Selection bias: occurs if the method for selecting the participants produces a sample that does not represent the population of interest ; **Nonresponse bias:** occurs when a representative sample is chosen for a survey, but a subset cannot be contacted or does not respond; **Response bias:** occurs when participants provide incorrect information

simple random sample without replacement (SRSWOR):all possible subsets are equally likely to be chosen for the sample

Stratified sampling: population is divided into subgroups (strata), and a simple random sample selected from each subgroup

Cluster sampling: population is divided into subgroups called clusters, a random sample of clusters is selected

Systematic sampling: population ordered into a list, and list divided into consecutive segments of the same length. A random starting point is selected from the first segment, and the same point is sampled in each successive segment.

Commonly used transformations: ‣ log / square root (positively skewed data) ‣ exponential / square (negatively skewed)

Simpson's paradox: When effect of confounding variable is strong enough to produce relationships in a different direction from when data are separated into categories according to confounding variable

- The blood cholesterol levels of a population of workers has mean 202 and standard deviation 14
- If a sample of 64 workers is selected, approximate the probability that the sample mean will lie between 198 and 206

$$\mathbb{E}[\bar{X}] = \mu = 202 \text{ and } \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} = \frac{14^2}{64}$$

$$\begin{aligned}\mathbb{P}(198 \le \bar{X} \le 206) &= \mathbb{P}\left(\frac{198-202}{14/\sqrt{64}} \le \frac{\bar{X}-202}{14/\sqrt{64}} \le \frac{206-202}{14/\sqrt{64}}\right)\\ &\approx \mathbb{P}(-2.286 \le Z \le 2.286)\\ &= \Phi(2.286) - \Phi(-2.286) = 0.978\end{aligned}$$

Chi-square distribution

- Definition: for $n \ge 1$, let $Z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ for $i = 1,\cdots,n$. Then the distribution of the random variable

$$X = \sum_{i=1}^{n} Z_i^2$$

is called the Chi-square (χ^2) distribution with n degrees of freedom.

- Key properties: $\mathbb{E}[X] = n$, $\mathrm{Var}(X) = 2n$

Point estimation of sample mean

- Suppose $X_1, \ldots, X_n$ are i.i.d. sampled from a population F, with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$
- What is the MSE of the sample mean for estimating μ?

$$\mathrm{MSE}(\bar{X}) = \mathrm{Var}(\bar{X}) + \left[\mathrm{bias}(\bar{X})\right]^2 = \frac{\sigma^2}{n} + 0 = \frac{\sigma^2}{n}$$

$$\text{Estimated SD} \rightsquigarrow \quad \mathrm{SE}(\bar{X}) = \frac{S}{\sqrt{n}}$$

Quality of a point estimator: bias and variance

- Bias of an estimator $\hat{\theta}$:

$$\mathrm{bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

- Variance and standard deviation of an estimator $\hat{\theta}$:

$$\mathrm{Var}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\right], \quad \mathrm{SD}(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta})}$$

- Ideally, both the bias and the variance are small

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

Example: Bernoulli sample mean calculation w/ normal approximation

Suppose $n = 20$ and $p = 0.5$. Compute the probability that the sample mean is smaller or equal to 0.4 w/ normal approximation

Step I: check the assumptions $np = 10,\ n(1-p) = 20 \ge 10$

Step II: compute the mean and variance

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = np = 10, \quad \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = np(1-p) = 20 \times 0.25 = 5$$

Step III: compute the probability

$$\mathbb{P}(\bar{X} \le 0.4) = \mathbb{P}\left(\sum_{i=1}^{n} X_i \le 8\right) = \mathbb{P}\left(\frac{\sum_{i=1}^{n} X_i - 10}{\sqrt{5}} \le \frac{8-10}{\sqrt{5}}\right) \approx \Phi\left(\frac{8-10}{\sqrt{5}}\right) = 0.1855$$

Normal approximation w/ continuity correction

- Previously

$$\mathbb{P}(\bar{X} \le 0.4) = \mathbb{P}\left(\sum_{i=1}^{n} X_i \le 8\right) = \mathbb{P}\left(\frac{\sum_{i=1}^{n} X_i - 10}{\sqrt{5}} \le \frac{8-10}{\sqrt{5}}\right) \approx \Phi\left(\frac{8-10}{\sqrt{5}}\right) = 0.1855$$

