STA 2210 Homework 5 (Due on Tuesday 3/3 by 11:59pm)

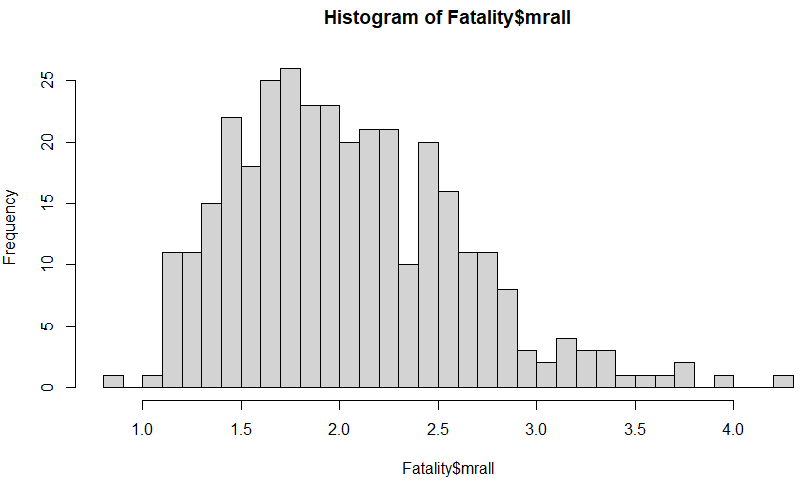
The data set **Fatality**, a .csv file, contains data from the study “Drunk Driving Laws and Traffic Deaths,” for 336 regions in the United States. Write your answer, and R codes, to the following exercises.

1. **Describe the distribution of *mrall*, in terms of its center, shape, spread, and modality.**

summary(Fatality$mrall)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.8212 1.6237 1.9560 2.0404 2.4179 4.2178

hist(Fatality$mrall, breaks = 25)

The distribution of the data is right skewed with a unimodal shape. The center appears to fall around 1.7-2.1, which encompasses where the mean and median are. The graph lacks in symmetry, and the spread of the histogram includes the entire range of the data, from 0.82 – 4.22.

1. **What is the mean and standard deviation of the distribution?**

mean(Fatality$mrall) = 2.040444

sd(Fatality$mrall) = 0.5701938

The mean is 2.04 and the standard deviation is 0.57.

1. **Select a random sample of size 30 from *mrall,* and find the mean and standard deviation of your sample? How do your point estimates compare to the values found in 2?**

TFR <- Fatality$mrall

samp <- sample(TFR, 30)

mean(samp) = 1.932236

sd(samp) = 0.5669702

The point estimate of the mean is 1.93, which is near the population mean of 2.04. The point estimate of the standard deviation is 0.566, which is very close to the population’s standard deviation of 0.570.

1. **Using your point estimates from 3, construct a 95% confidence interval for the true population mean.**

sample\_mean <- mean(samp)

se <- sd(samp) / sqrt(60)

lower <- sample\_mean - 1.96 \* se

upper <- sample\_mean + 1.96 \* se

c(lower, upper) = 1.788773, 2.075700

We are 95% confident that the true population mean is between 1.79 – 2.08 for the traffic fatality rate of the study.

1. **Did your interval capture the true population mean? Explain.**

Yes, the confidence interval captured the true population mean, as the interval is 1.79 – 2.08, and the true population mean, 2.04, falls in this range.

1. **Formally check all 3 conditions on pg. 178 in OpenIntro. These conditions help ensure the confidence interval we created is valid.**

• The sample observations are independent.

Yes, since each sample is independent of another.

• The sample size is large: n ≥ 30 is a good rule of thumb.

Yes, our sample size is 30, so this is checked off.

• The population distribution is not strongly skewed. This condition can be difficult to evaluate, so just use your best judgement.

Yes, I would not say the population distribution is strongly skewed, but it is slightly right skewed. I would say it is not skewed enough to not check this condition off.

1. **Select 50 different samples of size 30 from *mrall*. Using each sample, construct a 95% confidence interval for the true population mean. What percent of your 50 intervals captured the true population mean?**

samp\_mean <- rep(NA, 50)

samp\_sd <- rep(NA, 50)

n <- 30

for(i in 1:50){

samp <- sample(TFR, n) # obtain a sample of size n = 60 from the population

samp\_mean[i] <- mean(samp) # save sample mean in ith element of samp\_mean

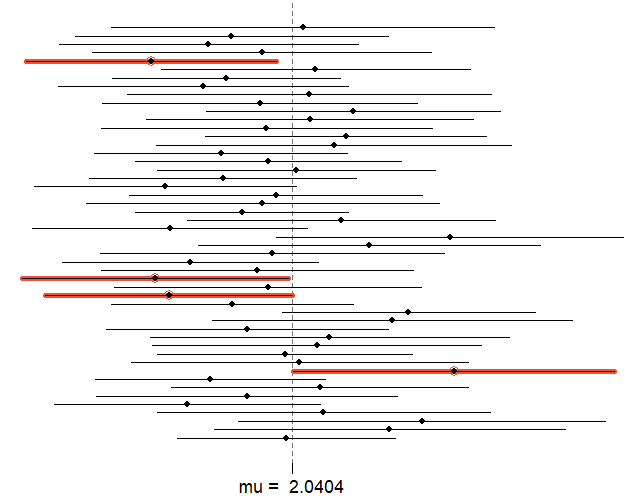
samp\_sd[i] <- sd(samp) # save sample sd in ith element of samp\_sd

}

lower\_vector <- samp\_mean - 1.96 \* samp\_sd / sqrt(n)

upper\_vector <- samp\_mean + 1.96 \* samp\_sd / sqrt(n)

c(lower\_vector[1], upper\_vector[1]) = 1.883349, 2.179544

**plot\_ci**(lower\_vector, upper\_vector, **mean**(TFR))

46/50 samples captured the true population mean.

46/50 = 0.92 = 92%.

This did not match up to the 95% in our confidence interval because we only had 50 random samples of a sample size of 30 traffic fatality rates. If we had more samples and a higher sample size, the confidence interval would capture the true population mean closer to 95% of the time.