

Solution to Homework 4

1. The spread sheet "birth weight.xlsx" contains the birth weights from 1000 consecutive deliveries at Boston City Hospital. Consider these 1000 birth weights as our population in consideration.
 - (a) Generate the sampling distribution of \bar{X} over 200 samples of size 10 and plot the distribution you generate. (Hint: You can use whatever programming language you feel comfortable with. Use a uniform random generator to obtain 200 random samples with each random sample having 10 birth weights. For each random sample, compute its sample mean. You will have a total of 200 sample means and plot the histogram of these 200 numbers.)
 - (b) For the 200 random samples you obtained in (a), generate the sampling distribution of the sample median.
 - (c) For the 200 random samples you obtained in (a), generate the sampling distribution of the average of the smallest and largest observation.
 - (d) What conclusion can you draw by comparing the three sampling distributions you generate in (a), (b), and (c)?

Solution: Figure 1 shows the histograms. Visual inspection of the three histograms shows that the variance of the sampling distribution of the mean is the smallest. This is because \bar{X} is the minimum variance unbiased estimator of μ . Sample R code to produce the histograms can be found in `Homework_4.R`.

2. Consider the 1000 birth weights as our population again. Generate the sampling distribution of the sample mean using 200 samples of different sizes.
 - (a) $n = 10$
 - (b) $n = 20$
 - (c) $n = 30$
 - (d) What conclusion can you draw by comparing the three sampling distributions you generate in (a), (b), and (c)?

Solution: Figure 2 shows the histograms. Visual inspection of the three histograms shows that the variance of the sampling distribution of the mean decreases as n increases. This is because

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

Sample R code to produce the histograms can be found in `Homework_4.R`.

3. (Illustration of the central limit theorem) Consider the 1000 birth weights as our population again. Generate the sampling distribution of the sample mean over 200 samples of different sizes.
 - (a) $n = 1$
 - (b) $n = 5$
 - (c) $n = 10$

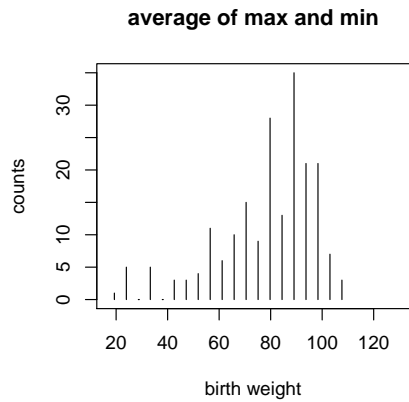
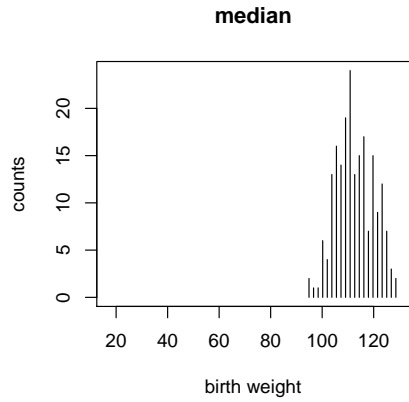
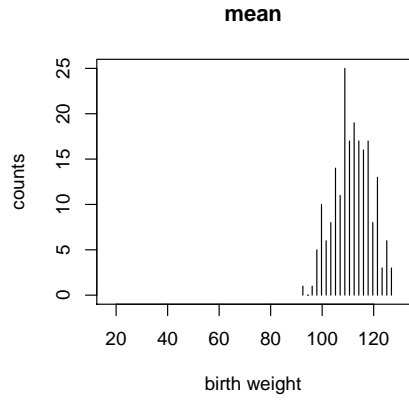


Figure 1: Histograms of Problem 1

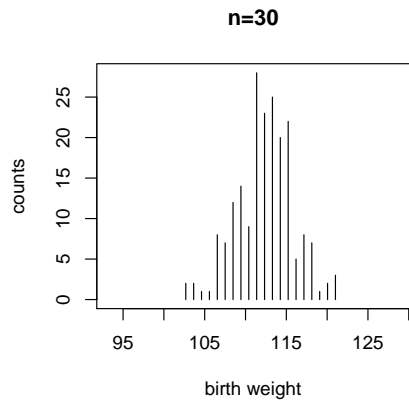
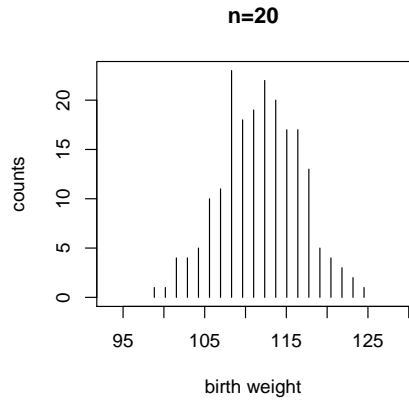
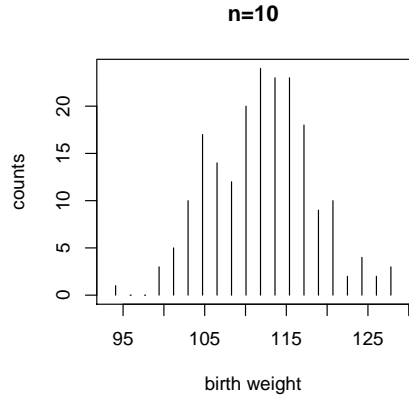


