

Assignment_1

Author: Rachakonda Hrithik Sagar

Roll_No: 2023900021

Email_id: hrithik.rachakonda@research.iiit.ac.in

Helper Function

```
set.seed(0)

run_trials <- function(dist, n, trials){
  means <- numeric(trials)
  sds <- numeric(trials)

  for (t in 1:trials) {
    x <- dist(n)
    means[t] <- mean(x)
    sds[t] <- sd(x)
  }
  list(means = means, sds = sds)
}
```

Question 1

Observation and Inference for Question 1

1. The sampling distribution of the sample mean is centered very close to the population mean of 100 across all sample sizes, confirming that the sample mean acts as an unbiased estimator of the population mean.
2. For smaller sample sizes, the distribution of sample means shows a wider spread, reflecting a larger standard error of the mean due to higher sampling variability.
3. As the sample size increases, the sampling distribution of the mean becomes increasingly narrow and sharply peaked around 100. This occurs because the standard error decreases at a rate proportional to $1/\sqrt{n}$, leading to more precise estimates of the population mean.
4. The sampling distributions of the sample standard deviation also become more tightly clustered around the true population standard deviation of 15 as sample size grows, indicating increased stability and reduced variability in the estimation of dispersion.
5. For large sample sizes, the sampling distribution of the mean closely resembles a normal distribution with minimal spread, demonstrating the effect of the Central Limit Theorem, where larger samples yield more reliable and consistent estimates of population parameters.

```

mu <- 100
sigma <- 15
trials <- 1000
ns <- c(10, 50, 100, 500, 1500)

rnorm_pop <- function(n) rnorm(n, mu, sigma)

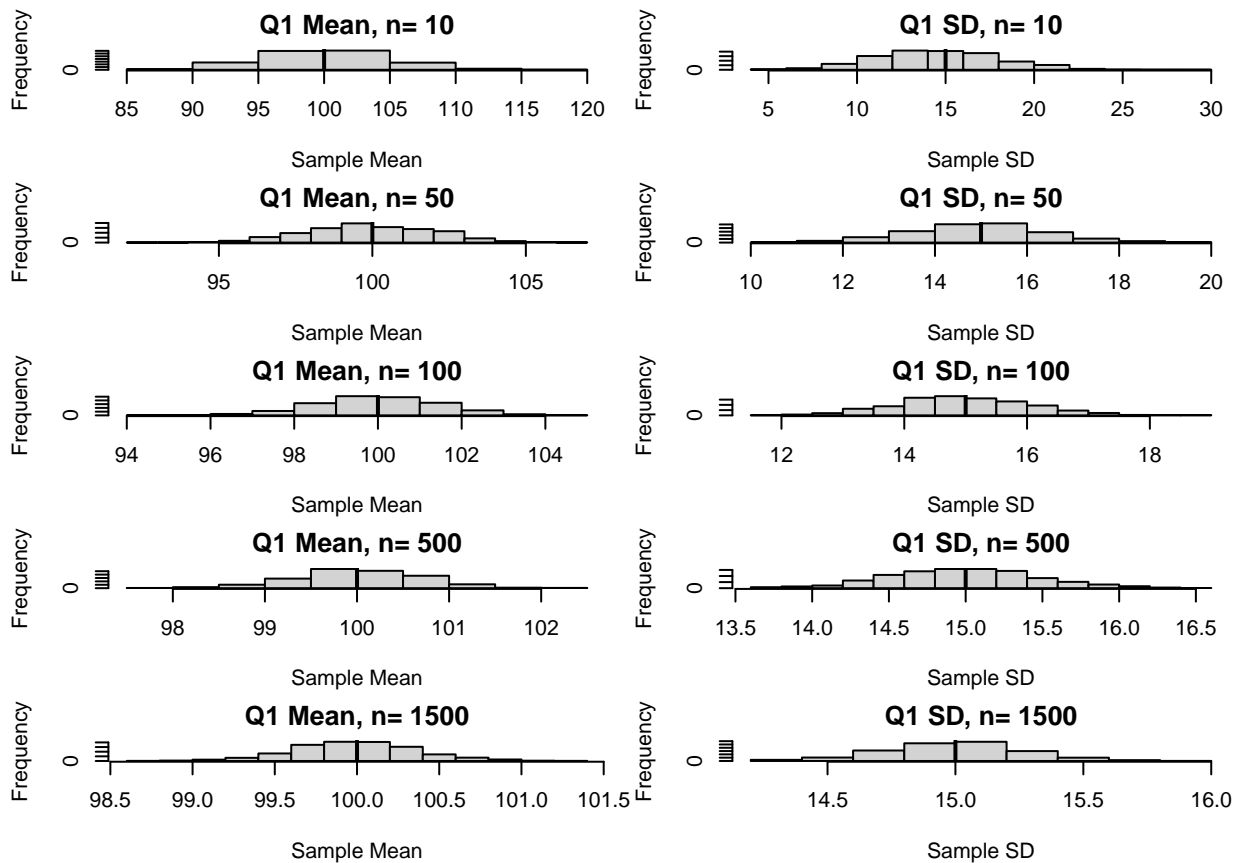
par(mfrow = c(length(ns), 2), mar = c(4,4,2,1))

for (n in ns) {
  res <- run_trials(rnorm_pop, n, trials)

  hist(res$means, main=paste("Q1 Mean, n=", n), xlab="Sample Mean")
  abline(v = mu, lwd = 2)

  hist(res$sds, main=paste("Q1 SD, n=", n), xlab="Sample SD")
  abline(v = sigma, lwd = 2)
}

