Problem Section 5

Interval Estimation

Learning Outcomes

The problems are designed to build conceptual understanding and problem-solving skills. The emphasis is on learning to find, evaluate and build confidence. The specific tasks include:

- Calculate probabilities using the Central Limit Theorem.
- Construct a confidence interval using Chebychev's inequality
- Calculate the sample size for a desired margin of error
- Back up and support work with relevant explanations

Instructions

- Please work in groups of three to answer the problems below.
- Each member of the group must write up their work on a separate sheet of paper. Be sure to clearly write your full name on the top as it appears in the gradebook.

Exercises

1. In a grocery store, 400 customers shop everyday. The number of half-gallons of nonfat milk bought by a randomly selected customer is a random variable X having PMF

$$f(x) = \begin{cases} 0.3 & x = 0 \\ 0.5 & x = 1 \\ 0.2 & x = 2 \end{cases}$$

Assume buying behaviors of different customers are independent. The grocer requests 390 half gallon containers per day from the distributor. Calculate, using the Central Limit Theorem, the probability that, on a given day that is enough. (Please practice defining random variables and setting up the problem mathematically before solving for a probability)

Let $X_1, X_2, \ldots, X_{400}$ denote the number of half-gallons purchased by the 400 customers. We are interested in calculating $P(S \leq 390)$ where $S = X_1 + X_2 + \ldots, +X_{400}$.

By the Central Limit Theorem for sums, we know that $S \approx N(mean = n\mu_0, sd = \sigma_0\sqrt{n})$ where $\mu_0 = E\,[X]$ and $\sigma_0 = SD\,[X]$. In other words, μ_0 and σ_0 are the mean and standard deviation of an individual X.

We have $\mu_0 = E[X] = 0 \times .3 + 1 \times .5 + 2 \times .2 = .9$.

To find the variance, let us first find the second moment, $E[X^2]$.

$E[X^2] = 0^2 \times .3 + 1^2 \times .5 + 2^2 \times .2 = 1.3$

So $\sigma_0^2 = Var[X] = 1.3 - (.9)^2 = .49$

and we get

$$S \approx Norm(mean = 360, sd = 14)$$

The probability that the grocer bought enough half-gallons of milk is calculated below using the approximate normal distribution for S:

```
pnorm(q = 390, mean = 360, sd = 14)  #P(S <= 390) without standardizing
```

```
## [1] 0.9839377
```

```
pnorm (q = (390-360)/14, mean = 0, sd = 1)  #P(S <= 390) with standardizing
```

```
## [1] 0.9839377
```

2. A manufacturer of toothpicks wonders what the mean width of a toothpick is under a new manufacturing method. For both parts below, assume the conditions of the Central Limit Theorem are satisfied, and note that toothpicks produced under the old method had a standard deviation of 0.4 mm and that this is unchanged with the new process.

a. Calculate a 90% large sample confidence interval estimate for the true mean based on a sample of size $n = 25$ and $\bar{x} = 1.68$ mm. Write your formulas first and then do calculations in a code chunk in R.

The formula for the 90% large sample confidence interval estimator for the true mean width is $\bar{X} \pm z_{0.05} \frac{\sigma_0}{\sqrt{n}}$ where $z_{0.05}$ is the 5th percentile of a standard normal distribution. The calculations for this particular sample are shown below.

```
z_05 <- qnorm(p = 0.05, mean = 0, sd = 1) ; z_05
```

```
## [1] -1.644854
```

```
n <- 25
xbar <- 1.68
sigma0 <- 0.4

xbar + c(-1,1)*z_05*sigma0/sqrt(n)
```

```
## [1] 1.811588 1.548412
```

We are 90% confident that the true mean width is in the range 1.81 and 1.55.

b. How many toothpicks must the manufacturer measure for the margin of error for 90% confidence to be no larger than 0.10 mm?

We want to find n so that the half length of a 90% (large sample) confidence interval for the mean width is less than or equal to 0.1 mm. Therefore we want to solve the equation

$$\begin{aligned} z_{.05} \times \frac{\sigma_0}{\sqrt{n}} &\leq 0.1, \\ n &\geq 1.645^2 \times \frac{0.4^2}{0.1^2}, \\ &\geq 43.296 \end{aligned}$$

Therefore, we need to measure $n = 44$ toothpicks.

To verify that the half length of the 90% confidence interval for the mean width is no larger than 0.1mm for this value of n, we can do the following calculation.

```
1.645 * 0.4/sqrt(44)
```

```
## [1] 0.09919723
```

3. Suppose $X_1, X_2, \dots, X_n \overset{i.i.d.}{\sim} Norm(\mu_0, 1)$. Then $\bar{X} \sim Norm\left(\mu_0, \frac{1}{\sqrt{n}}\right)$ and a 95% (exact) confidence interval for μ_0 is the interval $\bar{X} \pm \frac{1.96}{\sqrt{n}}$. This means that the probability that the true mean μ_0 will fall in this interval is 95%.

   Let p denote the probability that an additional independent observation - X_{n+1} - drawn from the normal distribution will fall in this interval.

a. Is p greater than, less than or equal to 0.95?

$$\begin{aligned} p &= P\left(\bar{X} - \frac{1.96}{\sqrt{n}} \le X_{n+1} \le \bar{X} + \frac{1.96}{\sqrt{n}}\right), \\ &= P\left(-\frac{1.96}{\sqrt{n}} \le (\bar{X} - X_{n+1}) \le \frac{1.96}{\sqrt{n}}\right). \end{aligned}$$

Then using Theorem 3.3 with $a = 1$, $b = -1$ and $c = 0$, we have the result that $\bar{X} - X_{n+1} \sim Norm\left(mean = 0, sd = \sqrt{\frac{1}{n} + 1}\right)$.

