ECON 7710
Homework 1: Statistic Review

General instructions

The `.Rmd` source for this document will be the template for your homework submission. You must submit your completed assignment as a single html document, uploaded to eLC by **11:59pm on January 16** using the filename `sectiontime_lastname_hw1.pdf` (e.g. `935_geddes_hw1.pdf`). After knitting to an html file, please save as a PDF to ensure all images upload to eLC correctly.

Notes:

- Do not alter the formatting code in this template.
- For questions requiring analytical solutions, you can type them in using markdown math code. Or, you can submit handwritten solutions, embedding them in the knitted document as clearly readable images.
- For questions requiring computation, some or all of the required code is included in associated chunks. Modify chunks where and how you are directed.
- For (almost) all questions about R Markdown, consult The Definitive Guide (https://bookdown.org/yihui/rmarkdown/).
- The `setup` chunk above indicates the packages required for this assignment.
- **Switch `eval` to `TRUE` in the global options command to execute code chunks**.

Part A: Bayes' Rule

A rare disease affects approximately 1 in 10,000 people in a specific population. There is a diagnostic test available that is quite accurate, but not perfect. The test has the following characteristics:

- $P(\text{Positive} \mid \text{Have Disease}) = .98$ (This is called sensitivity)
- $P(\text{Negative} \mid \text{Don't Have Disease}) = .995$ (This is called specificity)

a. Write down the definition of Bayes' Rule.
b. If someone tests positive for this disease, what is the probability they actually have the disease?
c. If someone tests negative for this disease, what is the probability they don't have the disease?

Answers

a. $Pr(A|B) = \frac{Pr(B|A) \cdot Pr(A)}{Pr(B)}$

Bayes' Rule is a mathematical formula for determining conditional probability. Conditional probability is the likelihood of an outcome occurring based on a previous outcome in similar circumstances. More specifically, it is a probability theory that describes how to update the probability of a hypothesis based on new evidence.

b.

$$P(\text{Positive}) = P(\text{Positive} \mid \text{Have Disease}) \cdot P(\text{Have Disease}) + P(\text{Positive} \mid \text{Don't Have Disease}) \cdot P(\text{Don't Have Disease})$$

$$P(\text{Have Disease} \mid \text{Positive}) = \frac{P(\text{Positive}|\text{Have Disease}) \cdot P(\text{Have Disease})}{P(\text{Positive})} = 0.0192$$

```
P_Positive_Have = 0.98
P_Negative_Dont = 0.995
P_Have = 0.0001

P_Positive_Dont = 1 - P_Negative_Dont
P_Dont = 1 - P_Have
P_Positive = (P_Positive_Have * P_Have) + (P_Positive_Dont * P_Dont)

P_Have_Positive = (P_Positive_Have *P_Have) / P_Positive

P_Have_Positive
```

```
## [1] 0.01922511
```

c.

$$P(\text{negative}) = P(\text{Negative} \mid \text{Have Disease}) \cdot P(\text{Have Disease}) + P(\text{Negative} \mid \text{Don't Have Disease}) \cdot P(\text{Don't Have Disease})$$

$$P(\text{Don't Have Disease}|\text{Negative}) = \frac{P(\text{Negative}|\text{Don't Have Disease}) \cdot P(\text{Don't Have Disease})}{P(\text{Negative})} = 0.999998$$

```
P_Negative_Have = 1 - P_Positive_Have
P_Negative = (P_Negative_Have * P_Have) + (P_Negative_Dont * P_Dont)

P_Dont_Negative = (P_Negative_Dont * P_Dont) / P_Negative

P_Dont_Negative
```

```
## [1] 0.999998
```

Part B: Conditional Expectations

Suppose you are interested in the relationship between sales and your advertising budget. You know that $E[Sales \mid Advertising] = \alpha + \beta \cdot Advertising$ (there is a linear relationship).

a. Write down the Law of Iterated Expectations.
b. If $E[Advertising] = 100$, what is $E[Sales]$?

Answers

a. $E[Sales] = E[E[Sales|Advertising]]$
b. $E[Sales] = \alpha + 100\beta$

Part C: Sampling Experiments

This problem asks you to conduct a series of sampling experiments and was adapted from the Mastering Metrics online materials. You should modify the below code.

