($E[\bar{y}]$):**

$$E[\bar{y}] = \frac{1}{4}(135.33) + \frac{1}{2}(147.33) + \frac{1}{4}(145.0) = 142.5$$

2. **Variance ($V[\bar{y}]$):**

$$V[\bar{y}] = 19.36$$

3. **Bias (Bias($\bar{y}$)):**

$$\text{Bias}(\bar{y}) = E[\bar{y}] - \bar{y}_U = 0.5$$

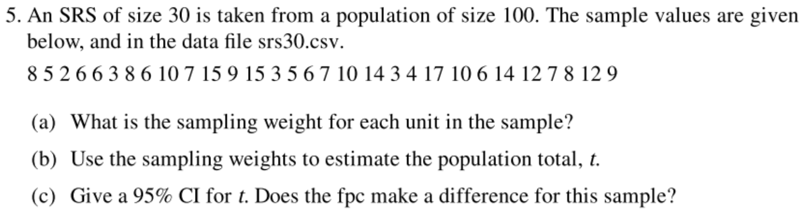
4. **Mean Squared Error (MSE($\bar{y}$)):**

$$\text{MSE}(\bar{y}) = V[\bar{y}] + \text{Bias}^2 = 19.61$$

1.4 Comparison

Plan 1 is better because it has no bias and a lower mean squared error (MSE = 18.94).

2 Question 5



5. An SRS of size 30 is taken from a population of size 100. The sample values are given below, and in the data file srs30.csv.

8 5 2 6 6 3 8 6 10 7 15 9 15 3 5 6 7 10 14 3 4 17 10 6 14 12 7 8 12 9

(a) What is the sampling weight for each unit in the sample?

(b) Use the sampling weights to estimate the population total, *t*.

(c) Give a 95% CI for *t*. Does the fpc make a difference for this sample?

Figure 2: This is the question 5

2.1 (a) Sampling Weight

The sampling weight for each unit is:

$$\text{Weight} = \frac{\text{Population Size}}{\text{Sample Size}} = \frac{100}{30} \approx 3.33$$

2.2 (b) Population Total Estimate (t)

The observed sample values are:

$$8, 5, 2, 6, 6, 3, 8, 6, 10, 7, 15, 9, 15, 3, 5, 6, 7, 10, 14, 3, 4, 17, 10, 6, 14, 12, 7, 8, 12, 9$$

The sum of the sample values is:

$$\sum y_i = 8 + 5 + 2 + \ldots + 9 = 232$$

The estimated population total is:

$$t = \text{Weight} \cdot \sum y_i = 3.33 \cdot 232 = 771.56$$

2.3 (c) 95% Confidence Interval for t

The sample mean is:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{232}{30} \approx 7.73$$

The sample variance is:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

The standard error is:

$$SE(t) = N \cdot \sqrt{\frac{S^2}{n}\left(1 - \frac{n}{N}\right)}$$

Using $N = 100$, $n = 30$, $\bar{y} = 7.73$, and S^2, the CI is:

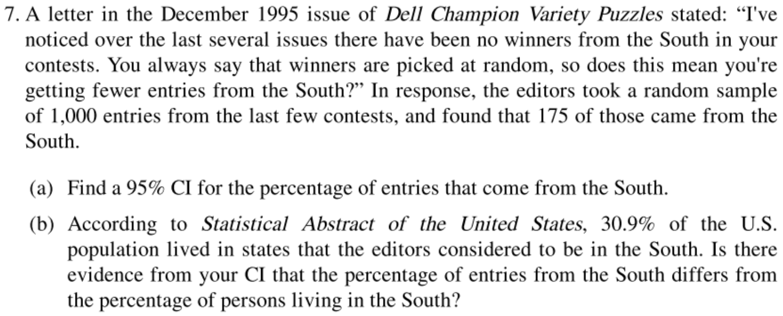
$$\text{CI} = \hat{t} \pm Z \cdot SE(t)$$

For $Z = 1.96$, calculate CI with and without the finite population correction (fpc):

$$\text{fpc} = \sqrt{1 - \frac{n}{N}} = \sqrt{1 - \frac{30}{100}} = \sqrt{0.7} \approx 0.837$$

Compare both results to evaluate the impact of the fpc.

