Problem Set 5

Due date: 23 October

Please upload your completed assignment to the ELMs course site (under the assignments menu). Remember to include an annotated script file for all work with R and show your math for all other problems (if applicable, or necessary). Please also upload your completed assignment to the Github repository that you have shared with us. *We should be able to run your script with no errors.*

Total points: 25

Question 1

Total points: 6

Use the data in the table below to answer the following questions.

Table 1: Voting by Age in 2000

Age group	Non-voters	Voters	Total
18-24	70	50	120
25-30	40	50	90
31 and up	220	570	790
TOTAL	330	670	1000

Part A

Points: 2

What is the probability of being 25-30 or a non-voter?

Ans A.:

Probability will be P(25-30) + P(Non-Voter) - P(Non-Voter and 25-30)

= 90/1000 + 330/1000 - 40/1000

= 380/1000 chances

Which is about 38% chance of being 25-30 or a non-voter

Part B

Points: 4

Assuming a normal distribution, report the 95% confidence intervals for the percentage of 18-to-24-year-olds who did not vote, and then the percentage of 25-to-30-year-olds who did not vote.

Ans B:

Assuming a normal distribution, and confidence interval of 95%

18-24 year old who did not vote: 70

For a proportion, the confidence interval can be given as,

$$Confidence = p + / - z * \sqrt{p * (1 - p)/n}$$

Here p = 70/1000 or 0.07%

Z score at 95% interval will be 1.96

N will be 1000

Hence the confidence interval will range from;

$$Confidence = 0.07 + / - 1.96 * \sqrt{0.07 * (1 - 0.07)/1000}$$

$$Confidence = 0.07 + / - 0.015$$

Hence the interval will be [0.085, 0.055]

Question 2

Total points: 7

Assume that the standard deviation for the population distribution of a state in which you want to conduct a poll is 200.

Part A

Points: 3

Calculate the spread of the sampling distribution for each of the following sample sizes: 1, 4, 25, 100, 250, 1000, 5,000, and 10,000.

Ans 2A:

SD = 200

Spread can be defined as the standard deviation or error of the sample space:

We can calculate the spread of a sample as:

$$SE = \sigma/\sqrt{n}$$

1. For a sample size of 1, the spread is 200/1 = 200
2. For a sample size of 4, the spread is 200/2 = 100
3. For a sample size of 25, the spread is 200/5 = 40
4. For a sample size of 100, the spread is 200/10 = 10
5. For a sample size of 250, the spread is 200/15.8 = 12.65 approx
6. For a sample size of 1000 the spread is 200/31.6 = 6.32 approx
7. For a sample size of 5000 the spread is 200/70.7 = 2.82 approx
8. For a sample size of 10,000 the spread is 200/100 = 2

As the sample size increases, we can see the error going down significantly

Part B

Points: 1

Describe specifically how the variability of the sampling distribution changes as the sample size varies. Considering the expense of running a poll, which sample size do you think is most optimal if conducting the poll?

Ans 2B:

As defined above , we can see that the sample error goes down the more the sample size increases, but we do hit a point of diminishing returns.

If we take

$$SE1 = \sigma/\sqrt{n}$$

and

$$SE2 = \sigma/\sqrt{n+1}$$

We can derive that

$$SE1/SE2 = \sqrt{(n+1)/n}$$

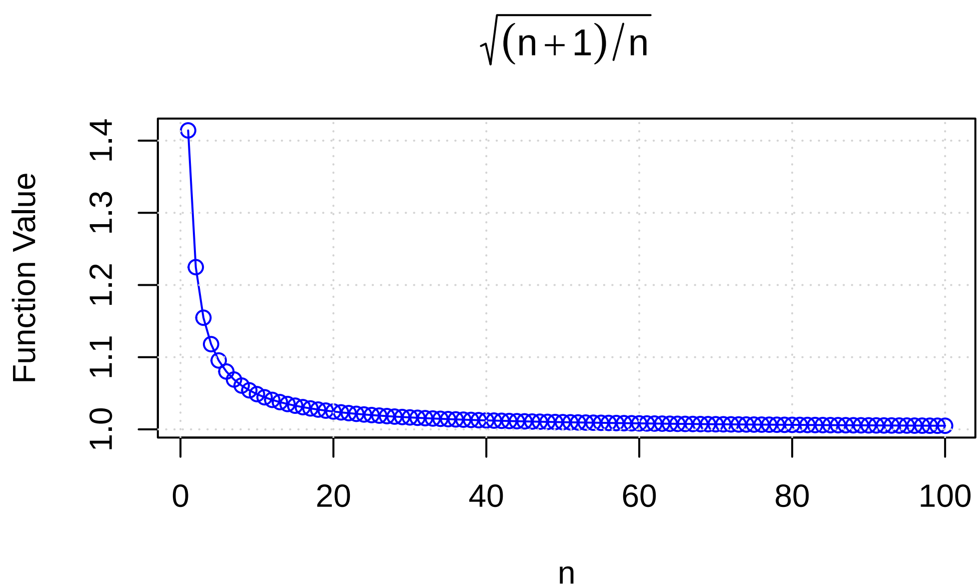
So a per unit increase in n does not drastically change the difference of the SE values.

```r
n <- c(1:100)

# Calculate the function values
result <- sqrt((n + 1) / n)

# Create a plot
plot(n, result, type = "o", col = "blue", xlab = "n", ylab = "Function Value",
     main = expression(sqrt((n + 1)/n)))

# Add gridlines
grid()
```



I would keep a drop off point of 1000 as it drops off after that.