Therefore

$$\begin{aligned} p &= P\left(-\frac{\frac{1.96}{\sqrt{n}}}{\sqrt{\frac{1}{n}+1}} \le \frac{(\bar{X} - X_{n+1})}{\sqrt{\frac{1}{n}+1}} \le \frac{\frac{1.96}{\sqrt{n}}}{\sqrt{\frac{1}{n}+1}}\right), \\ &= P\left(-\frac{\frac{1.96}{\sqrt{n}}}{\sqrt{\frac{1}{n}+1}} \le Z \le \frac{\frac{1.96}{\sqrt{n}}}{\sqrt{\frac{1}{n}+1}}\right). \end{aligned}$$

Since $(\frac{1}{n} + 1)$ is larger than 1, we have that $\sqrt{\frac{1}{n}+1}$ is also larger than 1 and therefore $\frac{1}{\sqrt{\frac{1}{n}+1}} < 1$. Hence

$$\frac{\frac{1.96}{\sqrt{n}}}{\sqrt{\frac{1}{n}+1}} < \frac{1.96}{\sqrt{n}}.$$

Hence we can see that p should be less than 95% since it is the area over a smaller interval than $\pm\frac{1.96}{\sqrt{n}}$ under a standard normal.[1]

b. How do you explain your answer from part a?

   It is not surprising that $p < 95\%$ since the interval $\bar{X} \pm \frac{1.96}{\sqrt{n}}$ is a confidence interval for a parameter value (fixed number). It is not a prediction interval. It does not have the same chance of capturing another random variable drawn from the distribution. Random variables have variability (unlike parameters) and the confidence interval only accounts for the variability in $\bar{X}$, not in X_{n+1}.

4. Suppose $X_1, X_2, \dots, X_n \overset{i.i.d.}{\sim} Unif(0, \theta_0)$. In this problem we will construct a 95% confidence interval for θ_0 using $X_{max} = \max\{X_1, X_2, \dots, X_n\}$ and its sampling distribution

$$f_{X_{max}}(x) = \frac{n\, x^{n-1}}{\theta_0^n} \quad 0 \le x < \theta_0.$$

a. Let q_α denote the α quantile of the PDF $f_{X_{max}}$. That is q_α is the number such that $P(X_{max} \le q_\alpha) = \alpha$.[2] Show that $q_\alpha = \theta_0\, \alpha^{1/n}$.

[1] Students: please plug in n = 10 in the above derivation and see why this is true if it is confusing

[2] As an example, 1.96 is the 0.975 quantile of a standard normal distribution. It is the number that has 97.5% area below it Therefore $q_{0.975} = 1.96$ for the standard normal.

We know that for $0 \le x < \theta_0$, the CDF of X_{max} is $P(X_{max} \le x) = \int_0^x f_{max}(u)du = \int_0^x \frac{n\,x^{n-1}}{\theta_0^n}dx = \frac{x^n}{\theta_0^n}$

Thus we are solving the equation:

$$P(X_{max} \le q_\alpha) = \frac{q_\alpha^n}{\theta_0^n} = \alpha$$

for q_α and we we simply get $q_\alpha = (\alpha\theta_0^n)^{\frac{1}{n}} = \theta_0(\alpha)^{\frac{1}{n}}$.

b. "Invert" the probability statement

$$P\left(q_{0.025} \le X_{max} \le q_{0.975}\right) = 0.95$$

to construct a 95% confidence interval for θ_0.

Hint: this just means rearrange terms inside the probability so you move θ_0 to the middle of the inequalities.

We have the expression:

$$P\left(q_{0.025} \le X_{max} \le q_{0.975}\right) = P(\theta_0(0.025)^{1/n} \le X_{max} \le \theta_0(0.975)^{1/n})$$

Remember, we wish to rework this inequality so θ_0 is in the middle. A simple thing to do here to remove θ_0 from the ends, is to divide the expression by θ_0.

This yields:

$$\begin{aligned} P(0.025)^{1/n} \le \frac{X_{max}}{\theta_0} \le (0.975)^{1/n}) &= P((0.025)^{-1/n} \ge \frac{\theta_0}{X_{max}} \ge (0.975)^{-1/n}), \\ &= P(X_{max}(0.025)^{-1/n} \ge \theta_0 \ge X_{max}(0.975)^{-1/n}) = .95 \end{aligned}$$

Thus:

$$P(X_{max}(0.975)^{-1/n} \le \theta_0 \le X_{max}(0.025)^{-1/n}) = .95$$

and $[X_{max}(0.975)^{-1/n}, X_{max}(0.025)^{-1/n}]$ is a 95% confidence interval estimator for θ_0.

c. Suppose our sample values are $x_1 = 0.2, x_2 = 0.9, x_3 = 1.9, x_4 = 2.2, x_5 = 4.7, x_6 = 5.1$. Calculate your interval for θ_0 from part b. for this sample.

In this case we have that n=6 and $x_{max} = 5.1$. Thus plugging in these values for our answer above we get:

$$(5.1 \times (.975)^{-1/6}, 5.1 \times (.025)^{-1/6}) = (5.122, 9.431)$$

```
lower <- 5.1*0.975^(-1/6); lower
```

```
## [1] 5.121566
```

```
upper <- 5.1*(0.025)^(-1/6); upper
```

```
## [1] 9.431487
```