- Continuity correction

$$\mathbb{P}(\bar{X} \le 0.4) = \mathbb{P}\left(\sum_{i=1}^{n} X_i \le 8\right) \approx \mathbb{P}(\mathcal{N}(10,5) \le 8.5) = \Phi\left(\frac{8.5-10}{\sqrt{5}}\right) = 0.2512$$

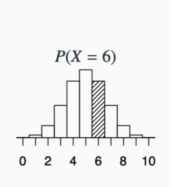
Compute the value of the CDF of Bin(20,0.5) at the value 8:

Normal approximation w/ continuity correction

Suppose $X \sim \mathrm{Bin}(10{,}0.5)$

- Approximate $\mathbb{P}(X = 6)$

$$\begin{aligned}\mathbb{P}(X=6) &\approx \mathbb{P}\left(5.5 \le \mathcal{N}(5{,}2.5) \le 6.5\right)\\ &= \mathbb{P}\left(\frac{5.5-5}{\sqrt{2.5}} \le Z \le \frac{6.5-5}{\sqrt{2.5}}\right)\\ &= \Phi\left(\frac{6.5-5}{\sqrt{2.5}}\right) - \Phi\left(\frac{5.5-5}{\sqrt{2.5}}\right)\\ &= 0.2045\end{aligned}$$



Compute $\mathbb{P}(\mathrm{Bin}(10, 0.5) = 6)$:

```
dbinom(6,size = 10, prob = 0.5)
```

```
## [1] 0.2050781
```

Sample variance with unknown mean

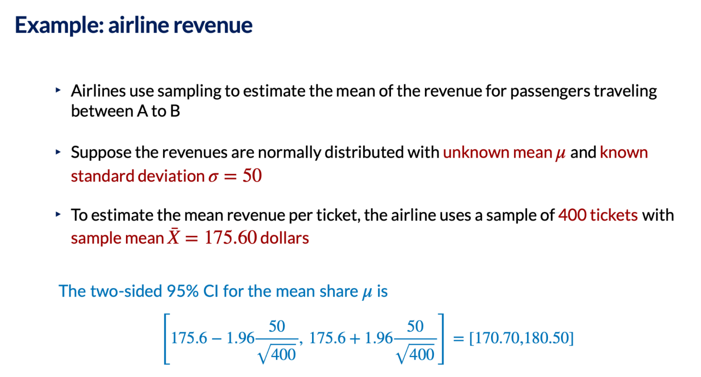
$$\boxed{X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)}$$

- μ is unknown; estimate σ^2
- An alternative estimator $S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

$$S_1^2 = \frac{n-1}{n}S^2 \Rightarrow \mathbb{E}[S_1^2] = \frac{n-1}{n}\mathbb{E}[S^2] = \frac{n-1}{n}\sigma^2, \quad \mathrm{Var}(S_1^2) = \frac{(n-1)^2}{n^2}\mathrm{Var}(S^2) = \frac{2(n-1)\sigma^4}{n^2}$$

$$\boxed{\begin{aligned}&\mathrm{MSE}(S_1^2) = \mathrm{Var}(S_1^2) + \left[\mathrm{bias}(S_1^2)\right]^2 = \frac{(2n-1)\sigma^4}{n^2}\\ &\mathrm{MSE}(S^2) = \mathrm{Var}(S^2) + \left[\mathrm{bias}(S^2)\right]^2 = \mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}\end{aligned}}$$

Can have smaller MSE with biased estimators!

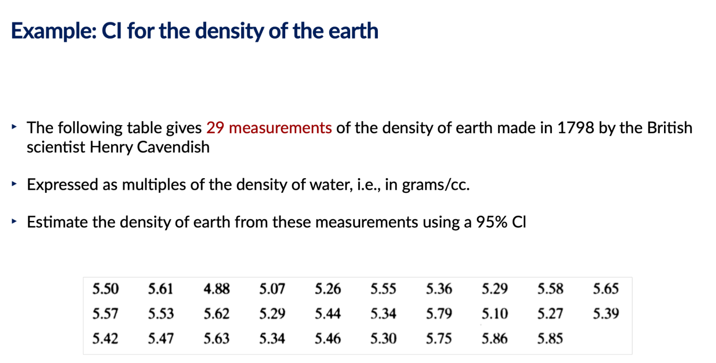


Example: airline revenue

- Airlines use sampling to estimate the mean of the revenue for passengers traveling between A to B
- Suppose the revenues are normally distributed with unknown mean μ and known standard deviation $\sigma = 50$
- To estimate the mean revenue per ticket, the airline uses a sample of 400 tickets with sample mean $\bar{X} = 175.60$ dollars