Figure 2: Histograms of Problem 2

(d) Can you see that your sampling distributions of the sample mean can be more closely approximated by a Gaussian distribution when n increases? Explain why.

Solution: Figure 3 shows the histograms. Visual inspection of the three histograms shows that the distribution of \bar{x} approximates a normal distribution when n increases. This is due to the central limit theorem. Sample R code to produce the histograms can be found in `Homework_4.R`.

4. The 1999-2000 National Health and Nutrition Examination Survey was used to estimate dietary intake of 10 key nutrients. One of those nutrients was calcium (mg). They found in all adults 60 years or older a mean daily calcium intake of 721 mg with a standard deviation of 454. Using these values for the mean and standard deviation for the U.S. population, find the probability that a random sample of size 50 will have a mean:

- (a) Greater than 800 mg
- (b) Less than 700 mg
- (c) Between 700 and 850 mg

Solution: First, we need to obtain the sampling distribution of \bar{X} . The following are given:

$$\begin{aligned}\mu &= 721 \\ \sigma &= 454 \\ n &= 50\end{aligned}$$

Thus,

$$\begin{aligned}\mu_{\bar{x}} &= \mu = 721 \\ \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} = \frac{454}{\sqrt{50}} = 64.2053 \\ z &= \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - 721}{64.2053}\end{aligned}$$

(a) The required probability can be obtained from either the distribution of \bar{x} or the distribution of z .

$$\begin{aligned}P[\bar{x} > 800] &= \int_{800}^{\infty} \frac{1}{\sigma_{\bar{x}}\sqrt{2\pi}} e^{-0.5\left(\frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}\right)^2} d\bar{x} \\ &= 1 - \int_{-\infty}^{800} \frac{1}{\sigma_{\bar{x}}\sqrt{2\pi}} e^{-0.5\left(\frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}\right)^2} d\bar{x} \\ &= 1 - .8907 = .1093\end{aligned}$$

$\int_{-\infty}^{800} \frac{1}{\sigma_{\bar{x}}\sqrt{2\pi}} e^{-0.5\left(\frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}\right)^2} d\bar{x}$ can be computed using R command

`pnorm(800,mean=721,sd=64.2053)`.

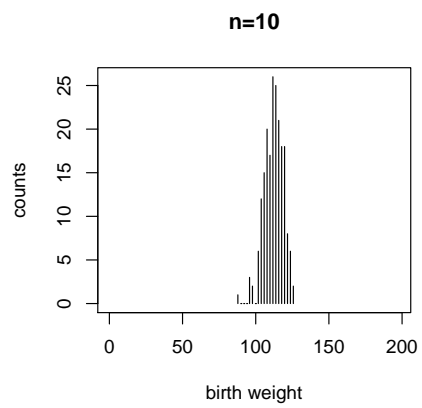
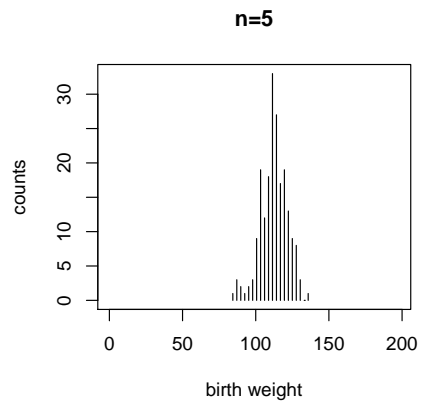
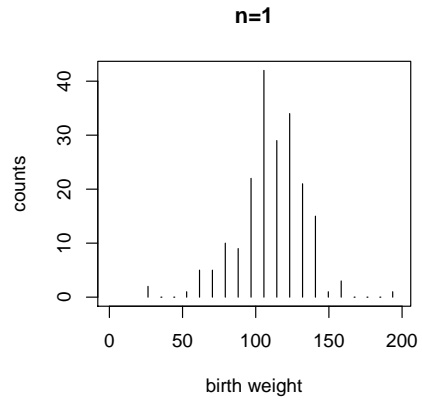