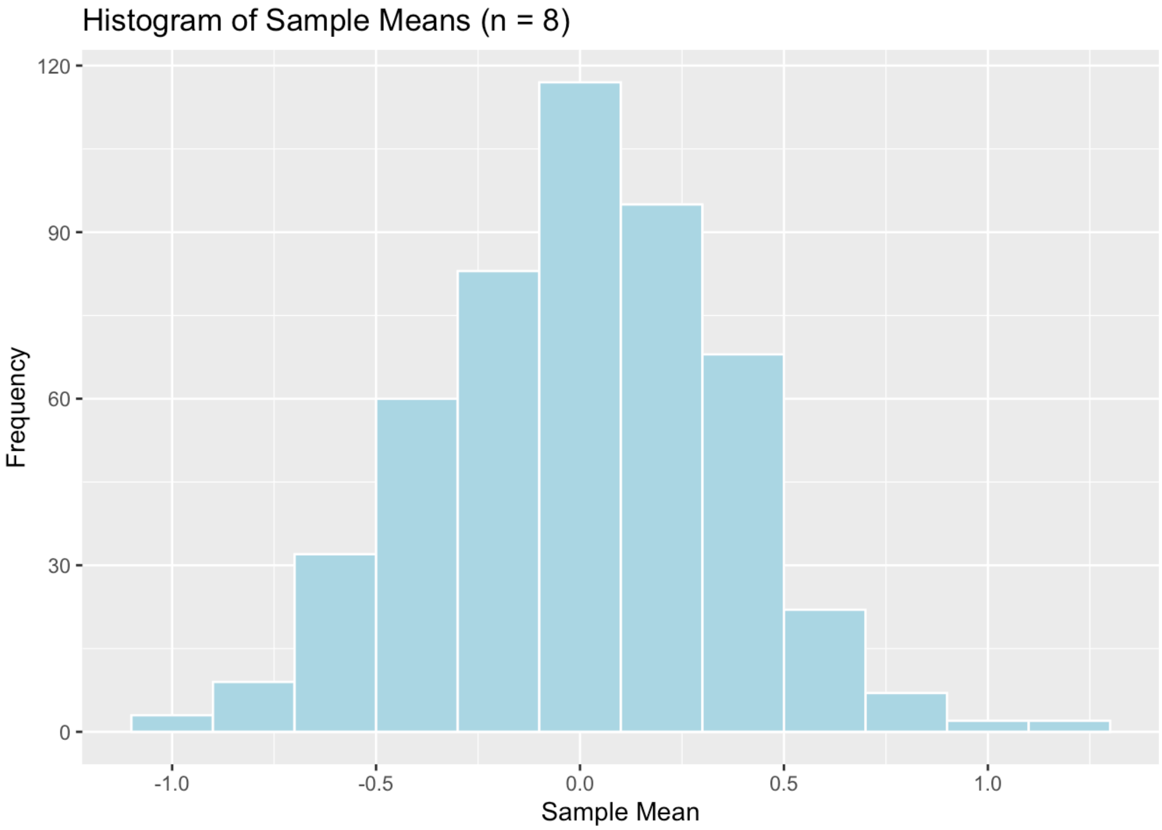
First, draw 500 random samples each with a sample size of 8 from a random number generator for a standard Normal distribution. Then, increase the sample size to 32 then increase the sample size to 128.

a. Plot histograms of the sampling distributions of (i) the sample mean and (ii) the sample variance, for each of these three sample sizes.
b. Your experiments produce "samples of sample means." What would you expect the mean and the variance of the sample means to be based on statistical theory?
c. Compute the mean and variance of the sample means generated by each experiment and compare them to what you predicted in part (c).

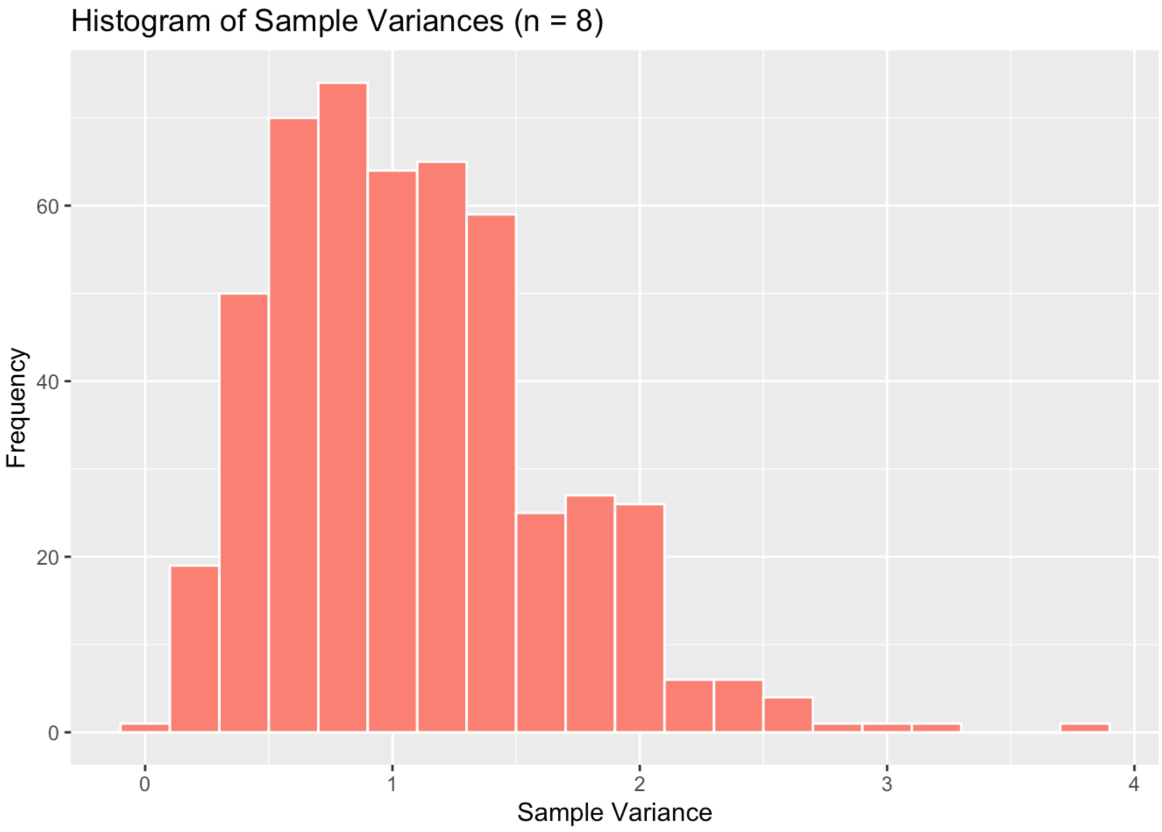
Answers

a.