The two-sided 95% CI for the mean share μ is

$$\left[175.6 - 1.96\frac{50}{\sqrt{400}},\ 175.6 + 1.96\frac{50}{\sqrt{400}}\right] = [170.70, 180.50]$$



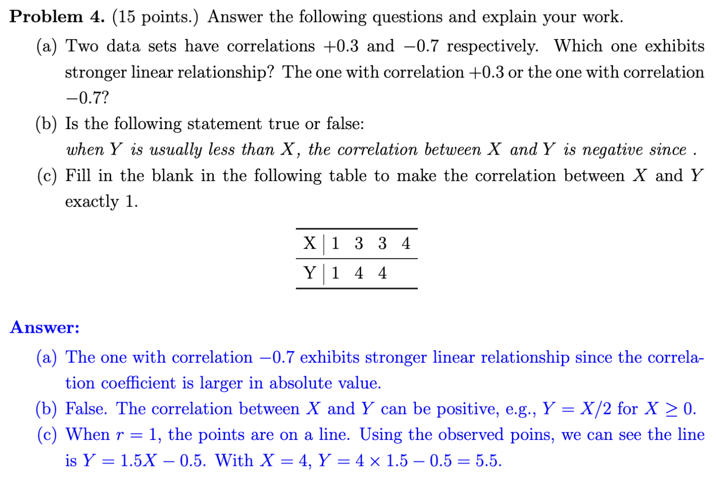
Example: CI for the density of the earth

- The following table gives 29 measurements of the density of earth made in 1798 by the British scientist Henry Cavendish
- Expressed as multiples of the density of water, i.e., in grams/cc.
- Estimate the density of earth from these measurements using a 95% CI

5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65
5.57	5.53	5.62	5.29	5.44	5.34	5.79	5.10	5.27	5.39
5.42	5.47	5.63	5.34	5.46	5.30	5.75	5.86	5.85	

mean time needed for the drug to enter the blood stream is less than 10m‣ NullH: mean time is at least 10 minutes

There are two hospitals in town, hospital A and hospital B. Hospital A has a higher success rate of both high-risk and low-risk surgical procedures than hospital B. Can we conclude that overall, hospital A has a higher success rate of all kinds of surgical procedures than hospital B? - SImpsons Paradox



Problem 4. (15 points.) Answer the following questions and explain your work.

(a) Two data sets have correlations +0.3 and −0.7 respectively. Which one exhibits stronger linear relationship? The one with correlation +0.3 or the one with correlation −0.7?

(b) Is the following statement true or false:
when Y is usually less than X, the correlation between X and Y is negative since .

(c) Fill in the blank in the following table to make the correlation between X and Y exactly 1.

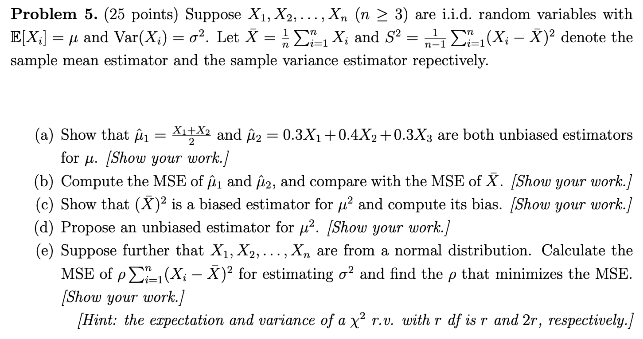
X	1	3	3	4
Y	1	4	4	

Answer:

(a) The one with correlation −0.7 exhibits stronger linear relationship since the correlation coefficient is larger in absolute value.

(b) False. The correlation between X and Y can be positive, e.g., $Y = X/2$ for $X \geq 0$.

(c) When $r = 1$, the points are on a line. Using the observed poins, we can see the line is $Y = 1.5X - 0.5$. With $X = 4$, $Y = 4 \times 1.5 - 0.5 = 5.5$.



