STAT430 Homework #2: Due Friday, February 8, 2019.

Name:

---

0. We are continuing with **Section 7.2** of the text in this homework: there is a lot to cover in that section. To get started, let's look at how to approximate the probability of an event via simulation.

In **Example 7.2** (pages 354-355), $Y_1, Y_2, \ldots, Y_9$ are iid $N(\mu, \sigma^2)$ with $\sigma^2 = 1$. The text shows how to compute

$$
\begin{aligned}
P(|\bar{Y} - \mu| \leq 0.3) &= P\left(\frac{-0.3}{\sqrt{1/9}} \leq Z \leq \frac{0.3}{\sqrt{1/9}}\right) \\
&= P\left(-0.9 \leq Z \leq 0.9\right) \\
&= P(Z \leq 0.9) - P(Z \leq -0.9),
\end{aligned}
$$

which evaluates to

```r
pnorm(0.9) - pnorm(-0.9)
```

```
## [1] 0.6318797
```

Let's consider how to approximate this probability via simulation. In `R`, if we make logical statements like

```r
x <- c(3 < 5, 6 < 5) # logical results "combined" into a vector with c()
x
```

```
## [1]  TRUE FALSE
```

then `R` evaluates those statements as `TRUE` or `FALSE`, which are treated numerically as `TRUE = 1` and `FALSE = 0`. So,

```r
mean(x) # mean of logical vector is the proportion TRUE
```

```
## [1] 0.5
```

In this little example, 1/2 of the statements are true. Now we just want to simulate a large number of standard normals, and check the empirical proportion that are between -0.9 and 0.9:

```r
Z <- rnorm(100000)
x <- (abs(Z) < 0.9) # big vector of logical results
mean(x)             # mean of logical vector is the proportion TRUE
```

```
## [1] 0.63344
```

This is quite close to the theoretical calculation. In general, we can approximate the probability of an event by (1) simulating a vector of logical outcomes that are `TRUE` when the event occurs; and (2) computing the mean of the logical vector. We can use this technique to check an analytical calculation, or to approximate the theoretical answer when the analytical calculation is hard or intractable. For example, suppose we wanted some weird probability like

$$P(\ln |Z| > 0.5).$$

We can get a quick, approximate answer via

```r
Z <- rnorm(100000)
x <- (log(abs(Z)) > 0.5) # big vector of logical results
mean(x)                  # mean of logical vector is the proportion TRUE
```

```
## [1] 0.09912
```

1. Complete **Exercise 7.11** of the text, using the `pnorm` function in `R` to compute the exact, theoretical probability. Then check your answer using the probability approximation above, via simulation of 10,000 $N(0,1)$ random variables. As usual, set your random number seed to 4302019.

Answer:

---

2. Complete **Exercise 7.12** of the text, finding the appropriate sample size, n. Then assess your results by simulation as follows. You know that since $Y_1, \ldots, Y_n$ iid $\mathrm{N}(\mu, \sigma^2)$, the exact sampling distribution of the sample mean is $\bar{Y} \sim N(\mu, \sigma^2/n)$. So reset your random number seed to 4302019 and use the `R` function `rnorm` with arguments `mean` = μ (you can choose any real number that doesn't break your computer) and `sd` = $\sqrt{\sigma^2/n} = \sqrt{4^2/n}$ to generate 10000 simulated $\bar{Y}$ values (since we know the exact sampling distribution, we don't have to start by simulating the raw data as in Homework 1: we can actually simulate from the $\bar{Y}$ distribution directly). With your value of n, is $\bar{Y}$ within 1 square inch of your μ at least 90% of the time? Convince yourself that changing the value of μ does not change your answer.

Answer:

---

3. In `R`, it is often useful to write your own functions for computations that you do repeatedly. For example, we can do the general computation

$$
\begin{aligned}
P(|\bar{Y} - \mu| \leq \delta) &= P\left(\frac{-\delta}{\sqrt{\sigma^2/n}} \leq Z \leq \frac{\delta}{\sqrt{\sigma^2/n}}\right) \\
&= P\left(Z \leq \frac{\delta}{\sqrt{\sigma^2/n}}\right) - P\left(Z \leq \frac{-\delta}{\sqrt{\sigma^2/n}}\right)
\end{aligned}
$$

by constructing the `R` function

```
my_prob <- function(delta, sigma, n){ # you can call your function anything you like
  prob <- pnorm(delta / sqrt(sigma ^ 2 / n)) - pnorm(-delta / sqrt(sigma ^ 2 / n))
  return(prob)
}
```

Then we can run our function on a problem like **Example 7.2**:

```
my_prob(0.3, 1, 9)
```

```
## [1] 0.6318797
```

Notice that we did not need to name the arguments, because they came in the order expected by the function. If for some reason we used a different order, we need to use the names:

```
my_prob(sigma = 1, delta = 0.3, n = 9)
```

```
## [1] 0.6318797
```

In `R`, we can often give a vector argument and get a vector response. For example, if we want to look at sample sizes $n = 9, 10, 11, 12$, we can use

```
my_prob(0.3, 1, c(9, 10, 11, 12))
```

```
## [1] 0.6318797 0.6572183 0.6802576 0.7013024
```

Use this new function to complete **Exercise 7.9 and 7.10** of the text. (For 7.9(d), give a better answer than "Yes"!)