Figure 3: Histograms of Problem 3

Transform $\bar{x} = 800$ to z , we have

$$z = \frac{800 - 721}{64.2053} = 1.23$$

Thus

$$P[\bar{x} > 800] = P[z > 1.23] = 1 - P[z \leq 1.23] = .1093$$

where $P[z \leq 1.23]$ can be obtained using the R command

```
pnorm(1.23,mean=0,sd=1).
```

(b) Transform $\bar{x} = 700$ to z :

$$z = \frac{700 - 721}{64.2053} = -.33$$

Thus

$$P[\bar{x} < 700] = P[z < -.33] = .3707$$

(c) Transform $\bar{x} = 850$ to z :

$$z = \frac{850 - 721}{64.2053} = 2.01$$

Thus,

$$\begin{aligned} P[700 \leq \bar{x} \leq 850] &= P[-.33 \leq z \leq 2.01] \\ &= P[z \leq 2.01] - P[z \leq -.33] = .9778 - .3707 = .6071 \end{aligned}$$

5. Suppose a population consists of the following values: 1,3,5,7,9. Construct the sampling distribution of \bar{x} based on samples of size 2 selected without replacement. Find the mean and variance of the sampling distribution.

solution: A total of $\binom{5}{2}$ samples can be drawn from the population. All the samples are shown in table 4. The mean of each sample is shown in red right below the sample.

Based on the sample means, we have:

$$\begin{aligned} \mu_{\bar{x}} &= \frac{2 + 3 + 4 + 5 + 4 + 5 + 6 + 6 + 7 + 8}{10} = 5 \\ \sigma_{\bar{x}}^2 &= \frac{(2 - 5)^2 + (3 - 5)^2 + 2(4 - 5)^2 + 2(5 - 5)^2 + 2(6 - 5)^2 + (7 - 5)^2 + (8 - 5)^2}{10} = 3 \end{aligned}$$

6. For a population of 17-year-old boys and 17-year-old girls, the means and standard deviations, respectively, of their subscapular skinfold thickness values are as follows: boys, 9.7 and 6.0; girls, 15.6 and 9.5. Simple random samples of 40 boys and 35 girls selected from the populations. What is the probability that the difference between sample means $\bar{x}_{girls} - \bar{x}_{boys}$ will be greater than 10?

Solution: First, we need to obtain the sampling distribution of $\bar{x}_G - \bar{x}_B$. Then, considering that the sample sizes are 40 and 35 for the boys and girls, respectively, we can conclude that the required sampling distribution has the following characteristics:

		second draw				
		1	3	5	7	9
first draw	1		(1,3)	(1,5)	(1,7)	(1,9)
			2	3	4	5
	3			(3,5)	(3,7)	(3,9)
				4	5	6
	5				(5,7)	(5,9)
					6	7
	7					(7,9)
						8
	9					

Figure 4: All the samples of size 2 for problem 5.

- $\mu_{\bar{x}_G - \bar{x}_B} = \mu_{\bar{x}_G} - \mu_{\bar{x}_B} = 15.6 - 9.7 = 5.9$
- $\sigma_{\bar{x}_G - \bar{x}_B} = \sqrt{\frac{\sigma_{\bar{x}_G}^2}{n_G} + \frac{\sigma_{\bar{x}_B}^2}{n_B}} = \sqrt{\frac{9.5^2}{35} + \frac{6.0^2}{40}} = 47.03$
- The distribution of $\bar{x}_G - \bar{x}_B$ is normal.

Transform $\bar{x}_G - \bar{x}_B = 10$ to z , we have

$$z = \frac{10 - 5.9}{47.03} = 2.20$$

Therefore,

$$P[\bar{x}_G - \bar{x}_B > 10] = P[z > 2.20] = 1 - P[z \leq 2.20] = .0139$$

7. We wish to estimate the average number of heartbeats per minute for a certain population. The average number of heartbeats per minute for a sample of 49 subjects was found to be 90. Assume that these 49 patients constitute a random sample, and that the population is normally distributed with a standard deviation of 10. Construct 90, 95, and 99 percent confidence intervals for the population mean, and state the practical and probabilistic interpretations of each. Explain why the three intervals that you construct are not of equal width. Indicate which of the three intervals you would prefer to use as an estimate of the population mean, and state the reason for your choice.