Problem 5. (25 points) Suppose $X_1, X_2, \ldots, X_n$ ($n \geq 3$) are i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ denote the sample mean estimator and the sample variance estimator repectively.

(a) Show that $\hat{\mu}_1 = \frac{X_1+X_2}{2}$ and $\hat{\mu}_2 = 0.3X_1 + 0.4X_2 + 0.3X_3$ are both unbiased estimators for μ. *[Show your work.]*

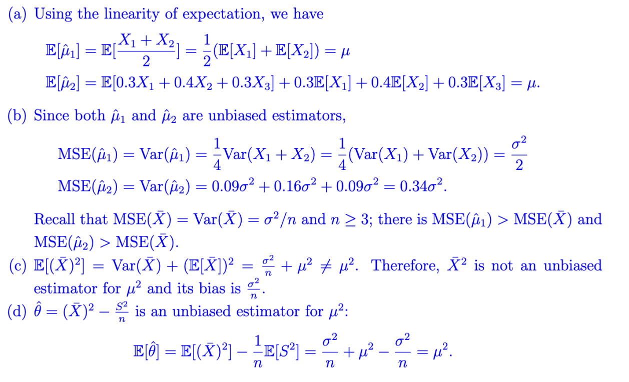
(b) Compute the MSE of $\hat{\mu}_1$ and $\hat{\mu}_2$, and compare with the MSE of $\bar{X}$. *[Show your work.]*

(c) Show that $(\bar{X})^2$ is a biased estimator for μ^2 and compute its bias. *[Show your work.]*

(d) Propose an unbiased estimator for μ^2. *[Show your work.]*

(e) Suppose further that $X_1, X_2, \ldots, X_n$ are from a normal distribution. Calculate the MSE of $\rho\sum_{i=1}^n (X_i - \bar{X})^2$ for estimating σ^2 and find the ρ that minimizes the MSE. *[Show your work.]*

[Hint: the expectation and variance of a χ^2 r.v. with r df is r and $2r$, respectively.]



(a) Using the linearity of expectation, we have

$$\mathbb{E}[\hat{\mu}_1] = \mathbb{E}[\frac{X_1 + X_2}{2}] = \frac{1}{2}(\mathbb{E}[X_1] + \mathbb{E}[X_2]) = \mu$$
$$\mathbb{E}[\hat{\mu}_2] = \mathbb{E}[0.3X_1 + 0.4X_2 + 0.3X_3] + 0.3\mathbb{E}[X_1] + 0.4\mathbb{E}[X_2] + 0.3\mathbb{E}[X_3] = \mu.$$

(b) Since both $\hat{\mu}_1$ and $\hat{\mu}_2$ are unbiased estimators,

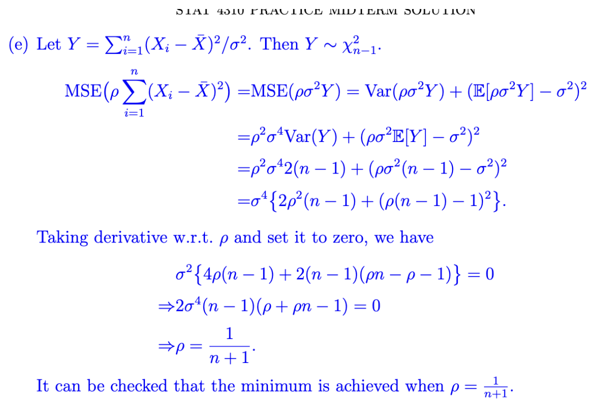
$$\mathrm{MSE}(\hat{\mu}_1) = \mathrm{Var}(\hat{\mu}_1) = \frac{1}{4}\mathrm{Var}(X_1 + X_2) = \frac{1}{4}(\mathrm{Var}(X_1) + \mathrm{Var}(X_2)) = \frac{\sigma^2}{2}$$
$$\mathrm{MSE}(\hat{\mu}_2) = \mathrm{Var}(\hat{\mu}_2) = 0.09\sigma^2 + 0.16\sigma^2 + 0.09\sigma^2 = 0.34\sigma^2.$$

Recall that $\mathrm{MSE}(\bar{X}) = \mathrm{Var}(\bar{X}) = \sigma^2/n$ and $n \geq 3$; there is $\mathrm{MSE}(\hat{\mu}_1) > \mathrm{MSE}(\bar{X})$ and $\mathrm{MSE}(\hat{\mu}_2) > \mathrm{MSE}(\bar{X})$.