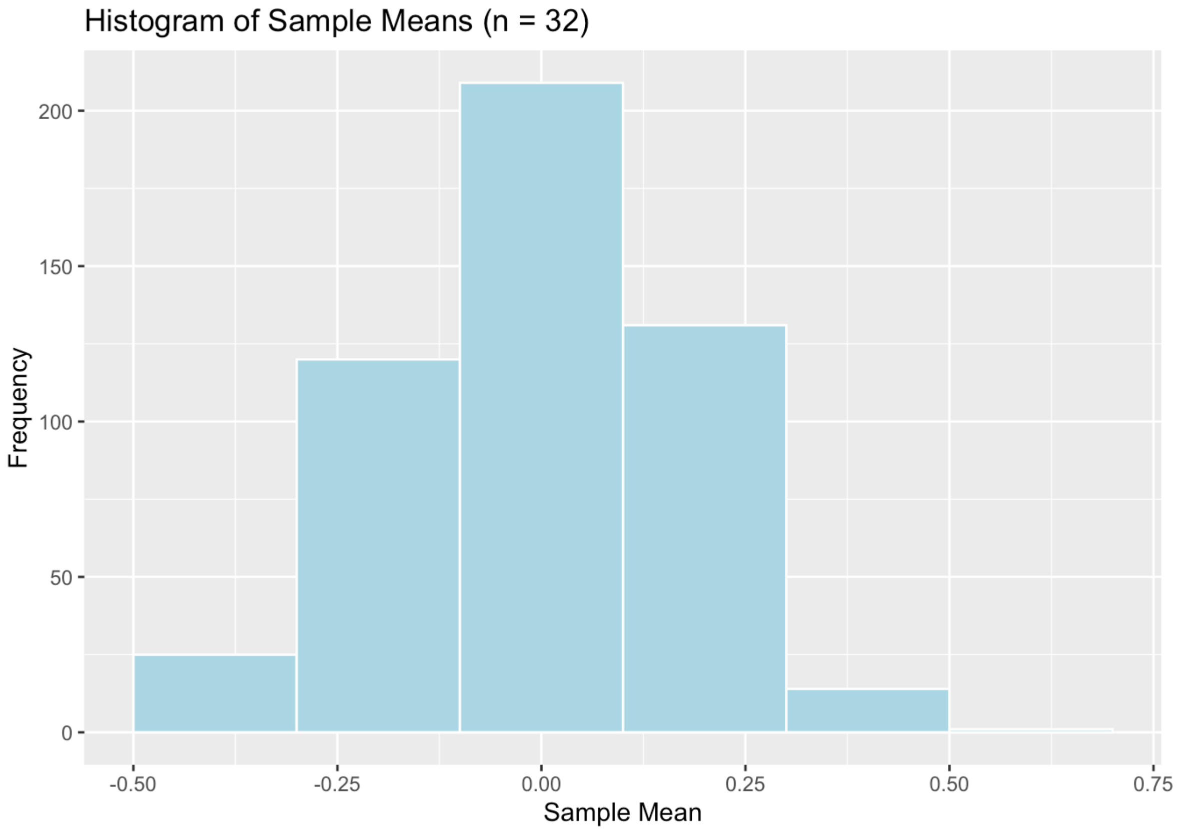
```
# Set the seed for reproducibility
set.seed(1)

# Set the number of samples
num_samples <- 500
sample_size_8 <- 8

# Create an empty matrix to store the samples
samples_array_8 <- matrix(NA, nrow = num_samples, ncol = sample_size_8)

# Generate the samples
for(i in 1:num_samples) {
  samples_array_8[i, ] <- rnorm(sample_size_8)
}

# Calculate the mean and variance
sample_means_8 <- apply(samples_array_8, 1, mean)
sample_variances_8 <- apply(samples_array_8, 1, var)

# Create a data frame for ggplot
df_means_8 <- data.frame(SampleMean = sample_means_8)
df_variances_8 <- data.frame(SampleVariance = sample_variances_8)

# Create a histogram for sample means
ggplot(df_means_8, aes(x = SampleMean)) +
  geom_histogram(binwidth = 0.2, fill = "lightblue", color = "white") +
  labs(title = "Histogram of Sample Means (n = 8)", x = "Sample Mean", y = "Frequency")
```



