Homework #8, Mei Maddox.

Mei Maddox

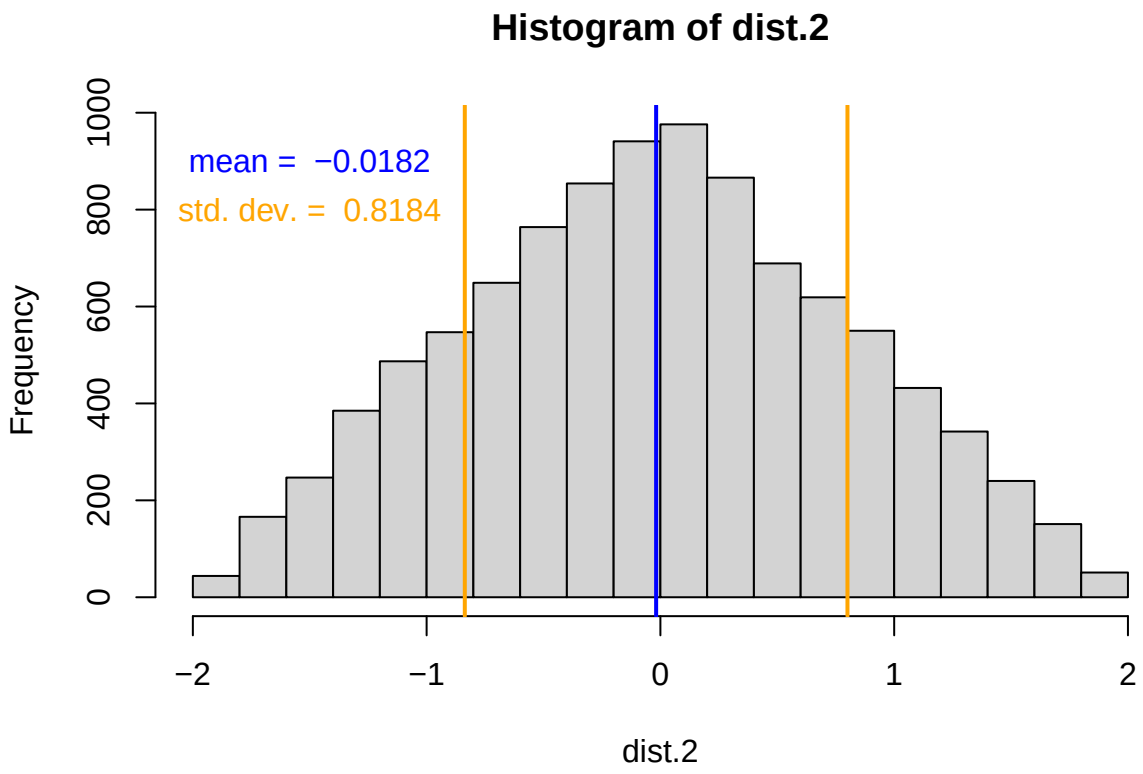
Submit the solution on Canvas into the corresponding assignment (e.g. "Homework #1") in the form of R Markdown report, knitted into either of the available formats (HTML, pdf or Word). Provide only code and output. NO NEED TO COPY THE PROBLEM FORMULATION (!)