3 Question 7



7. A letter in the December 1995 issue of *Dell Champion Variety Puzzles* stated: "I've noticed over the last several issues there have been no winners from the South in your contests. You always say that winners are picked at random, so does this mean you're getting fewer entries from the South?" In response, the editors took a random sample of 1,000 entries from the last few contests, and found that 175 of those came from the South.

(a) Find a 95% CI for the percentage of entries that come from the South.

(b) According to *Statistical Abstract of the United States*, 30.9% of the U.S. population lived in states that the editors considered to be in the South. Is there evidence from your CI that the percentage of entries from the South differs from the percentage of persons living in the South?

Figure 3: This is the question 7

3.1 (a) 95% Confidence Interval for Proportion

The proportion of entries from the South is:

$$\hat{p} = \frac{x}{n} = \frac{175}{1000} = 0.175$$

The standard error for the proportion is:

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Substituting the values:

$$SE = \sqrt{\frac{0.175 \cdot (1 - 0.175)}{1000}} = \sqrt{\frac{0.175 \cdot 0.825}{1000}} = \sqrt{0.000144375} \approx 0.01202$$

The critical value for a 95% confidence interval is $Z = 1.96$. The confidence interval is:

$$CI = \hat{p} \pm Z \cdot SE$$

Substituting the values:

$$CI = 0.175 \pm 1.96 \cdot 0.01202$$

$$CI = 0.175 \pm 0.02355$$

$$CI = (0.15145, 0.19855)$$

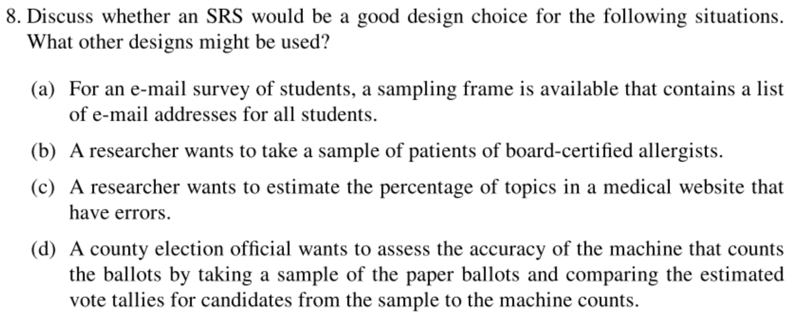
Thus, the 95% confidence interval for the percentage of entries from the South is approximately:

$$(15.15\%, 19.86\%)$$

3.2 (b) Comparison with U.S. Population Percentage

The percentage of the U.S. population living in the South is 30.9%. Since 30.9% does not fall within the confidence interval $(15.15\%, 19.86\%)$, there is strong evidence that the percentage of entries from the South differs from the percentage of persons living in the South.

4 Question 8



8. Discuss whether an SRS would be a good design choice for the following situations. What other designs might be used?

(a) For an e-mail survey of students, a sampling frame is available that contains a list of e-mail addresses for all students.

(b) A researcher wants to take a sample of patients of board-certified allergists.

(c) A researcher wants to estimate the percentage of topics in a medical website that have errors.

(d) A county election official wants to assess the accuracy of the machine that counts the ballots by taking a sample of the paper ballots and comparing the estimated vote tallies for candidates from the sample to the machine counts.

Figure 4: This is the question 8

4.1 (a) Email Survey of Students

Answer: SRS might not account for specific groups within the student body (e.g., year, major, or demographics). If these groups are important, stratified sampling might be better. We should use stratified Sampling as subsititute, i.e. divide students into groups (e.g., by year or department) and take a random sample from each group.

4.2 (b) Patients of Board-Certified Allergists

Answer: An SRS may not be ideal because patients could vary by location, allergist, or type of allergy. A single random sample might over-represent patients from one allergist or region. We either use cluster sampling(Randomly select board-certified allergists and sample all (or a random subset of) their patients.) or stratified sampling(Stratify by allergist, location, or patient type, then take random samples within each stratum.)