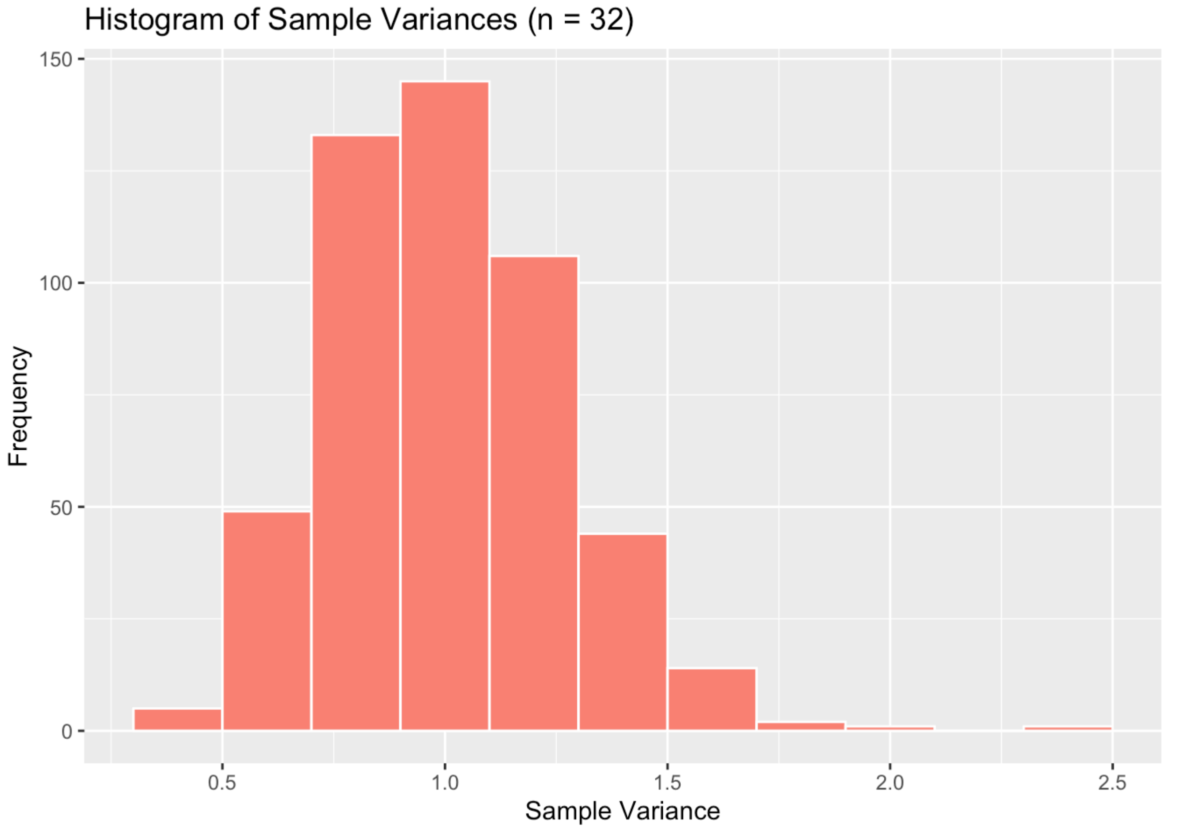
```
# Create a histogram for sample variances
ggplot(df_variances_8, aes(x = SampleVariance)) +
  geom_histogram(binwidth = 0.2, fill = "salmon", color = "white") +
  labs(title = "Histogram of Sample Variances (n = 8)", x = "Sample Variance", y = "Frequency")
```



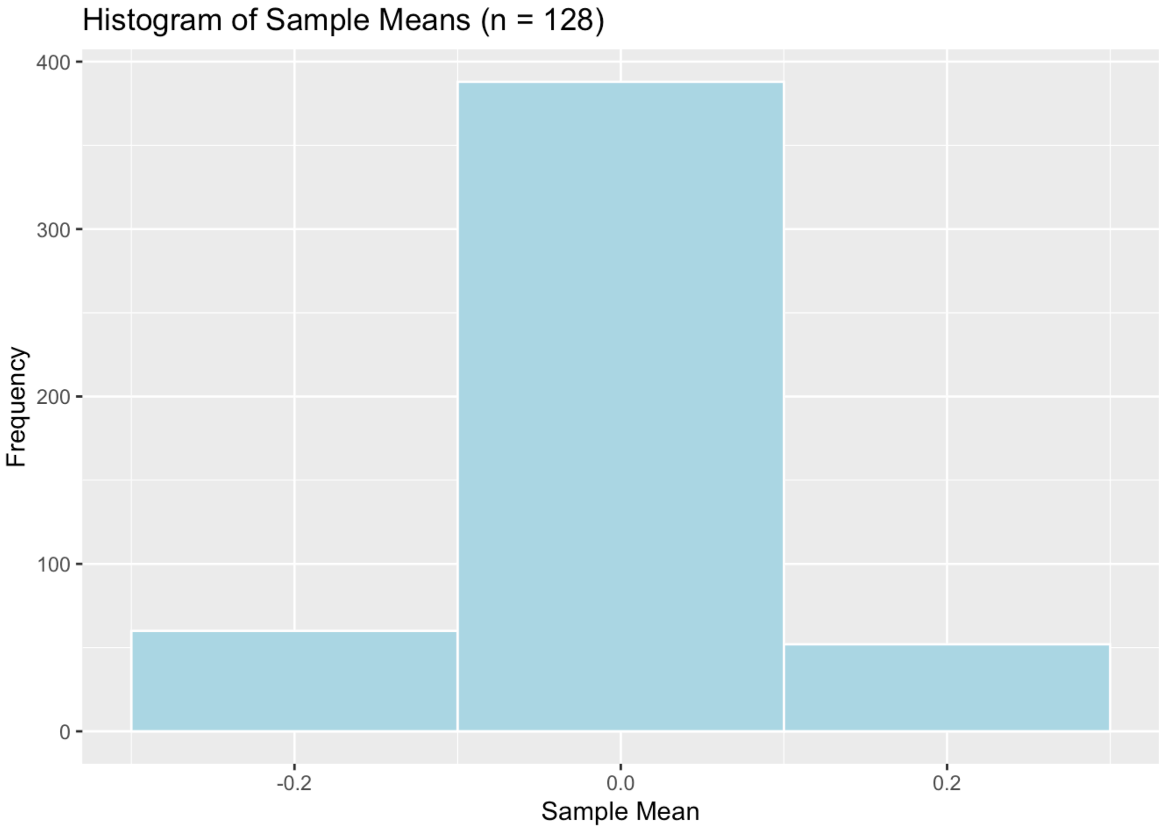
```
# Set the seed for reproducibility
set.seed(1)

# Set the number of samples
num_samples <- 500
sample_size_32 <- 32  # Change sample size to 32

# Create an empty matrix to store the samples
samples_array_32 <- matrix(NA, nrow = num_samples, ncol = sample_size_32)  # Adjusted for sample size 32

# Generate the samples
for(i in 1:num_samples) {
  samples_array_32[i, ] <- rnorm(sample_size_32)  # Using sample_size_32 here
}

# Calculate the mean and variance
sample_means_32 <- apply(samples_array_32, 1, mean)
sample_variances_32 <- apply(samples_array_32, 1, var)

# Create data frames for ggplot
df_means_32 <- data.frame(SampleMean = sample_means_32)
df_variances_32 <- data.frame(SampleVariance = sample_variances_32)

# Create a histogram for sample means
ggplot(df_means_32, aes(x = SampleMean)) +
  geom_histogram(binwidth = 0.2, fill = "lightblue", color = "white") +
  labs(title = "Histogram of Sample Means (n = 32)", x = "Sample Mean", y = "Frequency")
```