(c) $\mathbb{E}[(\bar{X})^2] = \mathrm{Var}(\bar{X}) + (\mathbb{E}[\bar{X}])^2 = \frac{\sigma^2}{n} + \mu^2 \neq \mu^2$. Therefore, $\bar{X}^2$ is not an unbiased estimator for μ^2 and its bias is $\frac{\sigma^2}{n}$.

(d) $\hat{\theta} = (\bar{X})^2 - \frac{S^2}{n}$ is an unbiased estimator for μ^2:

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}[(\bar{X})^2] - \frac{1}{n}\mathbb{E}[S^2] = \frac{\sigma^2}{n} + \mu^2 - \frac{\sigma^2}{n} = \mu^2.$$



STAT 4310 PRACTICE MIDTERM SOLUTION

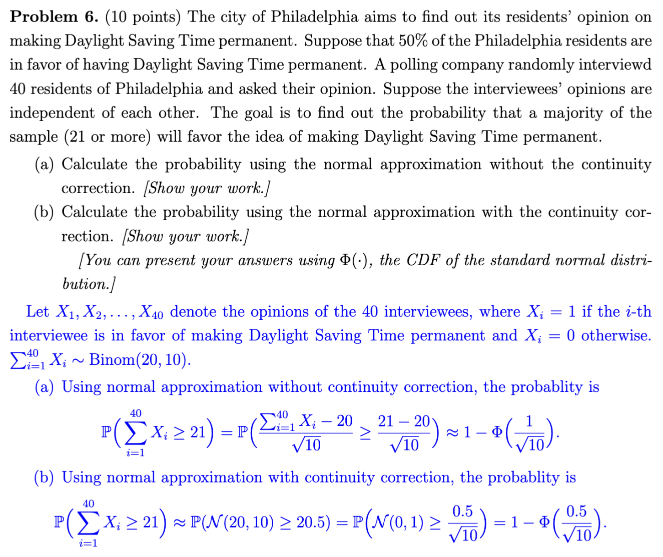
(e) Let $Y = \sum_{i=1}^n (X_i - \bar{X})^2/\sigma^2$. Then $Y \sim \chi^2_{n-1}$.

$$\begin{aligned}\mathrm{MSE}\Big(\rho\sum_{i=1}^n (X_i - \bar{X})^2\Big) =& \mathrm{MSE}(\rho\sigma^2 Y) = \mathrm{Var}(\rho\sigma^2 Y) + (\mathbb{E}[\rho\sigma^2 Y] - \sigma^2)^2 \\ =& \rho^2\sigma^4\mathrm{Var}(Y) + (\rho\sigma^2\mathbb{E}[Y] - \sigma^2)^2 \\ =& \rho^2\sigma^4 2(n-1) + (\rho\sigma^2(n-1) - \sigma^2)^2 \\ =& \sigma^4\big\{2\rho^2(n-1) + (\rho(n-1) - 1)^2\big\}.\end{aligned}$$

Taking derivative w.r.t. ρ and set it to zero, we have

$$\begin{aligned}&\sigma^2\big\{4\rho(n-1) + 2(n-1)(\rho n - \rho - 1)\big\} = 0 \\ \Rightarrow& 2\sigma^4(n-1)(\rho + \rho n - 1) = 0 \\ \Rightarrow& \rho = \frac{1}{n+1}.\end{aligned}$$

It can be checked that the minimum is achieved when $\rho = \frac{1}{n+1}$.



Problem 6. (10 points) The city of Philadelphia aims to find out its residents' opinion on making Daylight Saving Time permanent. Suppose that 50% of the Philadelphia residents are in favor of having Daylight Saving Time permanent. A polling company randomly interviewd 40 residents of Philadelphia and asked their opinion. Suppose the interviewees' opinions are independent of each other. The goal is to find out the probability that a majority of the sample (21 or more) will favor the idea of making Daylight Saving Time permanent.

(a) Calculate the probability using the normal approximation without the continuity correction. *[Show your work.]*

(b) Calculate the probability using the normal approximation with the continuity correction. *[Show your work.]*

[You can present your answers using $\Phi(\cdot)$, the CDF of the standard normal distribution.]