Problem 1

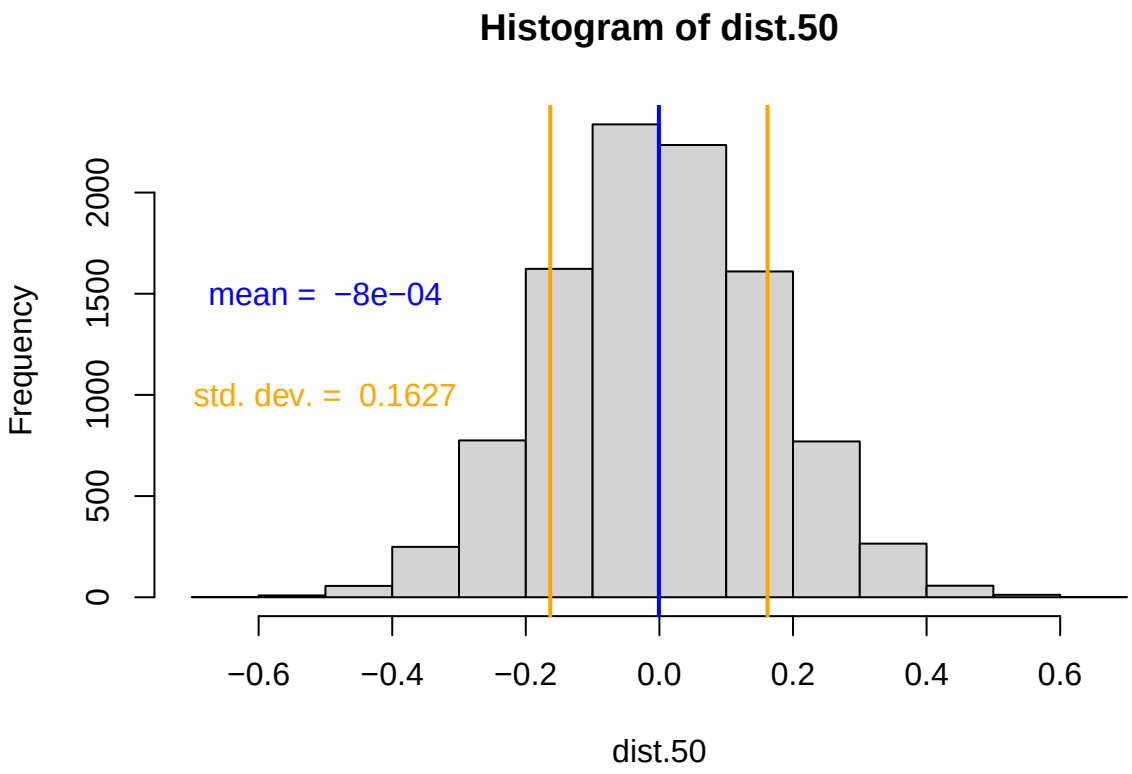
```
set.seed(1)
n.sim <- 10000
dist.2 <- rep(0, n.sim)
dist.50 <- rep(0, n.sim)

for(t in 1:n.sim){
  dist.2[t] <- mean(runif(2, -2, 2))
  dist.50[t] <- mean(runif(50, -2, 2))
}

hist(dist.2)
abline(v = mean(dist.2), col = "blue", lwd = 2)
text(-1.5, 900, paste("mean = ",round(mean(dist.2), 4)), col = "blue")
abline(v = mean(dist.2)-sd(dist.2), col = "orange", lwd = 2)
abline(v = mean(dist.2)+sd(dist.2), col = "orange", lwd = 2)
text(-1.5, 800, paste("std. dev. = ",round(sd(dist.2), 4)), col = "orange")
```




```
hist(dist.50)
abline(v = mean(dist.50), col = "blue", lwd = 2)
text(-.5, 1500, paste("mean = ",round(mean(dist.50), 4)), col = "blue")
abline(v = mean(dist.50)-sd(dist.50), col = "orange", lwd = 2)
abline(v = mean(dist.50)+sd(dist.50), col = "orange", lwd = 2)
text(-.5, 1000, paste("std. dev. = ",round(sd(dist.50), 4)), col = "orange")
```



More observations leads to a more bell-shaped curve. The mean of the sampling distribution is unbiased because the mean of the sampling distribution is effectively equivalent to the mean of the sample. Neither of the standard deviations correspond to the theoretical standard deviation of $\frac{2--2}{\sqrt{12}} = \frac{4}{\sqrt{12}} \approx 1.15$.