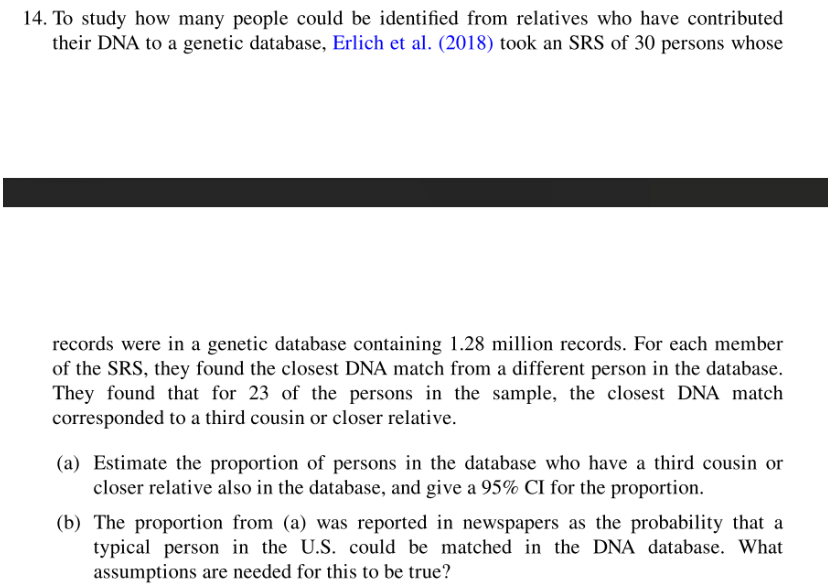
4.3 (c) Topics in a Medical Website

Answer: An SRS could work if every topic has an equal chance of being selected and if there is a complete list of topics. However, topics might vary in importance, size, or frequency of errors in this case. So we should use stratified sampling. (Group topics by type (e.g., symptoms, diseases, treatments) and sample within each group.)

4.4 (d) County Election and Machine Accuracy

Answer: An SRS could work but may not provide geographic representativeness. Ballots from one area might not reflect voting patterns across the county. We should use stratified sampling, which can divide ballots by precinct or geographic area and sample proportionally from each.

5 Question 14



14. To study how many people could be identified from relatives who have contributed their DNA to a genetic database, Erlich et al. (2018) took an SRS of 30 persons whose records were in a genetic database containing 1.28 million records. For each member of the SRS, they found the closest DNA match from a different person in the database. They found that for 23 of the persons in the sample, the closest DNA match corresponded to a third cousin or closer relative.

(a) Estimate the proportion of persons in the database who have a third cousin or closer relative also in the database, and give a 95% CI for the proportion.

(b) The proportion from (a) was reported in newspapers as the probability that a typical person in the U.S. could be matched in the DNA database. What assumptions are needed for this to be true?

Figure 5: This is the question 14

5.1 (a) Estimate the Proportion and Confidence Interval

The sample proportion is:

$$\hat{p} = \frac{x}{n} = \frac{23}{30} = 0.7667$$

The standard error is:

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Substitute the values:

$$SE = \sqrt{\frac{0.7667 \cdot (1-0.7667)}{30}} = \sqrt{\frac{0.7667 \cdot 0.2333}{30}} = \sqrt{\frac{0.1787}{30}} = \sqrt{0.00596} \approx 0.0772$$

The 95% confidence interval is:

$$CI = \hat{p} \pm Z \cdot SE$$

For $Z = 1.96$:

$$CI = 0.7667 \pm 1.96 \cdot 0.0772$$

$$CI = 0.7667 \pm 0.1513$$

$$CI = (0.6154, 0.9180)$$

Thus, the 95% confidence interval is approximately:

$$(61.54\%, 91.80\%)$$

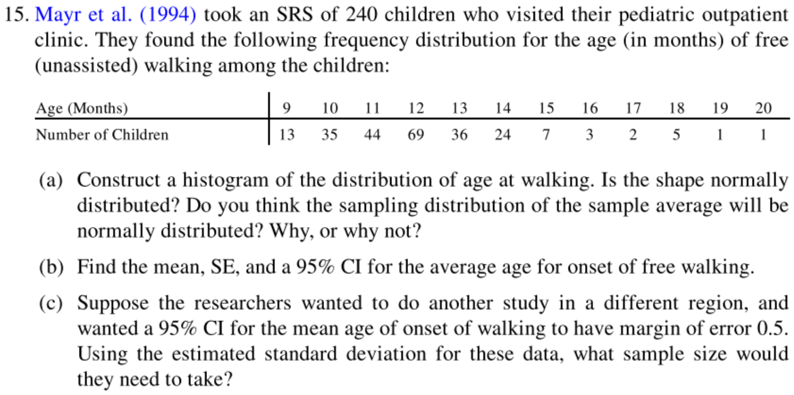
5.2 (b) Assumptions for Generalizing the Proportion