```



Question 2

Observation and Inference for Question 2

1. For smaller sample sizes, the sampling distribution of the sample mean is noticeably wider and exhibits visible skewness when compared to Question 1. This reflects the inherent asymmetry of the Beta(2,5) population distribution.
2. With an increase in sample size, the sampling distribution of the mean gradually becomes more symmetric and bell-shaped, while also concentrating more closely around the true population mean. This indicates a reduction in sampling variability as more observations are included.
3. The sampling distributions of the sample standard deviation also show substantial variability for small sample sizes, but as n increases, these distributions become more tightly clustered around the population standard deviation, indicating improved stability in dispersion estimates.
4. Unlike Question 1, where the population itself is normally distributed and the sampling distribution of the mean appears approximately normal even for small samples, the skewed nature of the Beta(2,5) population delays this normal approximation.
5. Consequently, a larger sample size is required in this case for the sampling distribution of the mean to closely follow a normal distribution, illustrating that the Central Limit Theorem applies more slowly for skewed populations.

In Question 1, since the population is normally distributed, the sampling distribution of the mean appears normal even for small sample sizes, whereas in Question 2 the population is skewed (Beta(2,5)), causing the sampling distribution of the mean to be less normal for small n and requiring larger sample sizes for it to become approximately normal.

```
a <- 2
b <- 5
trials <- 1000
ns <- c(10, 50, 100, 500, 1500)

beta_mean <- a / (a + b)
beta_sd <- sqrt((a * b) / ((a + b)^2 * (a + b + 1)))

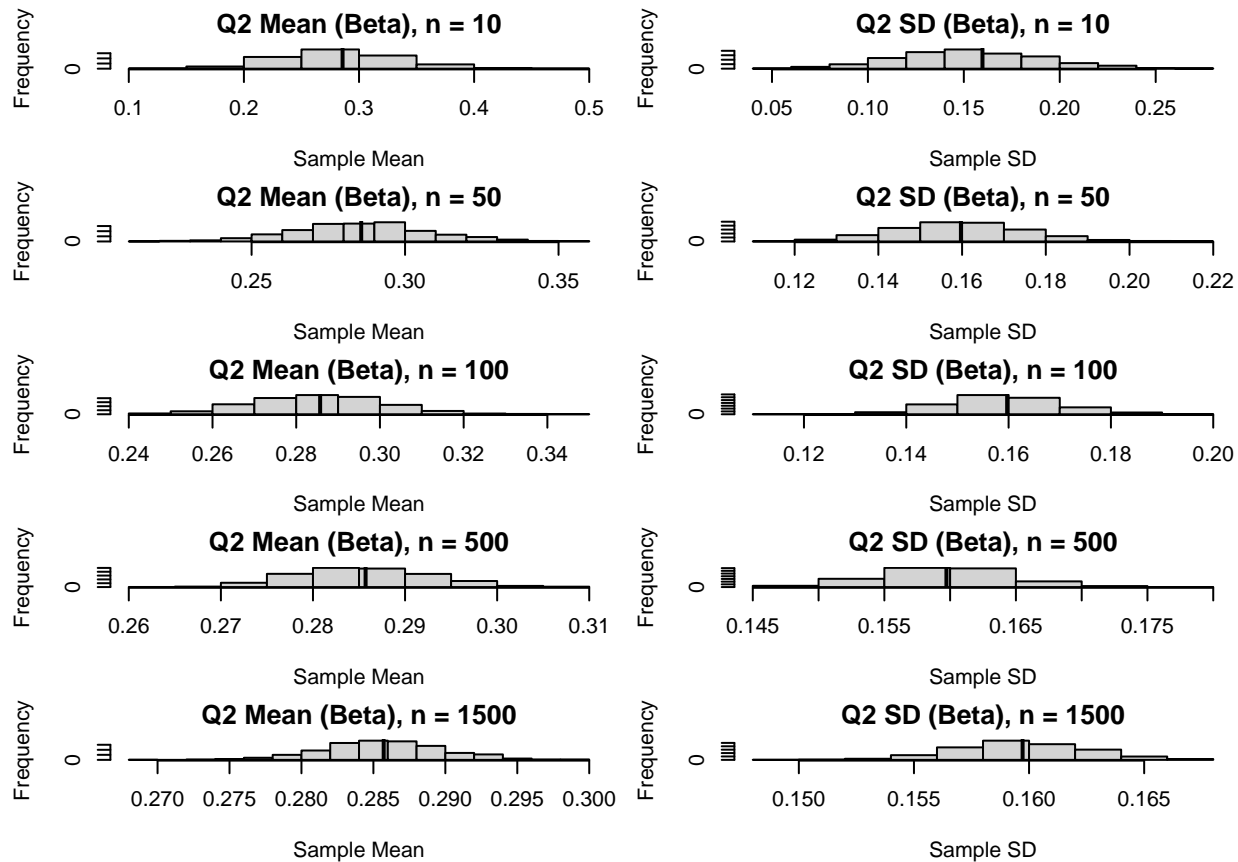
rbeta_pop <- function(n) rbeta(n, a, b)

par(mfrow = c(length(ns), 2), mar = c(4,4,2,1))

for (n in ns) {
  res <- run_trials(rbeta_pop, n, trials)

  hist(res$means,
       main = paste("Q2 Mean (Beta), n =", n),
       xlab = "Sample Mean")
  abline(v = beta_mean, lwd = 2)

  hist(res$sds,
       main = paste("Q2 SD (Beta), n =", n),
       xlab = "Sample SD")
  abline(v = beta_sd, lwd = 2)
}
```



Question 3

Observations and Interpretation for Question 3

Q3(d) For all three groups, the 95% confidence intervals become progressively narrower as the sample size increases from 30 to 70 and then to 100. This reduction in interval width occurs because the standard error of the mean decreases with increasing sample size ($SE = s/\sqrt{n}$), resulting in more precise estimates of the population mean.