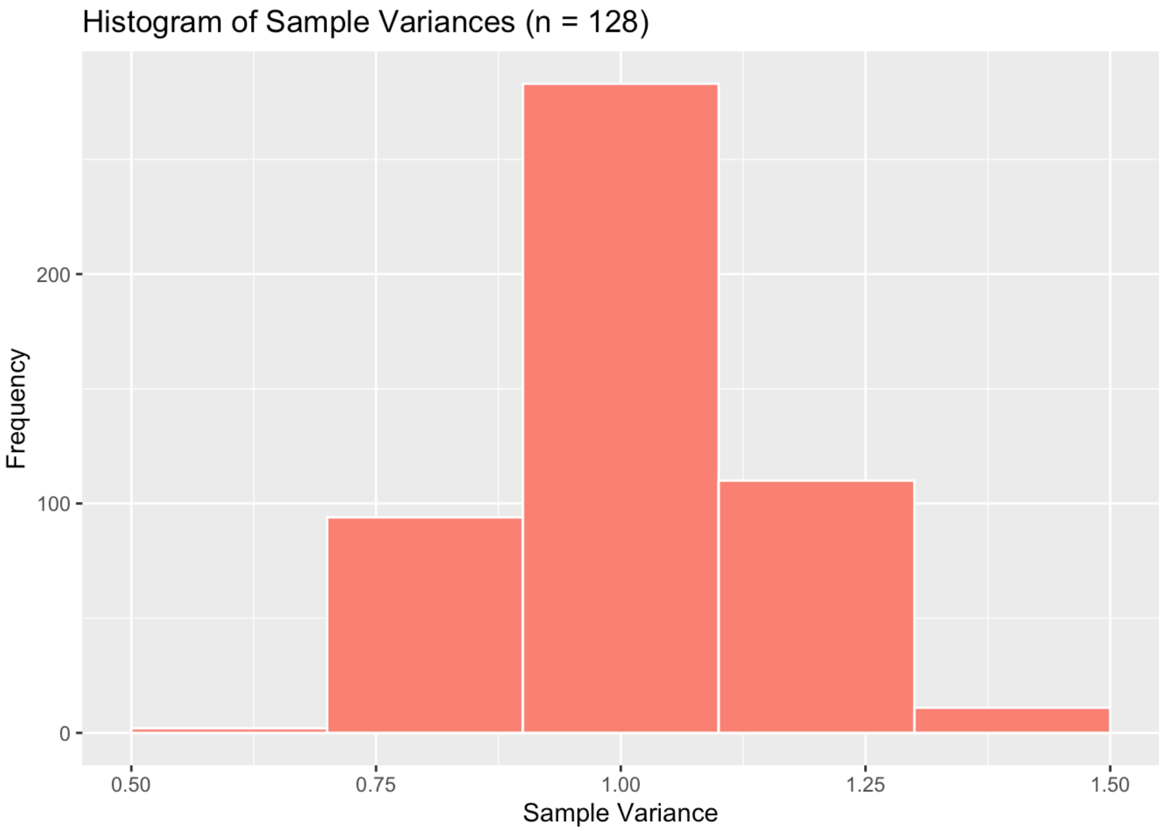
```
# Create a histogram for sample variances
ggplot(df_variances_32, aes(x = SampleVariance)) +
  geom_histogram(binwidth = 0.2, fill = "salmon", color = "white") +
  labs(title = "Histogram of Sample Variances (n = 32)", x = "Sample Variance", y = "Frequency")
```



```
# Set the seed for reproducibility
set.seed(1)

# Set the number of samples
num_samples <- 500
sample_size_128 <- 128  # Change sample size to 128

# Create an empty matrix to store the samples
samples_array_128 <- matrix(NA, nrow = num_samples, ncol = sample_size_128)  # Adjusted for sample size 128

# Generate the samples
for(i in 1:num_samples) {
  samples_array_128[i, ] <- rnorm(sample_size_128)  # Using sample_size_128 here
}

# Calculate the mean and variance
sample_means_128 <- apply(samples_array_128, 1, mean)
sample_variances_128 <- apply(samples_array_128, 1, var)

# Create data frames for ggplot
df_means_128 <- data.frame(SampleMean = sample_means_128)
df_variances_128 <- data.frame(SampleVariance = sample_variances_128)

# Create a histogram for sample means
ggplot(df_means_128, aes(x = SampleMean)) +
  geom_histogram(binwidth = 0.2, fill = "lightblue", color = "white") +
  labs(title = "Histogram of Sample Means (n = 128)", x = "Sample Mean", y = "Frequency")
```



```
# Create a histogram for sample variances
ggplot(df_variances_128, aes(x = SampleVariance)) +
  geom_histogram(binwidth = 0.2, fill = "salmon", color = "white") +
  labs(title = "Histogram of Sample Variances (n = 128)", x = "Sample Variance", y = "Frequency")
```