For the proportion $\hat{p} = 0.7667$ to represent the probability that a typical person in the U.S. could be matched in the DNA database, the following assumptions are required:

1. **Representative:** The sample of 30 persons must be a SRS and representative of the U.S. population.

2. **Independence:** The matches must be independent (no clustering effects, e.g., from family groups in the sample).

6 Question 15



15. Mayr et al. (1994) took an SRS of 240 children who visited their pediatric outpatient clinic. They found the following frequency distribution for the age (in months) of free (unassisted) walking among the children:

Age (Months)	9	10	11	12	13	14	15	16	17	18	19	20
Number of Children	13	35	44	69	36	24	7	3	2	5	1	1

(a) Construct a histogram of the distribution of age at walking. Is the shape normally distributed? Do you think the sampling distribution of the sample average will be normally distributed? Why, or why not?

(b) Find the mean, SE, and a 95% CI for the average age for onset of free walking.

(c) Suppose the researchers wanted to do another study in a different region, and wanted a 95% CI for the mean age of onset of walking to have margin of error 0.5. Using the estimated standard deviation for these data, what sample size would they need to take?

Figure 6: This is the question 15

6.1 (a) Histogram and Normality Assessment

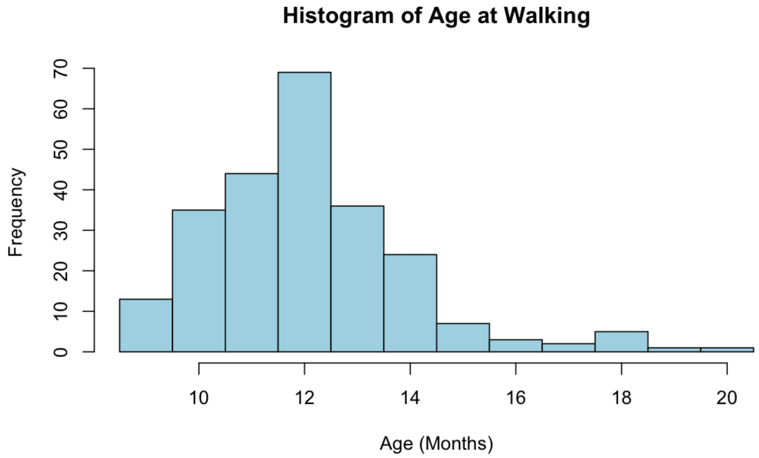


Figure 7: Histogram

The histogram of the age distribution is constructed using the frequency data provided. The distribution appears skewed due to the higher frequency of younger ages. However, according to the Central Limit Theorem, the sampling distribution of the sample mean will be approximately normal for large n.

6.2 (b) Mean, Standard Deviation, and 95% Confidence Interval

Mean ($\bar{x}$):

$$\bar{x} = \frac{\sum(x_i \cdot f_i)}{\sum f_i} = \frac{(9 \cdot 13) + (10 \cdot 35) + \ldots + (20 \cdot 1)}{240}$$

Variance and Standard Deviation (s):

$$s^2 = \frac{\sum f_i(x_i - \bar{x})^2}{n-1}$$

$$s = \sqrt{s^2}$$

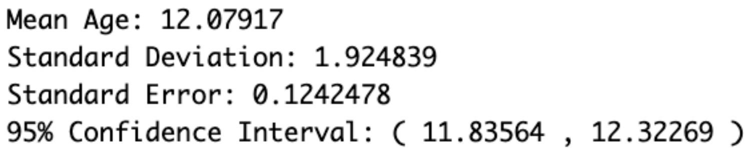
Standard Error (SE):

$$SE = \frac{s}{\sqrt{n}}$$

95% Confidence Interval (CI):

$$CI = \bar{x} \pm Z \cdot SE$$

According to R output, the results are:



```
Mean Age: 12.07917
Standard Deviation: 1.924839
Standard Error: 0.1242478
95% Confidence Interval: ( 11.83564 , 12.32269 )
```

