MSIS 5503 – Statistics for Data Science – Fall 2020 - Assignment 6

Solution

1. The length of a maternity stay in a U.S. hospital is said to be normally distributed with mean of 2.4 days with a standard deviation of 0.9 days. We randomly survey 80 women who recently bore children in a U.S. hospital.
   a. In words, what is $\overline{X}$? **Mean/Average length of maternity stay in a US Hospital (in days) in the sample.**
   b. Did you need to use the Central Limit Theorem for this example? Why or why not? **We don't need the CLT because the population is normally distributed.**
   c. **$\overline{X} \sim N(2.4, 0.9/\sqrt{80} = 0.1)$ because the population standard deviation is known.**
   d. What is the expected value of $\overline{X}$? 2.4 days Is $\overline{X}$ a biased or unbiased estimator? **$\overline{X}$ is always an unbiased estimator of μ the population mean, by theory that $E(\overline{X}) = \mu$.**
   e. What is the value of the standard error? Explain the meaning of the standard error in words. **The value of the standard error is 0.1, and represents the standard deviation of the sampling distribution of $\overline{X}$. It captures the variability of $\overline{X}$ across samples and is a measure of the precision of $\overline{X}$ as an estimator of μ.**
   f. What is the probability that *a woman stayed* more than 3 days in a hospital?
   To calculate this we use the population distribution $X \sim N(2.4, 0.9)$, convert it to a z-value and get the probability $P(X > 3) = P(Z > (3 – 2.4)/0.9) = P(Z > 0.6667) = 0.252$
   g. What is the probability that the *average length of stay of 80 women* in the hospital is more than 3 days?
   To calculate this we use the sampling distribution $\overline{X} \sim N(2.4, 0.1)$, convert it to a z-value and get the probability $P(\overline{X} > 3) = P(Z > (3 – 2.4)/0.1) = P(Z > 6) = 0.0001$

2. Suppose that 14 children, who were learning to ride two-wheel bikes, were surveyed to determine how long they had to use training wheels. The survey revealed that they used them an average of six months with a sample standard deviation of three months. *Assume that the underlying population distribution is normal.*
   a. In words, what is $\overline{X}$? **Mean length of time (in months) to learn to ride two-wheel bikes in children in the sample.**
   b. Did you need to use the Central Limit Theorem for this example? Why or why not? **We don't need the CLT because the population is normally distributed.**
   c. **$\overline{X} \sim t(13\ df)$ because the sample size is small (< 30)**
   d. What is the value of the standard error? Explain the meaning of the standard error in words.
   The value of the standard error is $3/\sqrt{14}$, and represents the standard deviation of the sampling distribution of $\overline{X}$. It captures the variability of $\overline{X}$ across samples and is a measure of the precision of $\overline{X}$ as an estimator of μ.
   e. Construct a 99% confidence interval for the population mean length of time using training wheels.
   For 99% CI, $\alpha = 0.01$ and $\alpha/2 = 0.005$, $t_{13,0.005} = 3.0123$.

The EBM = $(3/\sqrt{14})*3.0123 = 2.415$.
The 99% CI = (6 – EBM, 6 + EBM) = (3.585, 8.415)

f. Interpret this Confidence Interval
For all possible random samples of size 14, if we constructed the confidence interval as we did for the first sample, 99% of such confidence intervals will contain the true population mean of mean length of time (in months) that children take to learn to ride a two-wheel bike.

g. What is the error bound for the mean (EBM)?
EBM = $(3/\sqrt{14})*3.0123 = 2.415$

h. Repeat d, e and f for a 95% CI.
For 95% CI, $\alpha = 0.05$ and $\alpha/2 = 0.025$, $t_{13,0.025} = 2.1604$.
The EBM = $(3/\sqrt{14})*2.1604 = 1.7322$.
The 95% CI = (6 – EBM, 6 + EBM) = (4.2678, 7.7322)

i. Explain why the 99% CI is larger than the 95% CI.
The 99% CI is larger than the 95% CI because if we want an interval with higher confidence, it will necessarily be wider due to a larger EBM. It says that the larger the confidence *level*, wider, or less precise, the Confidence *Interval*.