b. As sample size increases, the variance of the sample means should decrease, leading to a tighter concentration of the sample means around the population mean. The mean of sample means is close to zero due to the large number of samples taken. According to the Central Limit Theorem, which states that as the sample size becomes large, the sampling distribution of the sample mean approaches a normal distribution, regardless of the shape of the original population distribution. That's why the histograms center around 0, and their spread decreases as the sample size increases.

c.

```
sample_size = c(8, 32, 128)

#compute mean and sample of variance means
mean_sample_means = c(mean(sample_means_8), mean(sample_means_32), mean(sample_means_128))
variance_sample_means = c(var(sample_means_8), var(sample_means_32), var(sample_means_128))

#now combine into a data frame
results_table = data.frame(
  `Sample Size` = sample_size,
  `Mean of Sample Means` = mean_sample_means,
  `Variance of Sample Means` = variance_sample_means
)

print(results_table)
```

```
##   Sample.Size Mean.of.Sample.Means Variance.of.Sample.Means
## 1           8          0.001030307              0.127841873
## 2          32         -0.011028046              0.028619832
## 3         128         -0.003069978              0.006837496
```

The data in the table supports my prediction based on the Central Limit Theorem. As the sample size increases, the variance of sample means decreases, making the distribution tighter and more centered around the population mean. This confirms that the sampling distribution becomes increasingly normal with larger sample sizes, consistent with theoretical expectations, even though the original distribution is already normal.

Part D: Hypothesis Testing

This question explores hypothesis testing. It uses data from Project STAR, a randomized control trial that evaluated the effects of class size on student outcomes. We will return to Project STAR later in the course.

