

Assignment #5 contains problems about sampling distributions and a first problem about confidence intervals. Problems which require the use of R have the symbol **(R)**.

Please include **your name** and **the discussion section (day/time) that you attend** on your homework. This assignment is worth 50 points in total. If you feel challenged by these problems, I encourage you to do additional problems on your own. Many problems have answers in the back of the textbook.

Your assignment must be turned in during lecture or to your TA's mailbox by 5pm on the due date. We will not grade late homework. If there are special circumstances, please speak to Professor Larget, preferably in advance, for consideration.

1. [5 points] Do Exercise 5.4.
2. [5 points] Do Exercise 5.22.
3. [5 points] Do Exercise 5.23.
4. [5 points] Do Exercise 5.34.
5. **(R)** [10 points] Load in the `prob.R` source file as on a previous assignment. (I made slight changes to this file, so you should reload it even if you already have done so in the past.)

```
> source("http://www.stat.wisc.edu/courses/st371-larget/prob.R")
```

Otherwise, you can download the file onto your computer by right-clicking (for a Windows computer) on the link to `prob.R` from the schedule on the course homepage. You source this data into R from an option from the File menu (Source File... on a Mac, Source R code... on a Windows machine).

Use R to plot the sampling distribution of the sample mean for a bimodal distribution with mean  $\mu = 200$ , standard deviation  $\sigma = 72$ , and additional parameter  $d = 65$ , for sample sizes  $n = 1, 2, 4, 16, 25, 36, 49, \text{ and } 100$ . For each sample size, draw a vertical line that is two standard errors to the right of the mean and compare the area under the true sampling distribution with that under the normal approximation.

The blue curve is the true sampling distribution while the red curve is a normal density with the same mean and variance. The blue curve when  $n = 1$  is the population density.

The R code for this for  $n = 1$  is as follows.

```
> n = 1
> gbimod(n,200,72,65)
> abline(v=200 + 1.96*72/sqrt(n))
```

As an alternative to examining the plots one at a time on screen, the next bit of R code will create a PDF file named `prob5-4.pdf` that puts four plots per page. You can look at these with your usual PDF viewer.

```
> n = c(1,2,4,16,25,36,49,100)
> pdf("bimodal.pdf")
> par(mfrow=c(2,2))
> for(k in n) { gbimod(k,200,72,65); abline(v=200+1.96*72/sqrt(k))}
> dev.off()
```

The last command finishes the file by closing the current graphics device.

Answer the following questions based on examining the graphs for each  $n$  visually.

- (a) For which values of  $n$  would you say that the area to the right of  $b$  under the actual sampling distribution is substantially different from the area under the normal curve?
  - (b) When the area is substantially different, is the area under the normal curve an overestimate or an underestimate of the true area?
6. **(R)** [10 points] Repeat the previous problem but use the skewed distribution `gskew` instead of the bimodal distribution `gbimod` with the same mean  $\mu = 200$  and standard deviation  $\sigma = 72$ , but using skewness coefficient 3.99.

The R code for creating a PDF file with all graphs is as follows.

```
> n = c(1,2,4,16,25,36,49,100)
> pdf("skewed.pdf")
> par(mfrow=c(2,2))
> for(k in n) { gskew(k,200,72,3.99); abline(v=200+1.96*72/sqrt(k))}
> dev.off()
```

7. [5 points] The central limit theorem says that the distribution of the sampling distribution of the sample mean is approximately normal when  $n$  is large enough. Which of the following two distributions would require a larger  $n$  for the approximation to be accurate: (1) a strongly skewed distribution, or (2) a symmetric distribution with a non-normal shape. Briefly explain using the previous two problems for guidance.
8. [5 points] Do Exercise 6.12.