Solution: The following are given:

$$\bar{x} = 90$$

$$n = 49$$

$$\sigma = 10$$

Therefore, $\sigma_{\bar{x}}$ can be computed as

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{49}} = 1.43$$

The confidence intervals are:

$$90\% : 90 \pm 1.645(1.43) = [88, 92]$$

$$95\% : 90 \pm 1.96(1.43) = [87, 93]$$

$$99\% : 90 \pm 2.58(1.43) = [86, 94]$$

Probabilistic interpretation of the 90% confidence interval: Among all the confidence intervals, 90% of them will contain the true population mean.

Practical interpretation of the 90% confidence interval: We are 90% confident that the interval [88,92] contains the true population mean.

The reason why the three intervals are of different length is because the standard deviation $\sigma_{\bar{x}}$ is multiplied by different reliability coefficients for different confidence intervals. I would prefer to use the 90% confidence interval because it is the shortest.

8. The maximal oxide diffusion rate in a sample of 15 asthmatic schoolchildren and 15 controls was reported as mean \pm standard error of the mean. For asthmatic children, they reported 3.5 ± 0.4 nL/s (nanoliters per second) and for control subjects they reported 0.7 ± 0.1 nL/s. For each group, determine the following:
- (a) What was the sample standard deviation?
 - (b) What is the 95% confidence interval for the mean maximal nitric oxide diffusion rate of the population?
 - (c) What assumptions are necessary for the validity of the confidence interval you constructed?
 - (d) What are the practical and probabilistic interpretations of the interval you constructed?
 - (e) If you were to construct a 90% confidence interval for the population mean from the information given here, would the interval be wider or narrower than the 95% confidence interval? Explain your answer without actually constructing the interval.
 - (f) If you were to construct a 99% confidence interval for the population mean from the information given here, would the interval be wider or narrower than the 95% confidence interval? Explain your answer without actually constructing the interval.

Solution: (a) For the asthmatic group: $(.4)\sqrt{15} = 1.549$

For the control group: $(.1)\sqrt{15} = .387$

(b) Assume that maximal oxide diffusion rate of the population of both asthmatic and non-asthmatic school children follow a normal distribution. Since sample size, 15, is not large (smaller than 30) and the population variance is unknown, we need to use s to estimate σ . Therefore, the distribution of

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

follows the t distribution with $15 - 1 = 14$ degrees of freedom. The corresponding reliability coefficient for a 95% confidence interval can be obtained using the R command

$$\text{qt}(.025, \text{df}=14)$$

as 2.1448. Therefore, the 95% confidence interval for the asthmatic group is

$$3.5 \pm 2.1448(.4) = [2.64, 4.36]$$

The 95% confidence interval for the control group is

$$.7 \pm 2.1448(.1) = [.49, .91]$$

(c) Nitric oxide diffusion rates are normally distributed in the population from which the sample was drawn.

(d) Practical: We are 95% confident the mean maximal nitric oxide diffusion rate for asthmatic schoolchildren is between 2.64 and 4.36 while for control subjects it is between .49 and .91.

Probabilistic: Approximately 95% of intervals constructed in a similar manner with samples of size 15 drawn from the populations of asthmatic and control schoolchildren contain the respective population means.

(e) Narrower because the t coefficient from which the interval is constructed is smaller.

(f) Wider because the t coefficient from which the interval is constructed is larger.

9. A study was performed to examine free fatty acid concentrations in 18 lean subjects and 11 obese subjects. The lean subjects had a mean level of 299 $\mu\text{Eq/L}$ with a standard error of the mean of 30, while the obese subjects had a mean of 744 $\mu\text{Eq/L}$ with a standard error of the mean of 62. Construct 90, 95, and 99 percent confidence intervals for the difference between population means. State the assumptions that make your method valid. State the practical and probabilistic interpretations of each interval that you construct. Consider the variables under consideration and state what use you think researchers might make of your results.

Solution: The samples constitute independent simple random samples from the two populations. The two populations of free fatty acid concentrations are normally distributed and the two population variances are equal.

$$s_1 = 30\sqrt{18} = 127.2792$$

$$s_2 = 62\sqrt{11} = 205.6307$$

$$s_p^2 = \frac{17(127.2792^2) + 10(205.6307^2)}{18 + 11 - 2} = 25860.7318$$

$$s_p = \sqrt{25860.7318} = 160.8127$$

$$90\% \text{ CI: } (299 - 744) \pm 1.7033\sqrt{\frac{25860.7318}{18} + \frac{25860.7318}{11}} = [-549.8, -340.2]$$

$$95\% \text{ CI: } (299 - 744) \pm 2.0518\sqrt{\frac{25860.7318}{18} + \frac{25860.7318}{11}} = [-571.3, -318.7]$$

$$99\% \text{ CI: } (299 - 744) \pm 2.7707\sqrt{\frac{25860.7318}{18} + \frac{25860.7318}{11}} = [-615.5, -274.5]$$