a. Check out the contents of the data set using the `glimpse` function. How many observations and variables does it contain? What is the gender of the third child in the sample? What is their kindergarten math test score and to what sort of “STAR” class were they assigned while in kindergarten?
b. Create a total test score variable (the sum of the reading and math score) for each grade, K-3. Plot the `scorek` distributions for the small and regular classes using `ggplot`. Write a sentence that summarizes what they indicate.
c. Use the `st` function to construct a table of test-score means and standard deviations for each class type (“small”, “regular” and “regular+aide”) by grade, based on the kindergarten assignment.
d. Write a sentence that states the difference between small and regular-class average scores for kindergartners. Write a sentence that expresses this difference in terms of the standard deviation of regular-class scores.
e. Using a t-test, evaluate whether these two means are statistically different from one another. Define what the null and alternative hypotheses that you are testing are. Write a sentence that interprets the p-value from this test.

```
# Load data.
data(STAR)

# (a)
# Insert a `glimpse` command to check the contents of `STAR`.
glimpse(STAR)
```

```
## Rows: 11,598
## Columns: 47
## $ gender      <fct> female, female, female, male, male, male, male, female, ma…
## $ ethnicity   <fct> afam, cauc, afam, cauc, afam, cauc, afam, cauc, cauc, cauc…
## $ birth       <yearqtr> 1979 Q3, 1980 Q1, 1979 Q4, 1979 Q4, 1980 Q1, 1979 Q3, …
## $ stark       <fct> NA, small, small, NA, regular+aide, NA, NA, NA, NA, NA, re…
## $ star1       <fct> NA, small, small, NA, NA, NA, NA, regular+aide, regular, r…
## $ star2       <fct> NA, small, regular+aide, NA, NA, regular, NA, regular+aide…
## $ star3       <fct> regular, small, regular+aide, small, NA, regular, regular+…
## $ readk       <int> NA, 447, 450, NA, 439, NA, NA, NA, NA, NA, 448, 447, 431, …
## $ read1       <int> NA, 507, 579, NA, NA, NA, NA, 475, NA, 651, 651, 533, 558,…
## $ read2       <int> NA, 568, 588, NA, NA, NA, NA, 573, NA, 596, 614, 608, 608,…
## $ read3       <int> 580, 587, 644, 686, NA, 644, NA, 599, NA, 626, 641, 665, 5…
## $ mathk       <int> NA, 473, 536, NA, 463, NA, NA, NA, NA, NA, 559, 489, 454, …
## $ math1       <int> NA, 538, 592, NA, NA, NA, NA, 512, NA, 532, 584, 545, 553,…
## $ math2       <int> NA, 579, 579, NA, NA, NA, NA, 550, NA, 590, 639, 603, 579,…
## $ math3       <int> 564, 593, 639, 667, NA, 648, NA, 583, NA, 618, 684, 648, 5…
## $ lunchk      <fct> NA, non-free, non-free, NA, free, NA, NA, NA, NA, NA, non-…
## $ lunch1      <fct> NA, free, NA, NA, NA, NA, NA, non-free, non-free, non-free…
## $ lunch2      <fct> NA, non-free, non-free, NA, NA, non-free, NA, non-free, no…
## $ lunch3      <fct> free, free, non-free, non-free, NA, non-free, free, non-fr…
## $ schoolk     <fct> NA, rural, suburban, NA, inner-city, NA, NA, NA, NA, NA, r…
## $ school1     <fct> NA, rural, suburban, NA, NA, NA, NA, rural, rural, rural, …
## $ school2     <fct> NA, rural, suburban, NA, NA, rural, NA, rural, rural, rura…
## $ school3     <fct> suburban, rural, suburban, rural, NA, rural, inner-city, r…
## $ degreek     <fct> NA, bachelor, bachelor, NA, bachelor, NA, NA, NA, NA, NA, …
## $ degree1     <fct> NA, bachelor, master, NA, NA, NA, NA, master, master, bach…
## $ degree2     <fct> NA, bachelor, bachelor, NA, NA, bachelor, NA, master, bach…
## $ degree3     <fct> bachelor, bachelor, bachelor, bachelor, NA, bachelor, bach…
## $ ladderk     <fct> NA, level1, level1, NA, probation, NA, NA, NA, NA, NA, lev…
## $ ladder1     <fct> NA, level1, probation, NA, NA, NA, NA, apprentice, level1,…
## $ ladder2     <fct> NA, apprentice, level1, NA, NA, notladder, NA, level1, lev…
## $ ladder3     <fct> level1, apprentice, level1, level1, NA, level1, notladder,…
## $ experiencek <int> NA, 7, 21, NA, 0, NA, NA, NA, NA, NA, 16, 5, 8, 17, NA, NA…
## $ experience1 <int> NA, 7, 32, NA, NA, NA, NA, 8, 13, 7, 11, 15, 0, 5, NA, 17,…
## $ experience2 <int> NA, 3, 4, NA, NA, 13, NA, 13, 6, 8, 31, 14, 9, NA, 4, 28, …
## $ experience3 <int> 30, 1, 4, 10, NA, 15, 17, 23, 8, 8, 7, 14, 8, NA, 19, 13, …
## $ tethnicityk <fct> NA, cauc, cauc, NA, cauc, NA, NA, NA, NA, NA, cauc, cauc, …
## $ tethnicity1 <fct> NA, cauc, afam, NA, NA, NA, NA, cauc, cauc, cauc, cauc, ca…
## $ tethnicity2 <fct> NA, cauc, afam, NA, NA, cauc, NA, afam, cauc, cauc, cauc, …
## $ tethnicity3 <fct> cauc, cauc, cauc, cauc, NA, cauc, afam, cauc, cauc, cauc, …
## $ systemk     <fct> NA, 30, 11, NA, 11, NA, NA, NA, NA, NA, 35, 41, 4, 11, NA,…
## $ system1     <fct> NA, 30, 11, NA, NA, NA, NA, 4, 40, 21, 35, 41, 4, 11, NA, …
## $ system2     <fct> NA, 30, 11, NA, NA, 6, NA, 4, 40, 21, 35, 41, 4, NA, 17, 2…
## $ system3     <fct> 22, 30, 11, 6, NA, 6, 11, 4, 40, 21, 35, 41, 4, NA, 17, 20…
## $ schoolidk   <fct> NA, 63, 20, NA, 19, NA, NA, NA, NA, NA, 69, 79, 5, 16, NA,…
## $ schoolid1   <fct> NA, 63, 20, NA, NA, NA, NA, 5, 77, 50, 69, 79, 5, 16, NA, …
## $ schoolid2   <fct> NA, 63, 20, NA, NA, 8, NA, 5, 77, 50, 69, 79, 5, NA, 41, 4…
## $ schoolid3   <fct> 54, 63, 20, 8, NA, 8, 31, 5, 77, 50, 69, 79, 5, NA, 41, 48…
```

```
third_child_gender <- STAR$gender[3]
third_child_math_score <- STAR$tmathssk[3]
third_child_class_type <- STAR$classtype[3]

# (b)
# Create total score variables by filling in this mutate command.
STAR2 <- STAR %>%
  mutate(scorek = readk + mathk,
         score1 = read1 + math1,
         score2 = read2 + math2,
         score3 = read3 + math3)


# Plot test-score distributions.
ggplot(data = subset(STAR2, stark=='small' | stark=='regular'),
       aes(x=scorek, fill=as.factor(stark), color=as.factor(stark))) +
  geom_density(alpha=0.4) +
  labs(title="Figure 1. Distribution of scorek by class type")
```

```
## Warning: Removed 351 rows containing non-finite outside the scale range
## (`stat_density()`).
```

