Introduction to Econometrics
PS #3 - Statistics Review

1 First moments of a continuous RV

1. Let $X \sim \text{Uniform}(0,1)$. Find and draw the PDF and CDF of X. Compute $E(X)$ and $\text{Var}(X)$.
2. Let Y have range $[0,2]$ and density $\frac{3}{8}y^2$. Draw the PDF of Y. Compute $E(Y)$.

2 Sampling

Assume that X_1, X_2, X_3 is a sample of observations drawn from $\mathcal{N}(\mu_x, \sigma_x^2)$, with sample average $\bar{X}$. Suppose further that the three RV's are not independent. In particular:

$$\text{Cov}(X_1, X_2) = \text{Cov}(X_2, X_3) = \text{Cov}(X_1, X_3) = 0.5\sigma_X^2$$

1. Compute $E(\bar{X})$.
2. Compute $\text{Var}(\bar{X})$.

3 Properties of estimators

Suppose $X_1, X_2, ..., X_n$ is a random sample drawn from $N(\mu, \sigma^2)$. Consider the following estimator of μ:

$$\hat{X} = \frac{X_1 + X_2 + X_3}{3}$$

1. Show that $\hat{X}$ is a linear estimator.
2. Show that $\hat{X}$ is an unbiased estimator.
3. Compute the variance of the estimator.

Consider now the following weighted estimator:

$$\tilde{X} = \frac{3X_1}{8} + \frac{X_2}{2} + \frac{X_3}{8}$$

4. Show that $\tilde{X}$ is also an unbiased estimator.
5. Compute the variance of this new estimator. Is it more or less efficient than $\hat{X}$?
6. Assuming $\sigma^2 = 9$, calculate for each estimator the probability it is within one unit on either side of μ. Compare and comment.
7. Could you think of an estimator that is more efficient than the ones proposed above?

4 Central Limit Theorem (CLT)

Given a random sample $X_1, X_2, .., X_n$. Under general conditions, the CLT states that the sample average $\bar{X}$ will be well approximated by a $\mathcal{N}(\mu, \sigma^2)$ as n becomes large.

1. (Use **R**) Assume that in the population $X \sim \mathcal{U}nif[0,1]$. Draw random samples of different sizes and plot the resulting sampling distribution of $\bar{X}$.
2. (Use **R**) Assume that in the population $X \sim \text{Exp}(\lambda)$, such that $f(x;\lambda) = \lambda e^{-\lambda x}$ with $E(X) = 1/\lambda$ and $V(X) = 1/\lambda^2$. Suppose $\lambda = 1$, draw random samples of different sizes and plot the resulting sampling distribution of $\bar{X}$.
3. Assume that $X \sim N(\mu, \sigma^2)$ is the population distribution. Do we still need to rely on the CLT to tell us about the sampling distribution of $\bar{X}$?

```r
#Clear workspace
rm(list=ls())

# Set seed for reproducibility
set.seed(123)

# Define sample sizes and draws
sample_sizes <- c(2, 5, 10, 30, 50, 100)
n_samples <- 1000                    # Number of samples to draw for each size
x_range <- seq(0, 1, length.out = 100)    # Range for x-axis

# Function to draw samples and calculate sample means
draw_sample_means <- function(sample_size, n_samples) {
  sample_means <- replicate(n_samples, mean(runif(sample_size, 0, 1)))
  return(sample_means)
}

# Initialize empty plot area
par(mfrow = c(2, 3))  # Arrange plots in a 2x3 grid

# Loop over each sample size, draw the samples, and plot the histogram
for (size in sample_sizes) {
  sample_means <- draw_sample_means(size, n_samples)
  hist(sample_means, breaks = 30, main = paste("n =", size),
       xlim = c(0, 1), col = "lightblue", xlab = "",
       freq = FALSE, ylab = "Density")
  curve(dnorm(x, mean = 0.5, sd = sqrt(1/(12*size))),
        col = "red", add = TRUE, lwd = 2)  # Add normal curve for comparison
}
```