Standard literature says 40. I am assuming that the expenses are telephone calls which is not very high for a sample size of this much.

Part C

Points: 3

Display your results graphically (using R) with the sample size on the x-axis and the standard error (of the sampling distribution) on the y-axis.

Ans 2C:

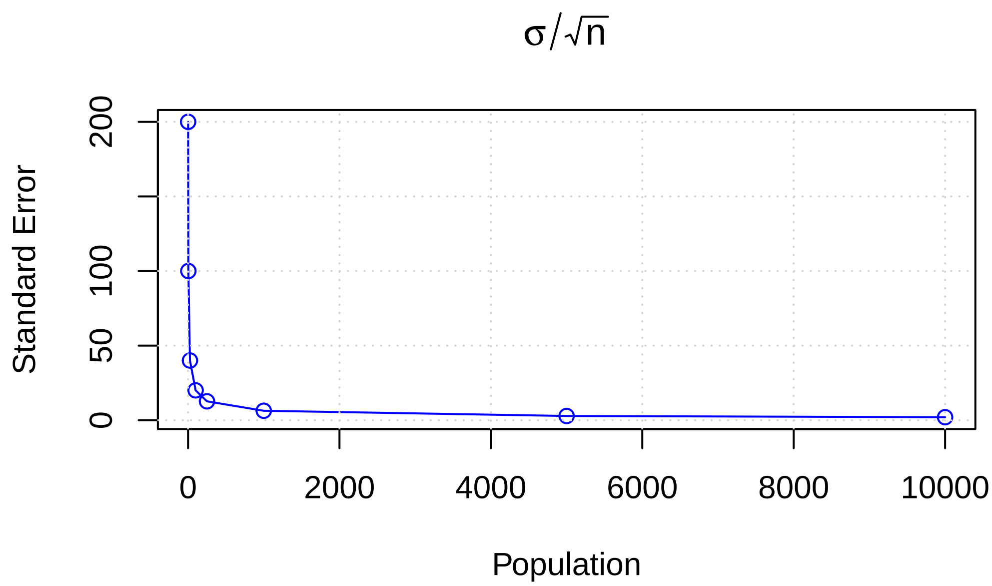
Just using the same logic as the above question,

```
n <- c(1, 4, 25, 100, 250, 1000, 5000, 10000)
sigma = 200

# Calculate the function values
result <- sigma/sqrt(n)

# Create a plot
plot(n, result, type = "o", col = "blue", xlab = "Population", ylab = "Standard Error",
     main = expression(sigma/sqrt(n)))


# Add gridlines
grid()
```



Question 3

Points: 4

Suppose you conduct a survey (to generate a sample mean of interest) and find that it has a margin of error of 4.5 with a sample size of 900 using a 95% confidence interval. What would the margin of error be for a 90% confidence interval?

Ans 4:

Margin of error = 4.5

Sample Size = 900

Confidence interval = 95%, which means the z-score = 1.96

Confidence interval = 90%, which means the z-score = 1.64

We can say that z_90/z_95 = ME_90/ME_90

Or 1.64/1.96 * 4.5 = ME_90

Or ME_90 = 3.76

Hence the margin of error will be 3.79 at 90% confid

Question 4

Points: 4

Assume that, in State A, the mean income in the population is $20,000 with a standard deviation of $2,000. If you took an SRS of 900 individuals from that population, what is the probability that you would get a sample mean income of $20,200 or greater? What would be the probability if the sample size was only 25?

> **Note**
>
> Assume a normal distribution for both questions.

Ans 4:

A_mean = 20000

A_sd = 2000

N = 900

Sample mean income = 20,200 or greater

Z score 1 = 3

Z score 2 = 4

We can write a Z score as:

$$Z = \overline{x} - \mu / sigma / \sqrt{n}$$

Or

$$Z = 20,200 - 20,000/2000/\sqrt{900}$$

Or,

$$Z = 200/66.6$$

Or approx 3

The probability if the Z score is 3 is 1-0.99865 = 0.00135. Which is a very small number for a single tailed test.

If the sample size is 25,

$$Z = 20,200 - 20,000/2000/\sqrt{25}$$

Or,

$$Z = 200/400$$

Or approx 0.5

The probability if the Z score is 0.5 is 1-0.69 = 0.37. Which is a reasonably higher but still quite small.

Question 5

Points: 4

Assume that a coin is fair. If I flip a coin 500 times, what is a 95% confidence interval for the range of the count of heads that I will get? What if I flip the coin 5,000 times? What about 50,000 times?

Ans 5:

For a binomial distribution, the equation can be represented as:

$$Confidence = p + / - z * \sqrt{p * (1 - p)/n}$$

Where in this case,

n = 500

Confidence interval = 95% or Z -> 1.96

Hence Confidence interval :

$$Confidence = 0.5 + / - 1.96 * \sqrt{0.5 * (1 - 0.5)/500}$$

Hence you should get it in the range of [0.402, 0.598]

If n = 5,000

$$Confidence = 0.5 + / - 1.96 * \sqrt{0.5 * (1 - 0.5)/5000}$$

Or a range of [0.486, 0.5138]

If n = 50,000

$$Confidence = 0.5 + / - 1.96 * \sqrt{0.5 * (1 - 0.5)/50000}$$

Or a range of [0.495, 0.504]

Hence we can see that there is a diminishing return in the data that we get if the sample size is increased.