STA371G Homework Assignment 3: Solutions

Problem 1

Read the report of Anzalone Liszt Grove Research on the NSA Surveillance Programs (http://www.algpolling.com/#!memos/cjna): "The public strongly supports... This memo is based on the results of a national poll conducted by Anzalone Liszt Grove Research. The poll of N=803 adults was conducted November 11-17, 2013. At least 30% of all interviews were conducted via cell phone and interviews were conducted in English and Spanish. The margin of error for the poll is plus or minus 3.5 percentage points at the 95% level of confidence."

(a) Can you explain why the margin of error for the poll is $\pm 3.5\%$ at the 95% level of confidence? (hint: the value of $p(1-p)$ is largest when $p = 0.5$)

The standard error associated with the sample proportion $\hat{p}$ as an estimate of the true proportion p is

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

where n is the sample size. We estimate the population proportion p with the sample proportion $\hat{p}$, and the (approximate) 95% confidence interval is

$$\hat{p} \pm 2\sqrt{\frac{p(1-p)}{n}}.$$

The margin of error is the largest when $p = 0.5$, which is

$$\pm 2\sqrt{\frac{0.5(1-0.5)}{n = 803}} = \pm 3.5\%$$

at the 95% level of confidence.

(b) If Anzalone Liszt Grove Research conducts another poll of $N = 8,000$ adults, what would be the margin of error at the 95% level of confidence?

The margin of error is the largest when $p = 0.5$, which is about

$$\pm 2\sqrt{\frac{0.5(1-0.5)}{n = 8000}} = \pm 1.1\%$$

at the 95% level of confidence.

Problem 2

According to the official Federal Election Commission report for the presidential election in 2012, out of a total of 7,993,851 votes in Texas, President Barack Obama received 3,308,124 votes.

(a) If you randomly survey 1000 Texas residents who had voted in the 2012 presidential election, can you predict the distribution of the number of votes for President Obama among these 1000 Texas voters? Will you be surprised to find out that more than 500 of them voted for President Obama? (Hint: using the normal approximation to the binomial distribution.)

The proportion of votes for President Obama in Texas in the 2012 election is

$$p = \frac{3,308,124}{7,993,851} = 0.414 \text{ (or } 41.4\%).$$

As the number of voters $7,993,851$ is much larger than the sample size $n = 1000$, we can consider the number of votes for President Obama in a random sample of $n = 1000$ to be binomial distributed as

$$X \sim \text{Binomial}(n = 1000, p = 0.414).$$

Since n is large and p is close to neither 0 nor 1, we may safely approximate the binomial distribution with a normal distribution with mean np and variance $np(1-p)$, which can be expressed as

$$X \sim \mathcal{N}(np, np(1-p)) = \mathcal{N}(413.8, 15.6^2).$$

Thus we are 95% confident that the number of votes for President Obama from a random sample of $n = 1000$ Texas voters is between

$$[382, 445]$$

Note that $X \geq 500$ is outside $[382, 445]$ and $\frac{500-413.8}{15.6} = 5.5$. As 500 is over 5 standard deviation away from the mean, $P(X \geq 500)$ would be very close to zero. Thus we would be very surprised to find out $X \geq 500$.

(b) According to http://www.politico.com/2012-election/results/president/texas/, In Dallas County, TX, President Obama received 57.1% of the votes in the 2012 presidential election. If you randomly survey 100 residents of Dallas County who had voted in the 2012 presidential election, can you predict the distribution of the number of votes for President Obama among them? Will you be surprised to find out no more than 50 votes for President Obama?

Similar to the analysis for Problem 6.(b), the number of votes X can be considered distributed as

$$X \sim \text{Binomial}(n = 100, p = 0.571)$$

Approximately, we have

$$X \sim \mathcal{N}(100 * 0.571, 100 * 0.571 * (1 - 0.571)) = \mathcal{N}(57.1, 4.95^2).$$

Thus we are 95% confident that the number of votes for President Obama from a random sample of $n = 100$ Dallas County voters is between

$$[47, 67]$$

Note that $X = 50$ is between [47, 67]. Further calculating $P(X \leq 50)$ using

$$pnorm(50, 57.1, 4.95)$$

in R, or

$$1 - NORMDIST(50, 57.1, 4.95, TRUE)$$

in Excel, we find that

$$P(X \leq 50) = 7.6\%.$$

As $P(X \leq 50)$ happen with a probability that is not small, we are usually not surprised to find out that X is no more than 50.