Figure 8: R outputs

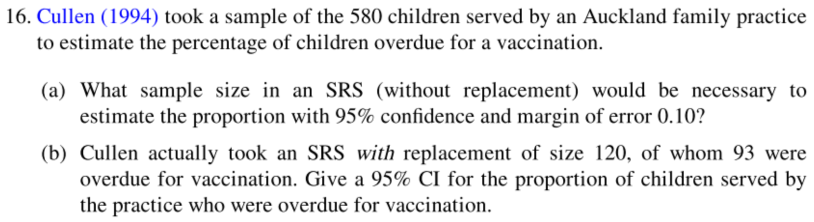
6.3 (c) Sample Size for Margin of Error 0.5

To achieve a 95% confidence interval with a margin of error of 0.5:

$$n = \left(\frac{Z \cdot s}{E} \right)^2$$

Substitute $Z = 1.96$, s (from above), and $E = 0.5$. We can get the result as 57.

7 Question 16



16. Cullen (1994) took a sample of the 580 children served by an Auckland family practice to estimate the percentage of children overdue for a vaccination.

(a) What sample size in an SRS (without replacement) would be necessary to estimate the proportion with 95% confidence and margin of error 0.10?

(b) Cullen actually took an SRS *with* replacement of size 120, of whom 93 were overdue for vaccination. Give a 95% CI for the proportion of children served by the practice who were overdue for vaccination.

Figure 9: This is the question 16

7.1 (a) Required Sample Size for Margin of Error 0.10

To determine the required sample size n in a Simple Random Sample (SRS) without replacement, we use the formula:

$$n = \frac{Z^2 \cdot \hat{p}(1 - \hat{p})}{E^2}$$

where:

- Z is the critical value for a 95% confidence level ($Z = 1.96$),
- $\hat{p}$ is the estimated proportion of children overdue for vaccination,
- E is the margin of error ($E = 0.10$).

Assuming no prior information about $\hat{p}$, we use $\hat{p} = 0.5$ to obtain the maximum sample size. Substituting the values:

$$n = \frac{(1.96)^2 \cdot 0.5 \cdot (1 - 0.5)}{(0.10)^2}$$

$$n = \frac{3.8416 \cdot 0.25}{0.01} = \frac{0.9604}{0.01} = 96.04$$

Since the sampling is without replacement, we apply the finite population correction (fpc):

$$n_{\text{adj}} = \frac{n}{1 + \frac{n-1}{N}}$$

Substitute $n = 96.04$ and $N = 580$:

$$n_{\text{adj}} = \frac{96.04}{1 + \frac{96.04-1}{580}} = \frac{96.04}{1 + \frac{95.04}{580}}$$

$$n_{\text{adj}} = \frac{96.04}{1.1639} \approx 82.55$$

Thus, the required sample size is approximately:

$$n \approx 83$$

7.2 (b) 95% Confidence Interval for Proportion

The confidence interval for the proportion is calculated as:

$$CI = \hat{p} \pm Z \cdot SE$$

where:

$$SE = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Given:

- Sample size $n = 120$,
- Number overdue $x = 93$,
- Proportion $\hat{p} = \frac{x}{n} = \frac{93}{120} = 0.775$,
- $Z = 1.96$ for a 95% confidence level.

Substitute into the formula for the standard error:

$$SE = \sqrt{\frac{0.775 \cdot (1 - 0.775)}{120}} = \sqrt{\frac{0.775 \cdot 0.225}{120}} = \sqrt{\frac{0.174375}{120}} = \sqrt{0.001453125} \approx 0.0381$$

Substitute into the confidence interval formula:

$$CI = 0.775 \pm 1.96 \cdot 0.0381$$

$$CI = 0.775 \pm 0.0747$$

$$CI = (0.7003, 0.8497)$$

Thus, the 95% confidence interval for the proportion is approximately:

$$(70.03\%, 84.97\%)$$