3. A survey of the mean number of cents off that coupons give was conducted by randomly surveying one coupon per page from the coupon sections of a recent San Jose Mercury News. The following data were collected: 22¢; 76¢; 54¢; 69¢; 31¢; 51¢; 41¢; 41¢; 30¢; 55¢; $1.40; 40¢; 66¢; 40¢. *Assume the underlying distribution is approximately normal.*

a. In words, what is $\overline{X}$? **Mean value of coupons (in cents) in the coupon section of San Jose Mercury News, in the sample.**

b. Did you need to use the Central Limit Theorem for this example? Why or why not?
We don't need the CLT because the population is normally distributed.

c. What is the value of the sample standard deviation (s)? Explain its meaning in words.
The sample standard deviation = 29.0733; It shows how much the observations vary within this single sample.

d. $\overline{X} \sim t(13\ df)$ because the sample size is small (< 30)

e. What is the value of the standard error? Explain the meaning of the standard error in words.
The standard error is $29.0733/\sqrt{14} = 7.7701$ and represents the standard deviation of the sampling distribution of $\overline{X}$. It captures the variability of $\overline{X}$ across samples and is a measure of the precision of $\overline{X}$ as an estimator of μ.
Find a 95% CI for the population mean worth of coupons.
For 95% CI, $\alpha = 0.05$ and $\alpha/2 = 0.025$, $t_{13,0.025} = 2.1604$.
The EBM = $(7.7701)*2.1604 = 16.787$.
$\overline{X} = 54.2143$ cents.
The 95% CI = (54.2143 – EBM, 54.2143 + EBM) = (37.427, 71.001)

f. Interpret the 95% CI for the population mean worth of coupons.
For all possible random samples of size 14, if we constructed the confidence interval as we did for the first sample, 95% of such confidence intervals will contain the true population mean worth of coupons (in cents) in the coupon section of San Jose Mercury News

4. In a recent sample of 84 used car sales costs, the sample mean was \$6,425. Assume that we know that the population standard deviation of used car sales cost is \$3,156. We wish to estimate the true population mean of used car sales cost.

   a. In words, what is $\overline{X}$? **Mean sales cost of used cars (in dollars) in the sample.**

   b. Do you need to use the Central Limit Theorem in this case? Why or why not?
   Yes, we need the CLT because we do not know the population distribution, we know the population standard deviation and the sample size is greater than 30.

   c. **$\overline{X} \sim N(6425, 3156/\sqrt{84} = 344.34)$**

   d. What is the value of the standard error? Explain the meaning of the standard error in words.
   The standard error is 344.34 and represents the standard deviation of the sampling distribution of $\overline{X}$. It captures the variability of $\overline{X}$ across samples and is a measure of the precision of $\overline{X}$ as an estimator of μ.

   e. Find a 95% CI for the population mean cost of a used car.
   For 95% CI, $\alpha = 0.05$ and $\alpha/2 = 0.025$, $z_{0.025} = 1.96$.
   The EBM = 1.96 * 344.24 = 674.92.
   The 95% CI = (\$6,425 – \$674.92, \$6,425 + \$674.92) = (\$5,750.08, \$7,099.92)

   f. If they need a 92% CI with an EBM of \$500 for the population mean cost of a used car, what is the minimum sample size they should use?
   For EBM=500 and for 92% CI, $\alpha = 0.08$ and $\alpha/2 = 0.04$, $z_{0.04} = 1.75$.
   Hence $(3156/\sqrt{n})*1.75 = 500$.
   So, $\sqrt{n} = 11.116$ and n = 122.

5. Public Policy Polling recently conducted a survey asking adults across the U.S. about music preferences. When asked, 80 of the 571 participants admitted that they have illegally downloaded music.

   a. What is the sample statistic and what is its distribution? Include the parameters of this distribution.
   The sample statistic here is the sample proportion p’ of adults across the U.S. who have illegally downloaded music’ p’ = 80/571 = 0.14.

   We are using the normal approximation to the binomial, since, n>=30, np’>=5, nq’>=5.
   $p' \sim N(p, \sqrt{((p'(1-p'))/n)})$, where p’ =0.14, 1-p’=0.86, n=571.
   $p' \sim N(p, 0.015)$

   b. Construct a 99% confidence interval the population percent of U.S. adults who have illegally downloaded music.
   EBM = $Z_{.005} * \sqrt{((p'(1-p'))/n)} = 2.575 * \sqrt{((0.14(0.86))/571)} = 2.575 * 0.015 =$ 0.039.
   The 99% CI = (0.14-0.039, 0.14+0.039) = (0.101, 0.179)

   c. Interpret the above confidence interval. i.e., in your own words explain what the confidence interval means.

If we take repeated samples, and for each sample we construct the confidence interval for proportion as above, then 99% of those confidence intervals will contain the true population proportion.

d. What is the minimum number you would need to survey to be 95% confident that the population proportion (who have illegally downloaded music) is estimated to within 0.03?

EBM = $Z_{.025}$ * $\sqrt{}$((0.14(0.86))/n) = 1.96*$\sqrt{}$((0.14(0.86))/n) <= 0.03
$\sqrt{}$n >= (1.96/0.03) *$\sqrt{}$(0.14(0.86))
n >= $(1.96/0.03)^2$(0.14)(0.86) >= 513.9 or 514