Problem 2

```
calculate.ci <- function(n.obs, n.succ, confidence = 0.95){
  p.hat <- n.succ/n.obs
  se <- sqrt(p.hat*(1-p.hat)/n.obs)
  return( c(p.hat - qnorm(0.5 + confidence/2)*se,
            p.hat + qnorm(0.5 + confidence/2)*se) )
}
```

```
set.seed(1)
n.sim <- 10000
prob <- 0.6
size <- 1000
# Placeholders for left (first column) and right (second column) ends
# of our CIs.
ci.95 <- matrix(0, nrow=n.sim, ncol=2)
ci.90 <- matrix(0, nrow=n.sim, ncol=2)
# Distribution of random variable x via generate the 10,000 values from Bin(1000,0.6),
x <- rbinom(n.sim, size, prob)
# Loop through those and feed them as input to your confidence level function from part 1 (for cases of
```

```
for (t in 1:n.sim){
  ci.95[t,] <- calculate.ci(size, x[t])
  ci.90[t,] <- calculate.ci(size, x[t], 0.90)
}
```

I would expect the percentage of times the 95% confidence interval contains the true parameter (0.9438) to be near .95. Likewise, I would expect the percentage of times the 90% confidence interval contains the true parameter (0.8943) to be near .90; both are.

Problem 3

7.7 Baseball player has 500 at-bats (times he is a hitter) and a 0.3 probability of getting a hit in an at-bat. His batting average at the end of the season is the number of hits divided by the number of at-bats.

a) Bell-shaped curve with a mean of $E[X] = p = 0.3$ and standard deviation of $\sqrt{\frac{0.3\times(1-0.3)}{500}} \approx 0.0205$.

b) Both batting averages of 0.320 and 0.280 are only about one standard deviation from the mean, which is not unusual.

7.14 Student running for student government believes 55% of student body will vote for her, but is worried about low voter turnout.

a) Assuming she truly has 55% support and only 200 people show up for voting, the mean is $E[X] = p = 0.55$ and the standard deviation is $\sqrt{\frac{0.55\times(1-0.55)}{200}} \approx 0.0352$.

b) It is reasonable to assume a normal shape for this sampling distribution because there are more than a standard number of samples (15 or 30).

c) The likelyhood that she will not get the majority of the vote is 0.0777377.

d) If n=1000 students, the likelihood that she will not get the majority of the vote is 7.2448723×10^{-4}.

7.15 Bell-shaped with mean $\mu = 70$ and population st. dev. $\sigma = 10$. You randomly sample $n = 12$

a) Due to sampling variety, it is not expected to get exactly 70.

b) The sampling distribution is indeed bell-shaped and centered at 70. Almost all sample means fall within 1 standard deviation.

c) The sampling distribution remains bell-shaped and centered at 70, except the variability has shrunk. In other words, more values fall closer to the mean than before making the bell-shape appear thinner.

7.20 A lottery option in Canada you bet on a 6 digit number between 000000 and 999999. For $1 bet, you win $100000 if you are correct. $\mu = 0.10$ (10 cents), $\sigma = 100$. Over the course of several years he bet 1 million times. Let $\bar{x}$ denote his average winnings.

a) Given a large n, the sampling distribution of sample mean $\bar{x}$ is $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. Therefore the mean is 0.10 and the standard deviation is $\sqrt{\frac{100^2}{1000000}} = 0.10$.

b) The probability that Joe's average winnings exceed $1, the amount to play each time, is 0.

8.6 For 5 paid games of \$1.09, \$4.99, \$1.99, \$1.99, and \$2.99.

a) The point estimate of the mean fee is $\bar{x} = 2.61$.

b) The margin of error at the 95% confidence level for this point estimate is \$1.85. With 95% confidence, the population mean μ of game price falls between \$1.85 above and \$1.85 below the sample mean $\bar{x}$.