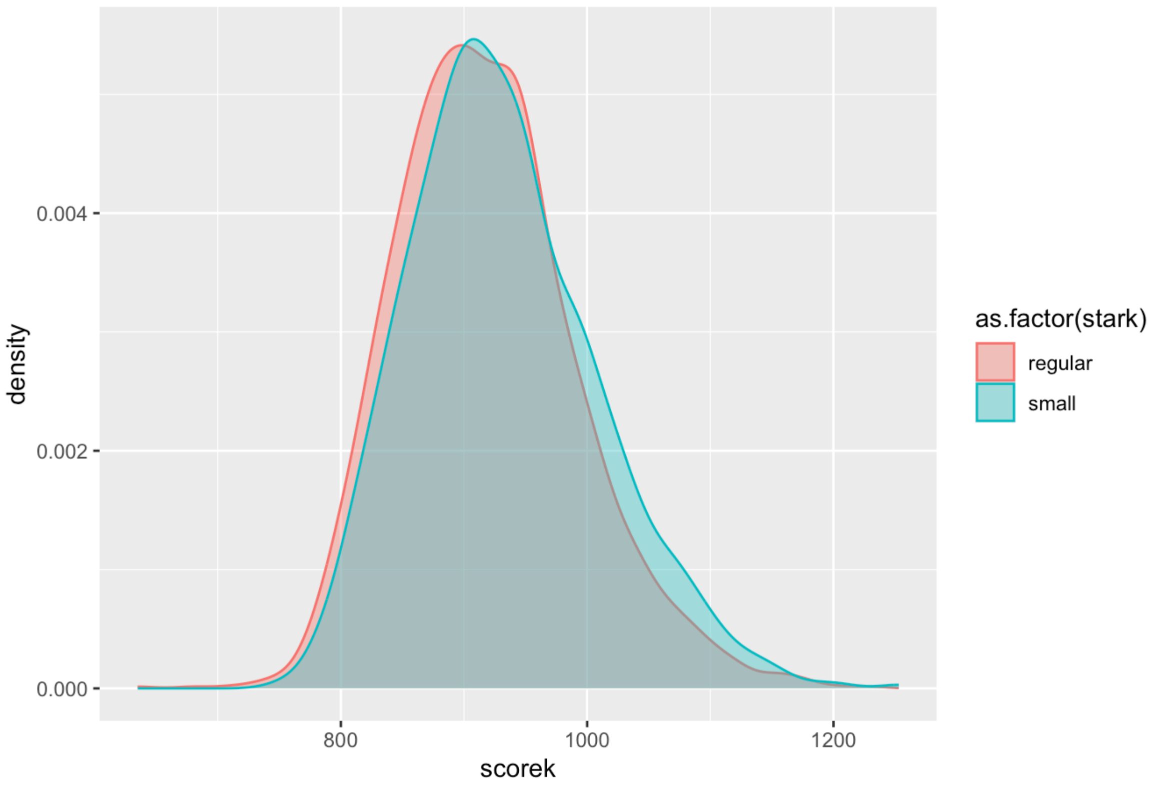


Figure 1. Distribution of scorek by class type

```
# (c)
# Report test-score summary statistics for each grade
# based on K assignment.

sumtable(STAR2, vars=c('scorek','score1', 'score2', 'score3'),
   group='stark', title="Test scores by class type")
```

Test scores by class type

stark	regular			small			regular+aide		
Variable	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**
scorek	2005	918	73	1738	932	76	2043	918	71
score1	1456	1057	91	1339	1076	95	1503	1054	91
score2	1201	1179	83	1080	1189	85	1183	1175	83
score3	1047	1247	70	937	1258	73	1021	1247	73

```
# (e)
# Use a t-test to calculate whether the kindergarten scores are statistically
# different from one another.

t.test(scorek ~ stark, data = subset(STAR2, stark=='small' | stark=='regular'))
```

```
##
##  Welch Two Sample t-test
##
## data:  scorek by stark
## t = -5.6635, df = 3616, p-value = 1.598e-08
## alternative hypothesis: true difference in means between group regular and group small is not equa
l to 0
## 95 percent confidence interval:
##  -18.710595  -9.087394
## sample estimates:
## mean in group regular   mean in group small
##              918.0429              931.9419
```

Answers

a. Observations: 11598, variables: 47, Third child: female, Math test score: 536, STAR class: small

b. This indicates that scores from small classes are slightly higher on average and exhibit a marginally narrower distribution, suggesting that students in small classes might perform slightly better and more consistently than those in regular classes. The overlap of the distributions shows that while class size may influence outcomes, other factors also may play significant roles in student performance.

c. Check the table output from part C.

d. The small-class average scores for kindergartners is 932 and regular-class average scores for kindergartners is 918. The difference is 14.The standard deviation for regular class size is 73. To express the difference in terms of standard deviation, 14/73 = 0.19

e. t = -5.6635, df = 3616, p-value = 1.598e-08

Given the p-value is significant at the 0.01 level, you reject the null hypothesis meaning there is an effect on class.