Stat 22000 Summer 2020 Homework 8 Solutions

Problems to Turn In: due **midnight of Tuesday, July 14, on Canvas**.

1. *[10 points in total]* This problem is essentially the "On Your Own" part in Lab #6.

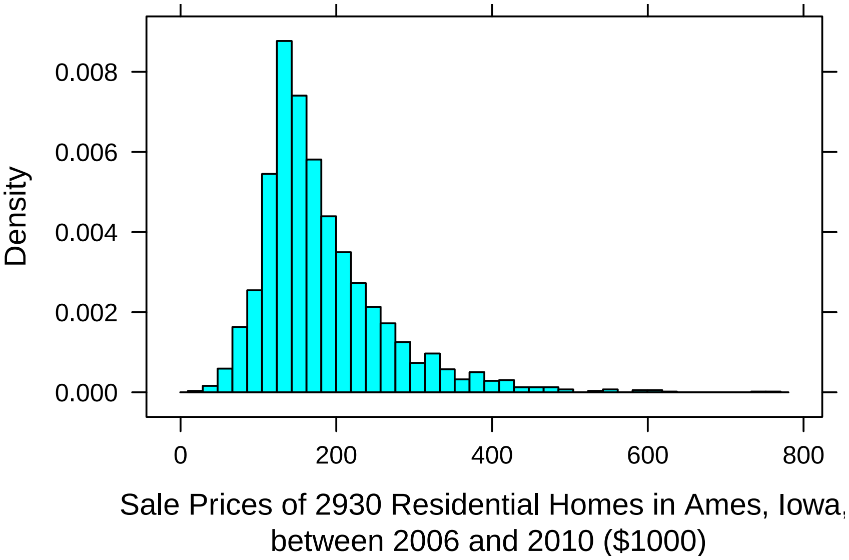
   `http://www.stat.uchicago.edu/~yibi/s220/labs/lab06.html`

   Please complete the lab and submit answers to the following questions.

   (a) Make a histogram of the `price`. Comment on the shape of the histogram.

   Answer: [*In all the parts below, as I divided* `price` *by 1000, the unit of* `price` *becomes $1000, not $1. This step is not necessary. Students may stick with the original scale and hence the plots and numbers will differ by a factor of 1000.*]

   [2pts = 1pt for the histogram + 1pt for the shape] The histogram of `price` is right-skewed. Note we divide `price` by 1000. After division, the unit of `price` is $1000.



   The R codes below are included for instructional purpose, not required.

```
library(mosaic)
download.file("http://www.openintro.org/stat/data/ames.RData", destfile = "ames.RData")
load("ames.RData")
price = ames$SalePrice
price = price/1000
histogram(price, nint=40, xlab="Sale Prices of 2930 Residential Homes in Ames, Iowa,\n
       between 2006 and 2010 ($1000)")
```

   (b) Let's treat the 2930 homes in the data set as the population. Find the mean μ and the SD σ of the sale price of the population.

   Answer: *[not graded]* The population mean μ is $180796.1 and the population SD σ is $79886.69.

```
> favstats(price)
    min    Q1 median    Q3 max     mean       sd    n missing
 12.789 129.5    160 213.5 755 180.7961 79.88669 2930       0
```

(c) Take a random sample of size 25 from `price`. Find the sample mean, and compare it with the population mean μ you found in the previous part.

Answer: [*not graded*] The answer may vary. For my sample, the mean is $175,726, which is pretty close to the population mean $180,796.1.

```
> samp1 = sample(price, 25)
> mean(samp1)
[1] 175.726
```

(Not required) By the 99.7% rule, 99.7% of the sample means should lie within 3 SE ($= \sigma/\sqrt{n}$) from the population mean μ, i.e., between

$$\mu \pm 3\frac{\sigma}{\sqrt{n}} = \$180796.1 \pm 3 \times \frac{\$79886.69}{\sqrt{25}} = \$180796.1 \pm \$47932.02 = (\$132{,}864.0, \$228{,}728.1)$$

So most students should get a number in the range above.

(d) Since we have access to the population, simulate the sampling distribution for the sample mean by taking 5000 samples each of size 25 from the population and computing 5000 sample means. Store these means in a vector called `sample_means25`. Make a histogram of the 5000 sample means.