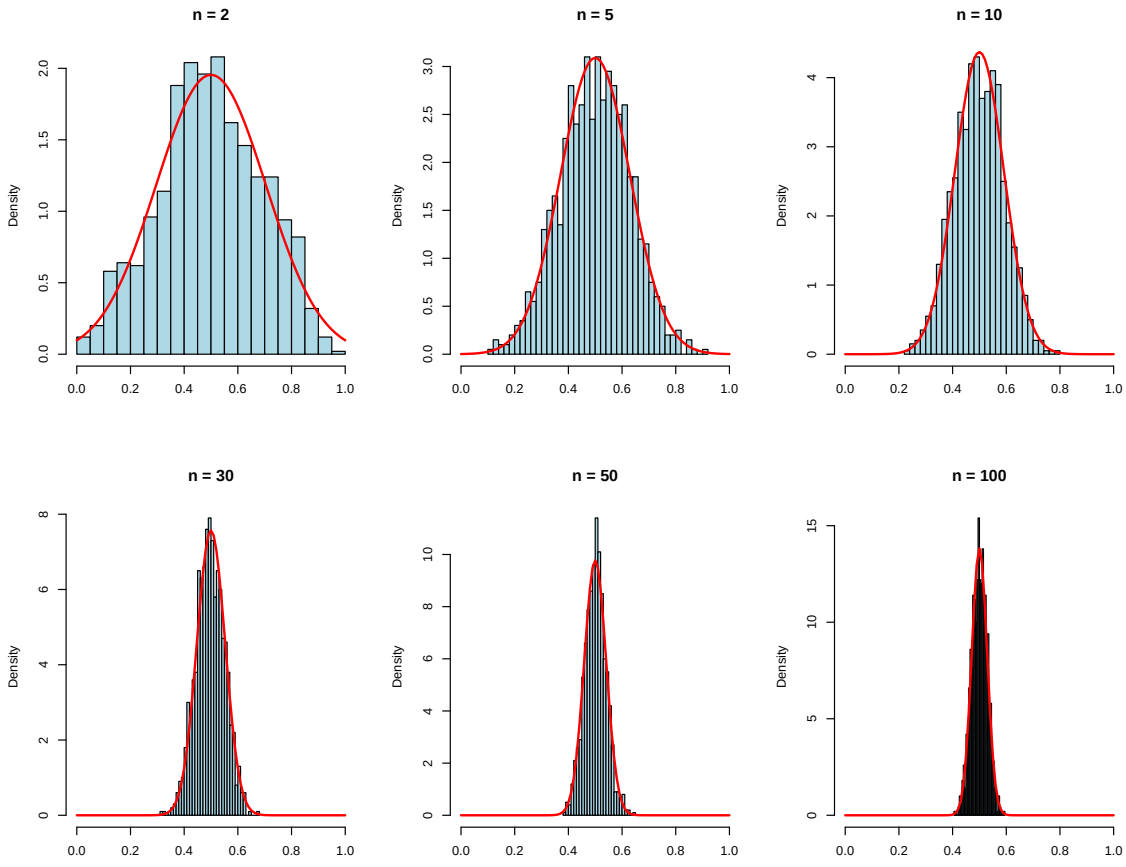


Figure 1: Sampling Distribution of $\bar{X}$ with varying n, $X \sim \mathcal{U}nif[0,1]$

```r
# Function to draw samples and calculate sample means
draw_sample_means_exp <- function(sample_size, n_samples) {
  sample_means <- replicate(n_samples, mean(rexp(sample_size, rate = 1)))  #
      Exponential dist with rate=1
  return(sample_means)
}

# Loop over each sample size, draw the samples, and plot the histogram
for (size in sample_sizes) {
  sample_means <- draw_sample_means_exp(size, n_samples)
  hist(sample_means, breaks = 30, main = paste("n =", size),
       xlim = c(0, 3), col = "lightblue", xlab = "Sample Means",
       freq = FALSE, ylab = "Density")
  curve(dnorm(x, mean = 1, sd = sqrt(1/size)),  # Theoretical normal distribution
        col = "red", add = TRUE, lwd = 2)  # Add normal curve for comparison
}
```