Problem 3 (50 points)

(a) Use R/Excel to simulate 1000 normal random numbers with mean 0.5 and standard deviation 0.6. Record the sample mean and sample variance of these 1000 simulated random numbers.
Sample mean = 0.4841. Sample variance = 0.3813.

```
# R code:
set.seed(12) # set the seed of random number generators.
x = rnorm(1000, mean = 0.5, sd = 0.6)
print(mean(x))
print(var(x))
```

(b) Simulate a random number u that is uniformly distributed between 0 and 1, record its value and write it down:
$u = 0.2655087$

```
# R code:
set.seed(1) # set the seed of random number generators.
u = runif(1)
print(u)
```

(c) Calculate the probability that $X \sim \mathcal{N}(0.5, 0.6^2)$ is larger than u obtained in (b).
Given $X \sim \mathcal{N}(0.5, 0.6^2)$ and $u = 0.2655087$, $P(X > u) = 0.6520$.

```
# R code:
print( 1 - pnorm(u, mean = 0.5, sd = 0.6) )
```

(d) Use these 1000 normal random numbers obtained in (a) to find an approximate answer to (c), record its value and write it down:
Calculate the proportion of the 1000 values of X that are greater than u. The proportion is an approximated probability of $P(X > u)$. $P(X > u) \approx 0.66$.

```
# R code:
print( mean(x>u) )
```

(e) Designing a simulation procedure to more accurately approximate $P(X > u)$.
Simulate more than 1000 $\mathcal{N}(0.5, 0.6^2)$ distributed random numbers. Calculate the proportion of theses values that are greater than u. For example, simulate 10,000 such random numbers. $P(X > u) \approx 0.6527$.

```
# R code:
set.seed(12) # set the seed of random number generators.
xx = rnorm(10000, mean = 0.5, sd = 0.6)
print( mean(xx>u) )
```

(f) Find out that X is smaller than which number with probability u (note u is between 0 and 1 and hence can be treated as a probability).
Find a such that $P(X < a) = u$. $a = 0.1241$.

```
# R code:
print(qnorm(u, mean = 0.5, sd = 0.6))
```

(g) Using these 1000 normal random numbers obtained in (a) to find an approximate answer to (f), record its value and write it down:
Sort the 1000 values by ascending order. Find the $1000 \times u \approx 266$th smallest number, which is 0.1425.

```
# R code:
print(sort(x)[266])
```

(h) Given the 1000 random numbers obtained in (a), provide your 95% confidence interval of the true mean of the probability distribution that these 1000 random numbers are simulated from.
By the central limit theorem, $\bar{X} \xrightarrow{d} \mathrm{N}(\mu, \sigma^2/n)$ as $n \to \infty$. The 95% confidence interval of μ is $(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}})$. If we know the true standard deviation $\sigma = 0.6$,, the CI= $(0.4469, 0.5213)$. Otherwise, we estimate σ by the sample standard deviation, and the CI= $(0.4485, 0.5198)$.

(i) To reduce the width of the 95% confidence interval of the true mean in (h) by about 50%, how many more random samples from the underlying distribution are needed?
For simplicity, suppose we know the true $\sigma = 0.6$. Let m be the size of the random samples that reduce the width of the 95% confidence interval by 50%.
$\frac{\sigma}{\sqrt{m}} / \frac{\sigma}{\sqrt{1000}} = 0.5 \to m = 4000$. So we need 3000 more random samples.

(j) Simulate these additional random numbers and combing them with the original 1000 random numbers to find the sample mean and sample variance, and provide your updated 95% confidence interval of the true mean.
The 95% CI = $(0.4832, 0.5204)$.

```r
# R code:
set.seed(123)
xx = c(x, rnorm(3000, mean = 0.5, sd=0.6))
print(paste(mean(xx)-1.96*0.6/sqrt(4000), mean(xx)+1.96*0.6/sqrt(4000)))
```