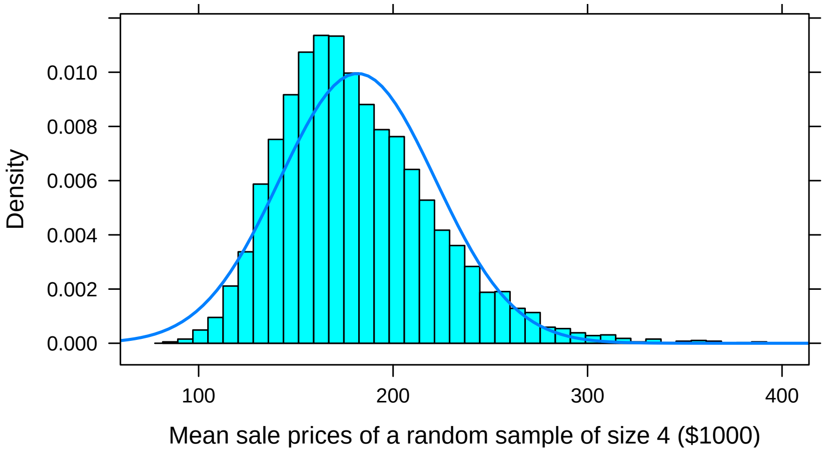
(e) Repeat the previous part but change the sample size from 25 to 4. Store the 5000 sample means in a vector called `sample_means4`.

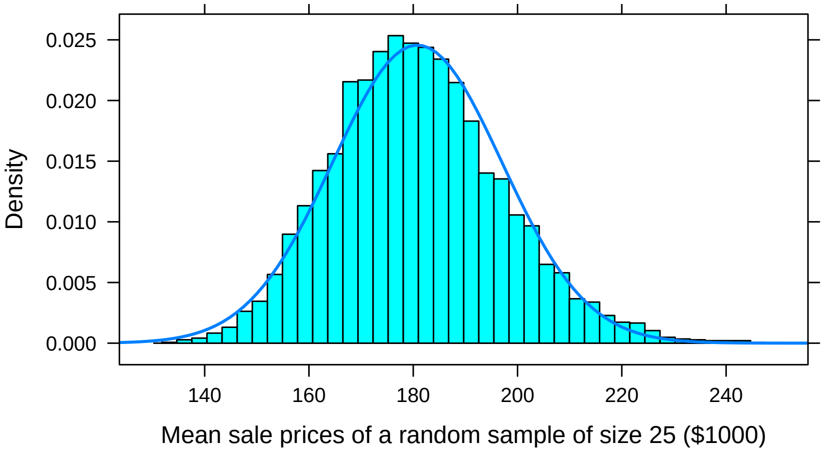
(f) Repeat the previous part again but change the sample size to 100. Store the 5000 sample means in a vector called `sample_means100`

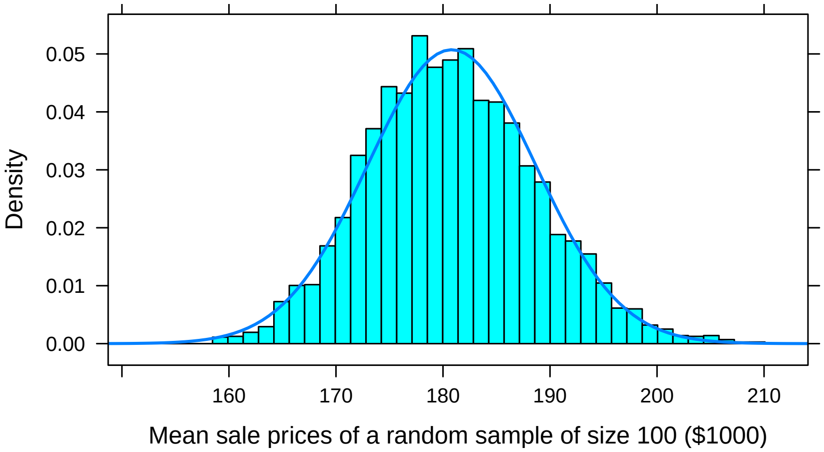
(g) The 3 histograms made above show (roughly) the sampling distributions of the sample mean for a sample of size 4, 25, and 100, respectively. Compare the center, spread, and the shape of the 3 sampling distributions. How do the center, spread, and shape of the sampling distributions change with the sample size?