8.13 Clinical study 3900 subjects vaccinated. Over course of 28 weeks, 24 subjects developed the flu.

a) $\hat{p} = \frac{24}{3900} \approx 0.00615$

b) Standard error of this estimate $\sqrt{\frac{\hat{p} \times (1-\hat{p})}{n}} \approx 0.00125$

c) The margin of error *moe* for a 95 confidence interval is approx. 0.00245.

d) The confidence interval is bounded by 0.00615 ± 0.00245, inclusive. We are 95% confident that the population proportion of people who receive the flu vaccine but develop the flu is between 0.0037 and 0.00861.

8.16 Number of respondents in favor 1183 out of 1824 total respondents. Therefore $\hat{p} \approx 0.648575$ with a CI of $(0.626665, 0.670484)$.

a) $\hat{p}$ was obtained via dividing those in favor from the total respondents: $\hat{p} = \frac{X}{n} = \frac{1183}{1824}$

b) We are 95% confident that the population proportion of respondents in favor is between 0.626665 and 0.670484.

c) 95% confident refers to interval converge of hypothetically infinite trials. In other words, out of infinite trials, 95% of them resulted in the population parameter p falling within the confidence interval.

d) One can conclude that more than half of all American adults were in favor because 50 is outside the confidence interval.

8.29 "What is the ideal number of children?" 590 female respondents gave numeric response from 0 to 6, median of 2, mean of 2.56, and std. dev. of 0.84

a) The sample mean $\bar{x} = 2.56$.

b) The standard error of the sample mean is $\sqrt{\frac{0.84^2}{590}} \approx 0.03458$.

c) We are 95% confident that the population mean ideal number of children for females is between 2.49 and 2.62.

d) It is not plausible that the population mean $\mu = 2$ because it falls outside of the confidence interval.

8.34 Heights (in mm) for seedlings 14 days after germination is 55.5, 60.3, 60.6, 62.1, 65.5, and 69.2.

```
x <- c(55.5, 60.3, 60.6, 62.1, 65.5, 69.2)
```

a) Use the web app to verify that the confidence interval for the population mean is (57.3, 67.1)

```r
moe <- qt(p=0.05/2, df=5,lower.tail=F)*(sd(x)/sqrt(6))
c(mean(x)-moe, mean(x)+moe)
```

```
## [1] 57.25628 67.14372
```

b) To narrow the interval, one could increase the sample size (number of seedlings) or decrease the confidence level.

c) The 99% confidence interval is wider than the 95% confidence interval because more confidence requires the interval to capture more variety, which therefore widens the interval.

d) It is assumed that the distribution for part a is t-distribution. This is important because it affects the calculation of the confidence interval. It is also assumed that the sample is less than 10% of all seedlings in the world, which allows us to therefore calculate a confidence interval. Furthermore, it is assumed that seedling height is independent of each other which is important because it allows us to construct a distribution.

8.37 14 females in the GSS of age at least 80 reported that they spend 0, 0, 0, 0, 1, 1, 1, 2, 2, 6, 6, 7, 7, 10 hours on email per week.

```r
x <- c(0, 0, 0, 0, 1, 1, 1, 2, 2, 6, 6, 7, 7, 10)
```

a) The sample mean $\bar{x} \approx 3.07$. The standard deviation of the sample mean $s \approx 3.38$. The standard error of the sample mean $se = \frac{s}{\sqrt{14}} \approx 0.905$.

b) The confidence interval is bounded by $\bar{x} \pm t_{1-0.1/2} \times se = 3.07 \pm 1.7709334 \times 0.905$, inclusive. We are 90% confident that the population mean number of hours spent on email per week for females in the GSS of age at least 80 is between 1.4694543 and 4.6734029.

c) The population distribution may be skewed right. Assuming that the sample is representative of the population, then most woman spend two or less hours on email a week. A few spend significantly more than that. This would lead to a right skewed distribution. The interval in part b is still valid because the t-distribution confidence intervals are robust against violations of normality assumptions for small sample sizes. Furthermore, there were no outliers in the sample which is ideal as the t-distribution does not work well when the data contains extreme outliers.