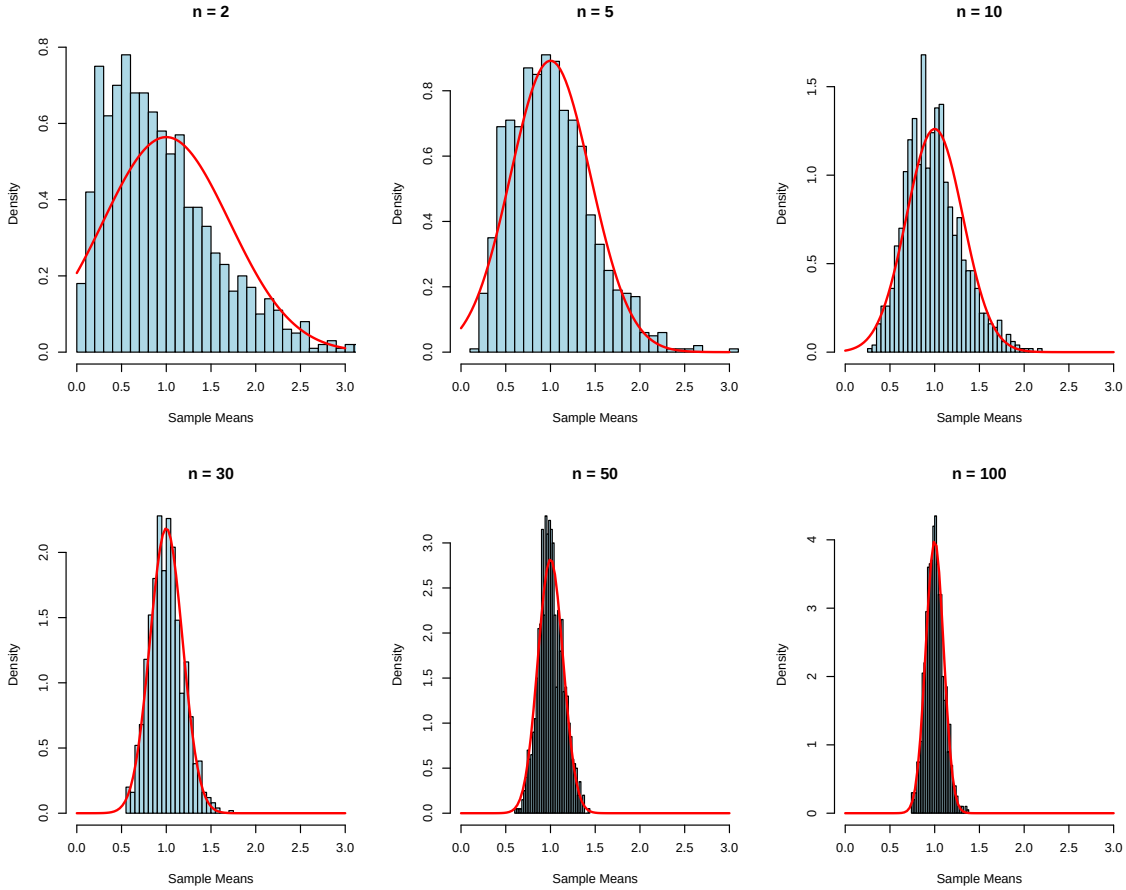


Figure 2: Sampling Distribution of $\bar{X}$ with varying n, $X \sim \text{Exp}\ (\lambda = 1)$,

5 Hypothesis testing

The same Instructor of Statistics you had previously met (in Problem Set #1) is now wondering whether his students are as diligent as he would expect. In particular, he expects them to solve more than three exercises by themselves for each exercise he solves in class. This means that for six exercises he solves in class, the students are expected to solve more than 18 by themselves. The instructor randomly selects eight students from the class and asks how many exercises they solved by themselves. Their answers are 3, 9, 12, 12, 18, 18, 24, 36.

1. Assuming that the number of exercises students solve by themselves follows a normal distribution, can the professor conclude at the 0.05 level of significance that the students are solving on average more than 18 exercises per class?

(a) State the null and alternate hypothesis *i.e.*, H_o and H_a

(b) Define the rejection region. What is the Type-I error?

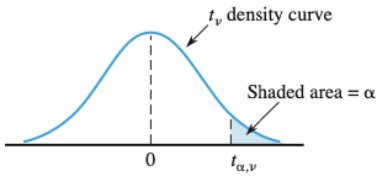
(c) Using $\alpha = 0.05$, what is your conclusion from the hypothesis test?

2. What would be the Type-II error in this case?

3. State the p-value decision rule for hypothesis testing. Use this approach to test the null hypothesis H_0, is the conclusion same as before?

4. Construct a 95% confidence interval for the mean number of exercises solved by the students. Based on this interval, can the instructor conclude that students solve more than 18 exercises on average?

Table A.5 Critical Values for *t* Distributions



	α						
ν	**.10**	**.05**	**.025**	**.01**	**.005**	**.001**	**.0005**
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
32	1.309	1.694	2.037	2.449	2.738	3.365	3.622
34	1.307	1.691	2.032	2.441	2.728	3.348	3.601
36	1.306	1.688	2.028	2.434	2.719	3.333	3.582
38	1.304	1.686	2.024	2.429	2.712	3.319	3.566
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	1.299	1.676	2.009	2.403	2.678	3.262	3.496
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Figure 3: t table