Answer:

---

4. In class, we denoted by $z_{\alpha/2}$ the value such that for $Z \sim N(0,1)$,

$$P(Z > z_{\alpha/2}) = P(Z \leq -z_{\alpha/2}) = \frac{\alpha}{2}.$$

In `R`, you can compute $\pm z_{\alpha/2}$ with `qnorm`; for example,

```
round(qnorm(c(0.025, 0.05, 0.95, 0.975)), 3)
```

```
## [1] -1.960 -1.645  1.645  1.960
```

are the values I told you to memorize for $\alpha = 0.05, 0.10$. Also in `R`, the "next largest integer" function is `ceiling`:

```
ceiling(c(3.9, 41.1, 2.7))
```

```
## [1]  4 42  3
```

Use these functions and follow the example above to write your own function that calculates the minimum (integer) value of n needed to guarantee that if $Y_1, \ldots, Y_n$ iid $N(\mu, \sigma^2)$, then $P(|\bar{Y} - \mu| \leq \delta) \geq 1 - \alpha$. Check your function by repeating **Example 7.3** of the book, repeating **Exercise 7.12** above, and completing **Exercise 7.14** using your function (we did this one by hand in class).

Answer:

---

5. Complete **Exercise 7.15** of the text, showing the steps in (a) and (b). For (c), you can use the function you created above.

Answer:

---

6. In class, I mentioned the paper by Student (1908) and the simulation experiment that involved writing the "height and left middle finger measurements of 3000 criminals" on "3000 pieces of cardboard, which were then very thoroughly shuffled and drawn at random." Let's do this experiment without using cardboard! The data are available in `R` in a slightly strange form, so run the following bit of code to extract the 3000 heights into a vector:

```
require(stats)
tmp <- as.numeric(colnames(crimtab)) / 2.54
height_inches <- as.numeric(rep(tmp, colSums(crimtab)))
```

(a) Draw a histogram of the criminals' heights in inches and comment: do they look approximately normal? (b) Set your seed as usual and use the method of simulation from Homework #1 to draw 10,000 simulated samples of size $n = 4$ with replacement from the 3000 criminals. (c) For each simulated sample, compute the sample variance, S^2, using the `var()` function in `R`. (d) Approximate the true variance, σ^2, by `var(height_inches)`. Is the mean of your 10,000 sample variances close to σ^2? (e) Recall that, if Y_1, Y_2, Y_3, Y_4 are iid $N(\mu, \sigma^2)$, then the sample variance satisfies

$$\frac{(4-1)}{\sigma^2} S^2 \sim \chi^2_3.$$

Rescale your 10,000 sample variances by multiplying by $3/\sigma^2$, plot the histogram of your rescaled sample variances (but set `breaks = 40` to get more bins in your histogram than the default number), and add the theoretical pdf of χ^2_3 using the `dchisq` function in `R`. (e). As an alternative to a histogram, use the *empirical cumulative distribution function*, implemented in `R` by `plot(ecdf(your_rescaled_variances))`. Add to your ecdf plot the theoretical cumulative distribution function using the `pchisq` function.

Answer:

---

7. We can avoid using the tables in the back of the book by using `R` functions. For this problem, use the `R` function `qchisq` to do the following computations: (a). Reproduce the *row* of quantiles (or "percentage points") in Table 6, pages 850-851, for 10 degrees of freedom. (b). Reproduce the *column* of quantiles in Table 6 on page 851 that corresponds to an upper tail area of $\alpha = 0.05$. (Small approximation errors might make the book's answer slightly different in some cases.)

Answer:

---

8. Complete **Exercise 7.19** in the text. You can use the `pchisq` function to complete the probability. Then check your answer by setting the usual seed, simulating 10,000 sample variances under the stated conditions, and checking the empirical proportion of sample variances that are bigger than 0.065. Also, we know that in theory, $E(S^2) = \sigma^2$, so check that the sample mean of your 10,000 simulated sample variances is approximately $\sigma^2 = 0.04$, as given in the problem.

Answer:

---

9. Suppose that $Y_1, \dots, Y_n$ are iid $N(\mu, \sigma^2)$ and let S^2 denote the sample variance, as usual. (a). Use the fact that $(n-1)S^2/\sigma^2 \sim \chi^2$ with $n-1$ df to determine the theoretical variance of the sample variance, $V(S^2)$. Show your work. (You can use the fact that a χ^2_ν random variable has $V(\chi^2_\nu) = 2\nu$.) (b). Compute the value of your theoretical variance formula using the numbers given in **Exercise 7.19**. (c). Check your computed theoretical value in (b) against the empirical variance of your simulated variances from the previous problem, (#8 above).

Answer:
