Introduction to Econometrics
PS #2 - Probability review

1 Variance and Covariance Algebra

1. Using $V(X) = E\Big\{[X - E(X)]^2\Big\}$ show that $V(X) = E(X^2) - E(X)^2$
2. Using $Cov(X,Y) = E\Big\{[X - E(X)][Y - E(Y)]\Big\}$ show that $Cov(X,Y) = E(XY) - E(X)E(Y)$
3. Using the above results, show that $V(X+Y) = V(X) + V(Y) + 2cov(X,Y)$

2 Joint Distributions (Refresher)

Consider the following joint distribution of weather conditions and commuting times:

	Rain ($X=0$)	No Rain ($X=1$)	$P(Y)$
Long commute ($Y=0$)	0.15	0.07	0.22
Short commute ($Y=1$)	0.15	0.63	0.78
$P(X)$	0.3	0.7	1.00

1. Compute $E(X)$ and $E(Y)$
2. Compute $V(X)$ and $V(Y)$
3. Compute $cov(X,Y)$ and $Corr(X,Y)$.
4. Define $p \equiv Pr(Y=1)$, show that $E(Y) = p$, and $V(Y) = p(1-p)$.

We now define a new random variable: $U = 3X + 2Y + 5$.

7. Compute $E(U)$ and $V(U)$

3 Law of Iterated Expectations

Consider the following data on the fitness levels (F) and gym membership (G) for some hypothetical population made up of only 3 individuals:

i	F	G
1	10	Y
2	15	N
3	20	Y

1. Compute the average fitness level in the population *i.e.*, $E(F)$
2. Recompute the average fitness level in the population using the law of iterated expectations.

4 Population Parameters and Sample Statistics

Assume that we have a class of 1000 students. The instructor is interested in estimating the mean height of the students in the class. However, due to time constraints, the instructor cannot measure the height of every student. Instead, the instructor decides to gather data from a smaller randomly chosen group of students.

Let X be the random variable representing the height of a randomly selected student. Suppose that the instructor collects a sample of $n \leq 1000$ students from the class.

4.1 The sample average

1. What is the population and sample in this scenario?
2. Let X_i be the RV which denotes the height of the i-th randomly selected student. Are the random variables $X_1, X_2, .., X_n$ independently and identically distributed?
3. Suppose that the instructor draws a random sample of 5 students, with the following realisations:

X_1	X_2	X_3	X_4	X_5
180	140	210	160	190

   What is the sample average height *i.e.*, $\bar{X}$?
4. The instructor is careless and forgets where he stored the data collected above. He recollects a sample of 5 students again, with the following realisations:

X_1	X_2	X_3	X_4	X_5
195	170	180	220	200

   What is the sample average height *i.e.*, $\bar{X}$ from this newly drawn sample? What can you conclude about $\bar{X}$ from one random sample to another?
5. Characterize the sampling distribution of $\bar{X}$ by stating the $E(\bar{X})$ and $V(\bar{X})$.

4.2 The law of large numbers (LLN)

Suppose that the instructor is extremely lazy. Consequently, he decides to hire someone to estimate the mean height of the class. This person has a lot of free time and can draw large sample of students.

1. How do you expect the sample average $\bar{X}$ to change if this person draws a sample of $n = 100, 500$ or 800?
2. What happens to the $V(\bar{X})$ as the sample size n increases?

4.3 Example in R

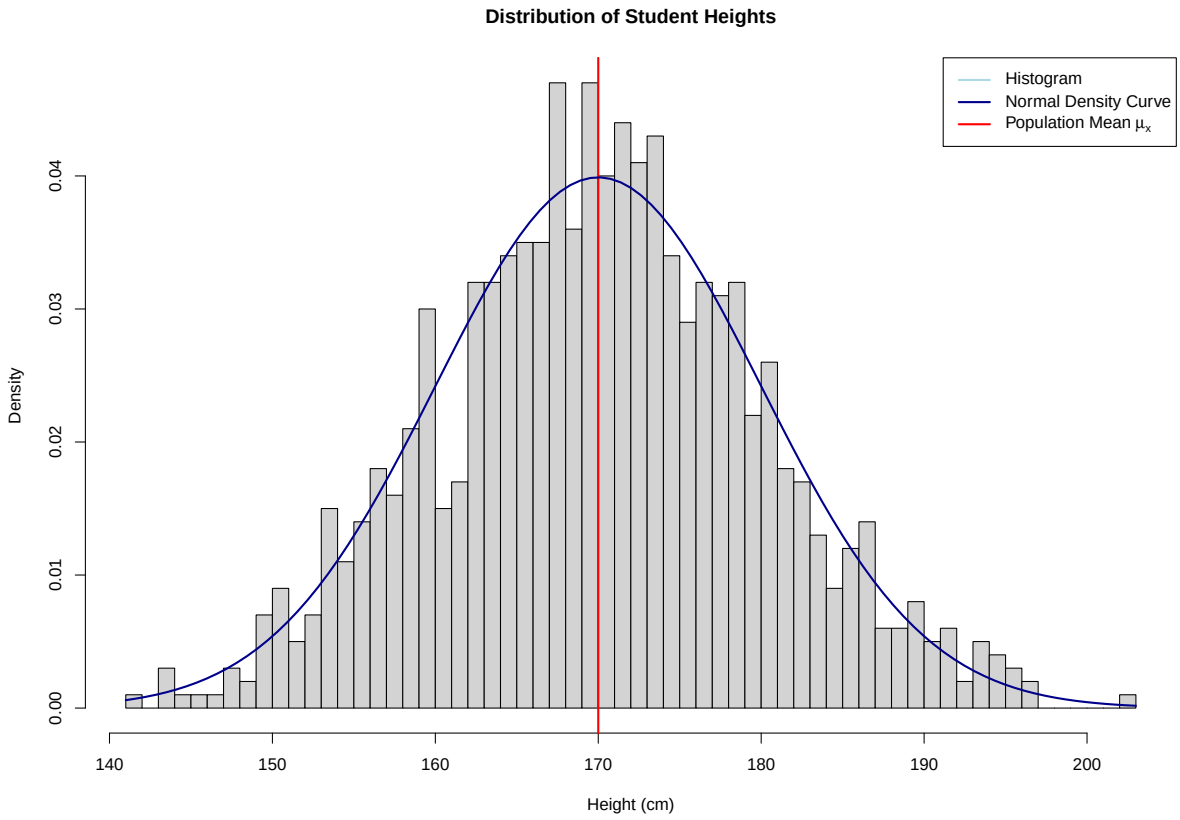
Suppose that the height of the class is normally distributed with $\mu_x = 170$ and $\sigma_x^2 = 10$ *i.e.*, $X \sim \mathcal{N}(170, 10)$

1. Simulate the population of student heights

```
# Set seed for reproducibility
set.seed(123)

# Simulate the population of student heights (in cm)
population_size <- 1000
population_heights <- rnorm(population_size, mean = 170, sd = 10)
```

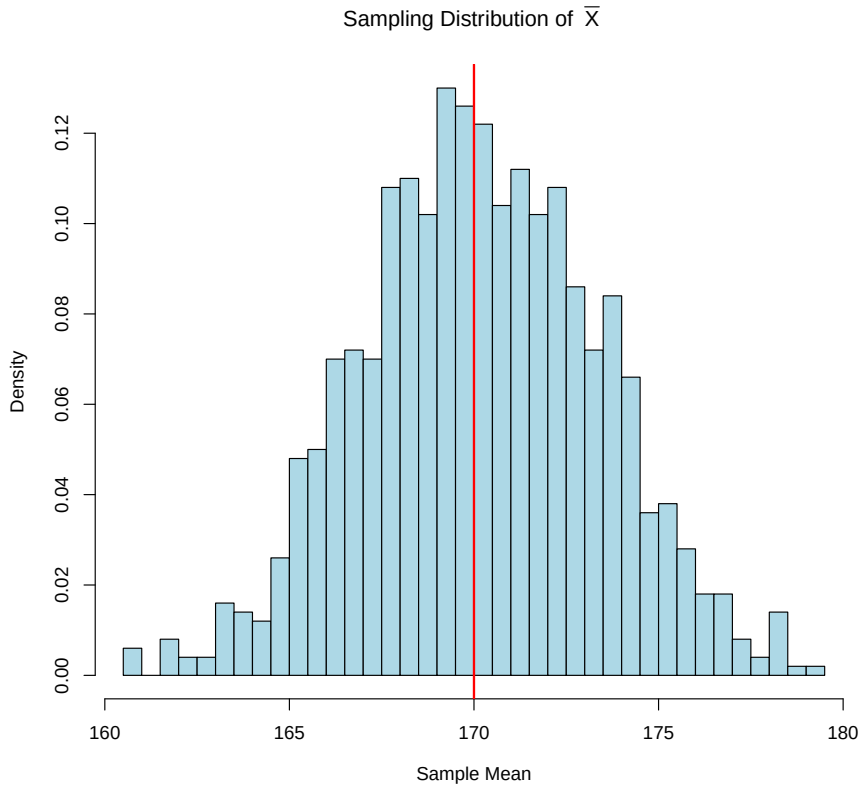
2. Plot the population distribution

```
# Plot the histogram of population heights
hist(population_heights, breaks = 50, probability = TRUE, col = "lightgrey",
     main = "Distribution of Student Heights", xlab = "Height (cm)",
     ylab = "Density")

# Add a vertical line at the population mean
abline(v = 170, col = "red", lwd = 2)

# Overlay a normal density curve
curve(dnorm(x, mean = 170, sd = 10), add = TRUE, col = "darkblue", lwd = 2)

# Add a legend
legend("topright", legend = c("Histogram", "Normal Density Curve", expression("
    Population Mean" ~ mu[x])),
       col = c("lightblue", "darkblue", "red"), lty = c(1, 1, 1), lwd = 2)
```



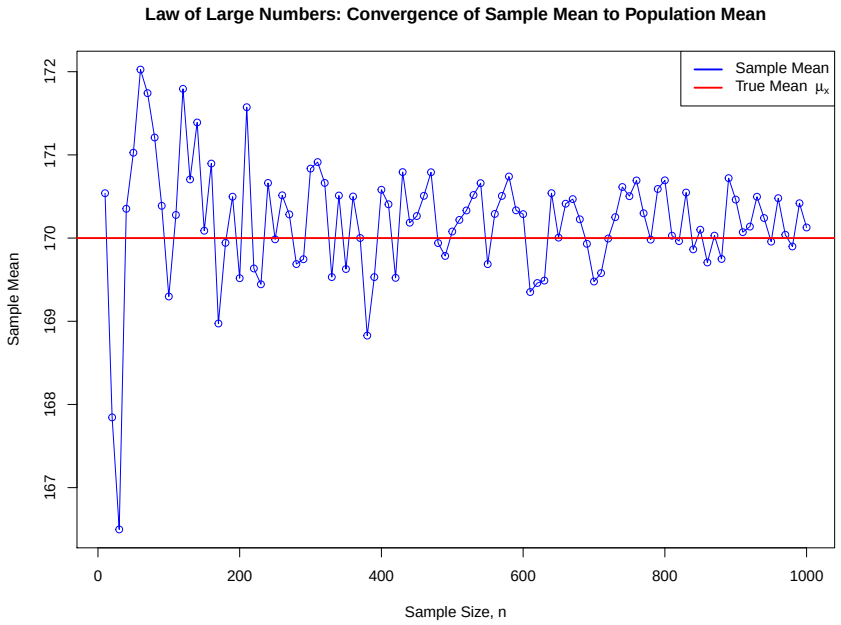
3. Draw samples of size 10 for 1000 times and plot the sampling distribution of the sample average $\bar{X}$

```r
# Number of samples to draw
num_samples <- 1000
sample_size <- 10

# Initialize a vector to store sample means
sample_means <- numeric(num_samples)

# Draw 1000 samples of size 10 and compute the sample means
for (i in 1:num_samples) {
  sample <- sample(population_heights, size = sample_size, replace = TRUE)
  sample_means[i] <- mean(sample)
}

# Plot the sampling distribution of the sample means
hist(sample_means, breaks = 50, probability = TRUE, col = "lightblue",
     main = expression("Sampling Distribution of " ~ bar(X) ~ ""),
     xlab = expression("Sample Mean"), ylab = "Density")

# Add a vertical line at the population mean
abline(v = mean(population_heights), col = "red", lwd = 2)
```



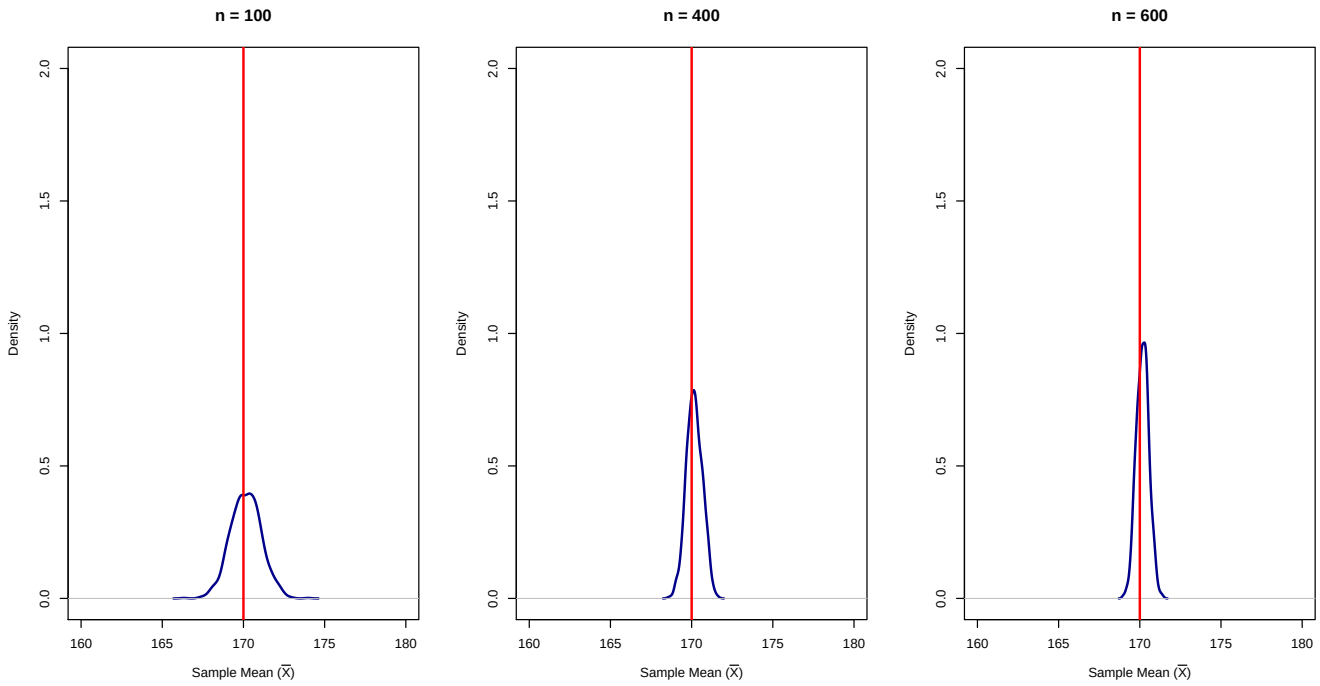
4. Draw random samples of size $n = 10, 20, ..., 1000$. Compute the sample averages for each, and compare with the population mean $\mu_x = 170$

```r
# Set seed for reproducibility
set.seed(123)

# True population mean
true_mean <- 170

# Simulate taking random samples of increasing size
sample_sizes <- seq(10, 1000, by = 10) # Sample sizes from 10 to 1000, in steps
    of 10

# Initialize a vector to store sample means
sample_means <- numeric()

# For each sample size, take a sample from the population and compute the mean
for (n in sample_sizes) {
  sample <- sample(population_heights, size = n, replace = TRUE)
  sample_means <- c(sample_means, mean(sample))
}

# Plot the sample means and show how they converge to the true population mean
plot(sample_sizes, sample_means, type = "o", col = "blue",
     xlab = "Sample Size, n", ylab = "Sample Mean",
     main = "Law of Large Numbers: Convergence of Sample Mean to Population Mean
         ")
abline(h = true_mean, col = "red", lwd = 2) # Add a horizontal line for the true
     population mean
legend("topright", legend = c("Sample Mean", expression("True Mean " ~ mu[x] ~ "
    ")),
       col = c("blue", "red"), lty = 1, lwd = 2)
```



5. Plot the sampling distribution of $\bar{X}$ for $n = 100, 400$ and 600. Comment on the variance.

```r
# Function to plot only the density curve for a given sample size 'n'
plot_density_curve <- function(n, num_samples, xlim_range, ylim_range) {
  # Initialize a vector to store sample means
  sample_means <- numeric(num_samples)

  # Draw 'num_samples' samples of size 'n' and compute the sample means
  for (i in 1:num_samples) {
    sample <- sample(population_heights, size = n, replace = TRUE)
    sample_means[i] <- mean(sample)
  }

  # Plot only the empirical density curve based on the sample means
  plot(density(sample_means), col = "darkblue", lwd = 2, xlim = xlim_range, ylim
       = ylim_range,
       main = paste("n =", n),
       xlab = expression(paste("Sample Mean (", bar(X), ")")), ylab = "Density")

  # Add a vertical line at the population mean
  abline(v = 170, col = "red", lwd = 2)

}

# Set number of samples to draw
num_samples <- 1000

# Set the same x and y limits for all plots
xlim_range <- c(160, 180)  # Define x-axis limits based on your data
ylim_range <- c(0, 2)     # Adjusted y-axis limits for the density

# Create a 1x3 grid for the plots
par(mfrow = c(1, 3))  # 1 row, 3 columns

# Plot the density curve for n = 50
plot_density_curve(100, num_samples, xlim_range, ylim_range)

# Plot the density curve for n = 300
plot_density_curve(400, num_samples, xlim_range, ylim_range)

# Plot the density curve for n = 600
plot_density_curve(600, num_samples, xlim_range, ylim_range)
```


n = 100
n = 400
n = 600
Density
0.0
0.5
1.0
1.5
2.0
160
165
170
175
180
Sample Mean ($\bar{X}$)