Answer: [*4pts = 1pt for the histograms + 1pt for the center + 1pt for the spread + 1pt for the shape of the histogram*]

The histograms of the 5000 sample means of size 4, 25, and 100 are as follows.







R codes (not required):

```
sample_means4 = do(5000) * mean(sample(price, 4))
histogram(sample_means4$mean, nint=40, fit='normal')

sample_means25 = do(5000) * mean(sample(price, 25))
histogram(sample_means25$mean, nint=40, fit='normal')

sample_means100 = do(5000) * mean(sample(price, 100))
histogram(sample_means100$mean, nint=40, fit='normal')
```

[*1pt*] All three histograms center at the population mean $180, 796.1

[*1pt*] The sample means range from 100K to 300K for a sample of size 4, 140K to 220K for a sample of size 25, and 160K to 200K for a sample of size 100. We can see the spread of the histogram (variability of the sample mean) decrease as the the sample size increase.

[*1pt*] All three histograms are slightly right-skewed. The degree of skewness decreases as the sample size increases. The histogram is very very close to normal when the sample size reaches 100.

(h) Use the CLT to find the (approximate) probability of getting a sample mean below $170,000 for a sample of size 100.

Answer: [*2pts*] By the CLT, $\overline{X} \sim N(\mu = 180796.1, SE = 79886.69/\sqrt{100})$.

$$P(\overline{X} < 170000) = P\left(Z < \frac{170000 - 180796.1}{79886.69/\sqrt{100}}\right) \approx P(Z < -1.3514) = pnorm(-1.3514) \approx 0.0883 \approx 8.83\%.$$

Or one can find the answer in R:

```
> pnorm(170000, m=180796.1, s=79886.69/sqrt(100))
[1] 0.0882794
```

(i) What percentage of the 5000 sample means in `sample_means100$mean` are below \$170,000? Is the percentage close to the probability computed in the previous part using CLT?

Answer: [*0pt*] The answer may vary. In my simulation, 377 of the 5000 sample means are below 170 thousands. The percentage is $377/5000 = 0.0754 = 7.54\%$, which is somewhat close to the probability 8.85% calculated by CLT. The CLT works decently well here.

```
> table(sample_means100 < 170)

FALSE  TRUE
 4623   377
```

(j) Use the CLT to find the (approximate) probability of getting a sample mean between \$130,000, and \$190,000 for a sample of size 4.

Answer: [*2pts*] By the CLT, $\overline{X} \sim N(\mu = 180796.1, SE = 79886.69/\sqrt{4})$.

$$
\begin{aligned}
P(130000 < \overline{X} < 190000) &= P\left(\frac{130000 - 180796.1}{79886.69/\sqrt{4}} < Z < \frac{190000 - 180796.1}{79886.69/\sqrt{4}}\right) \\
&\approx P(-1.27 < Z < 0.23) = P(Z < 0.23) - P(Z < -1.27) = pnorm(0.23) - pnorm(\\
&\approx 0.5910 - 0.1020 = 0.489 \approx 48.9\%.
\end{aligned}
$$

Or one can find the answer in R:

```
> pnorm(190000, m=180796.1, s=79886.69/sqrt(4))-pnorm(130000, m=180796.1, s=79886.69/sqrt(4
[1] 0.4893796
```

(k) What percentage of the 5000 sample means in `sample_means4$mean` are between \$130,000 and \$190,000? Hint: Use the following R commands.

```
table(sample_means4 < 190000 & sample_means4 > 130000)
table(sample_means4 < 190000 & sample_means4 > 130000)/5000
```

Is the percentage close to the probability computed in the previous part using CLT? Does the CLT work well when the sample size is only 4?

Answer: The answer may vary. In my simulation, 2839 of the 5000 sample means are between 130 and 190 thousands. The percentage is $2839/5000 = 0.5678 = 56.78\%$, which is quite a bit higher than probability 48.9% calculated by CLT. The CLT doesn't work decently well here.

```
> table(sample_means4 < 190 & sample_means4 > 130)

FALSE  TRUE
 2161  2839
```

---

2. In the Southern Ocean food web, the krill species Euphausia superba is the most important prey species for many marine predators, from seabirds to the largest whales. Body lengths of the species are normally distributed with a mean of 40 mm and a standard deviation of 12 mm[1].