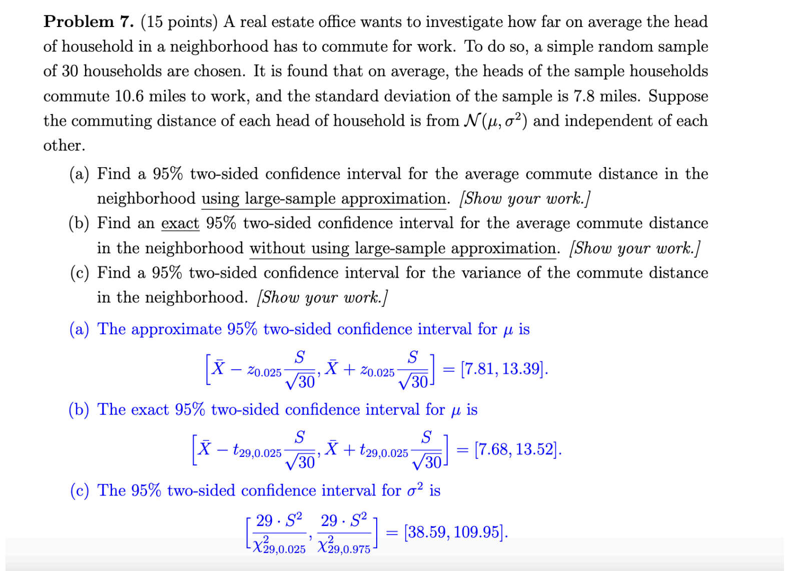
Let $X_1, X_2, \ldots, X_{40}$ denote the opinions of the 40 interviewees, where $X_i = 1$ if the i-th interviewee is in favor of making Daylight Saving Time permanent and $X_i = 0$ otherwise. $\sum_{i=1}^{40} X_i \sim \mathrm{Binom}(20, 10)$.

(a) Using normal approximation without continuity correction, the probablity is

$$\mathbb{P}\Big(\sum_{i=1}^{40} X_i \geq 21\Big) = \mathbb{P}\Big(\frac{\sum_{i=1}^{40} X_i - 20}{\sqrt{10}} \geq \frac{21-20}{\sqrt{10}}\Big) \approx 1 - \Phi\Big(\frac{1}{\sqrt{10}}\Big).$$

(b) Using normal approximation with continuity correction, the probablity is

$$\mathbb{P}\Big(\sum_{i=1}^{40} X_i \geq 21\Big) \approx \mathbb{P}(\mathcal{N}(20, 10) \geq 20.5) = \mathbb{P}\Big(\mathcal{N}(0,1) \geq \frac{0.5}{\sqrt{10}}\Big) = 1 - \Phi\Big(\frac{0.5}{\sqrt{10}}\Big).$$



Problem 7. (15 points) A real estate office wants to investigate how far on average the head of household in a neighborhood has to commute for work. To do so, a simple random sample of 30 households are chosen. It is found that on average, the heads of the sample households commute 10.6 miles to work, and the standard deviation of the sample is 7.8 miles. Suppose the commuting distance of each head of household is from $\mathcal{N}(\mu, \sigma^2)$ and independent of each other.

(a) Find a 95% two-sided confidence interval for the average commute distance in the neighborhood using large-sample approximation. *[Show your work.]*

(b) Find an exact 95% two-sided confidence interval for the average commute distance in the neighborhood without using large-sample approximation. *[Show your work.]*

(c) Find a 95% two-sided confidence interval for the variance of the commute distance in the neighborhood. *[Show your work.]*

(a) The approximate 95% two-sided confidence interval for μ is

$$\Big[\bar{X} - z_{0.025}\frac{S}{\sqrt{30}}, \bar{X} + z_{0.025}\frac{S}{\sqrt{30}}\Big] = [7.81, 13.39].$$

(b) The exact 95% two-sided confidence interval for μ is

$$\Big[\bar{X} - t_{29,0.025}\frac{S}{\sqrt{30}}, \bar{X} + t_{29,0.025}\frac{S}{\sqrt{30}}\Big] = [7.68, 13.52].$$

(c) The 95% two-sided confidence interval for σ^2 is

$$\Big[\frac{29 \cdot S^2}{\chi^2_{29,0.025}}, \frac{29 \cdot S^2}{\chi^2_{29,0.975}}\Big] = [38.59, 109.95].$$