   (a) What is the probability that a randomly selected krill is longer than 46 mm?

   (b) Describe the distribution of the mean length of a sample of four krill.

   (c) What is the probability that the mean length of a sample of four krill is more than 46 mm?

   (d) Could you estimate the probabilities from parts (a) and (c) if the lengths of krill had a skewed distribution?

Answer: *[8 points in total]*

(a) *[2pts]* Let X denote the body length of a randomly chosen krill of the species. Then, $X \sim N(\mu = 40, \sigma = 12)$.

$$P(X > 46) = P\left(Z > \frac{46-40}{12}\right) = P(Z > 0.5) = 1 - pnorm(0.5) \approx 1 - 0.6915 = 0.3085.$$

Or in R:

```
> pnorm(46, m=40, s=12,lower.tail=F)
[1] 0.3085375
```

(b) *[2pts]* Since the population distribution is normal, the distribution of the sample mean will also be normal regardless of the sample size. So $\bar{X} \sim N(\mu = 40, SE = 12/\sqrt{4} = 6)$.

(c) *[2pts]* Let $\bar{X}$ denote the mean weight of the four krill. Then,

$$P(\bar{X} > 46) = P\left(Z > \frac{46-40}{12/\sqrt{4}}\right) = P(Z > 1) = 1 - pnorm(1) \approx 0.1587.$$

Or in R:

```
> pnorm(46, m=40, s=12/sqrt(4), lower.tail=F)
[1] 0.1586553
```

(d) *[2pts]* We could not estimate (a) without a nearly normal population distribution. We might not be able to estimate (c) since the sample size is not sufficient to yield a nearly normal sampling distribution if the population distribution is not nearly normal.

---

3. A study of rush-hour traffic in San Francisco counts the number of people in each car entering a freeway at a suburban interchange. Suppose that this count has mean 1.55 and standard deviation 0.85 in the population of all cars that enter at this interchange during 7-9 am.

   (a) Could the exact distribution of the count be normal? Why or why not?

   (b) Find the (approximate) probability that 800 randomly selected cars at this freeway interchange between 7-9 am will carry a total of 1200 people or more. Show your work. (Hint: Restate this event in terms of the mean number of people $\overline{x}$ per car.)

   (c) Is it possible to find the (approximate) probability that 4 randomly selected cars at this freeway interchange between 7-9 am will carry more than 5 people? Why or why not?

[1] Source : K. Reid et al., "Krill Population Dynamics at South Georgia 1991-1997 Based on Data From Predators and Nets",*Marine Ecology Progress Series*, Vol. 177, pp. 103-14

Answer: *[7 points in total]*

(a) *[2pts]* Not normal since the count is integer-valued, 1, 2, 3, 4,. . . . Normal distributions are continuous distributions. There is no car carrying 1.1 to 1.9 passengers. For a normal distribution, we can calculate $P(1.1 < X < 1.9)$ and it's never 0.

(b) *[4pts = 1pt for the reasons that normal curve can be used + 1pt for normal distribution + 2pts for the normal probability]* We are looking for the probability that the total of 800 cars carry more than 1200 people. This means that one car on average should cover 1200/800 = 1.5 persons. Here we can use the CLT because of the large sample size 800 even though the population distribution is not normal.

Let $\bar{X}$ be the mean number of people in a sample of 800 cars. By the CLT, we know $\bar{X} \sim N(\mu = 1.55, \text{SE} = 0.85/\sqrt{800})$.

$$P(\bar{X} \geq 1.5) = P\left(Z \geq \frac{1.5 - 1.55}{0.85/\sqrt{800}}\right) = P(Z \geq -1.66) = 1 - pnorm(-1.66) = 1 - 0.0484 = 0.9515$$

Or in R:

```
> pnorm(1.5, m=1.55, s=0.85/sqrt(800), lower.tail=F)
[1] 0.9519219
```

(c) *[1pt]* Probably not. As the population distribution is obviously not normal, for a small sample size of $n = 4$, the sampling distribution of the sample mean is probably not normal.