Q3(e) A 95% confidence interval provides a range of plausible values for the population mean such that, over repeated sampling, approximately 95% of the constructed intervals would contain the true mean. As the sample size increases, the confidence intervals shrink, indicating improved precision and reduced uncertainty in the mean estimate. Additionally, populations with higher variability exhibit wider confidence intervals, implying that larger sample sizes are required to achieve the same level of estimation reliability.

```
ci95 <- function(x) {
  m <- mean(x)
  s <- sd(x)
  n <- length(x)
  err <- qt(0.975, n - 1) * s / sqrt(n)
  c(mean = m, lower = m - err, upper = m + err)
}

groups <- list(
```

```

list(name="N(0,1)", mu=0, sd=1),
list(name="N(2,0.5)", mu=2, sd=0.5),
list(name="N(3,2)", mu=3, sd=2)
)

ns <- c(30, 70, 100)

for (n in ns) {

  means <- numeric(3)
  lower <- numeric(3)
  upper <- numeric(3)

  par(mfrow=c(1,3))

  for (i in 1:3) {
    x <- rnorm(n, groups[[i]]$mu, groups[[i]]$sd)
    ci <- ci95(x)

    means[i] <- ci["mean"]
    lower[i] <- ci["lower"]
    upper[i] <- ci["upper"]

    hist(x,
         main=paste(groups[[i]]$name, "\n n =", n),
         xlab="Values",
         col="grey")
    abline(v=means[i], lwd=2)
  }

  par(mfrow=c(1,1))

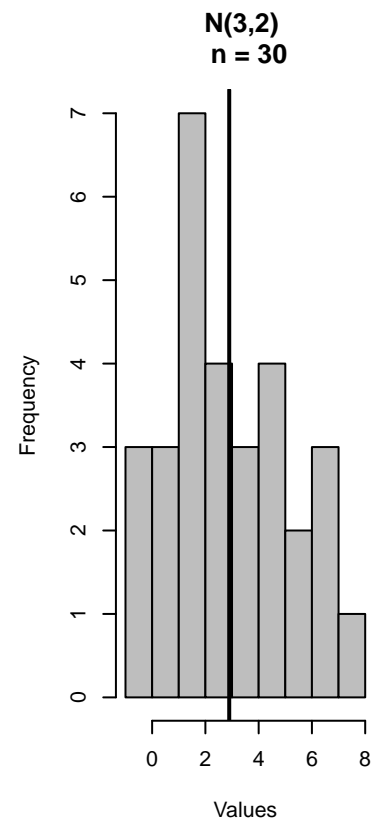
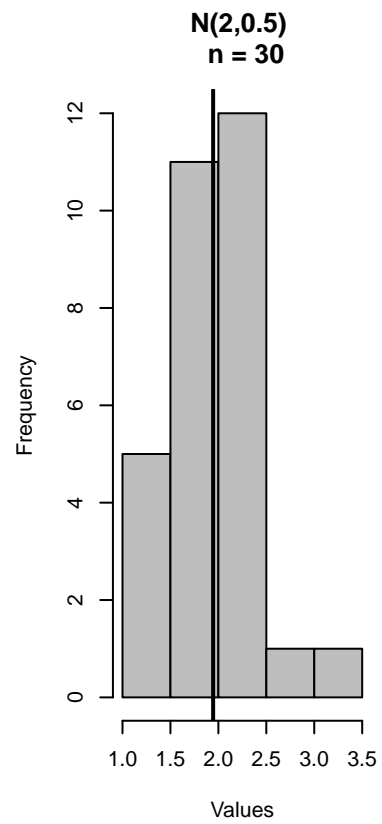
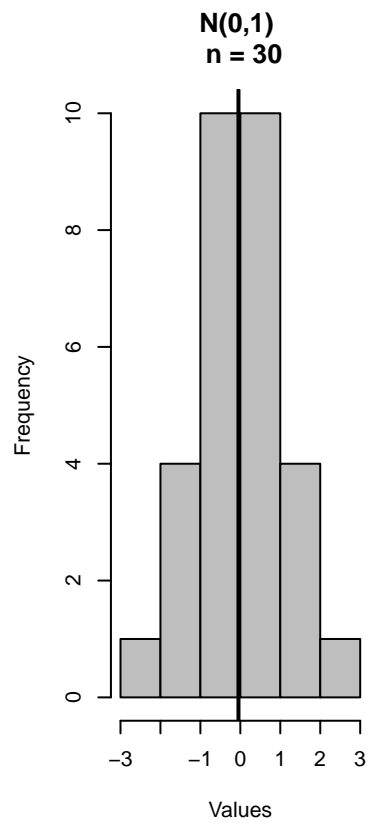
  plot(1:3, means,
       ylim=range(lower, upper),
       xaxt="n",
       xlab="Group",
       ylab="Sample Mean",
       main=paste("95% CI for Means (n =", n, ")"),
       pch=19)

  axis(1, at=1:3, labels=sapply(groups, `[`, "name"))

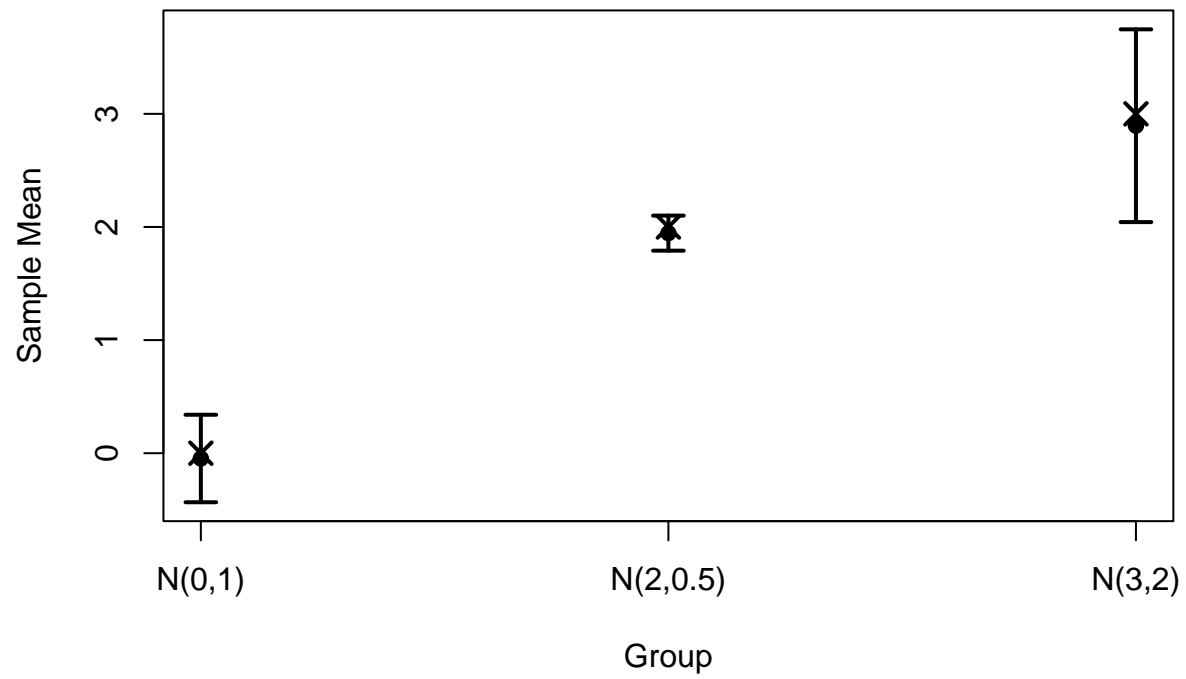
  arrows(1:3, lower, 1:3, upper,
        angle=90, code=3, length=0.08, lwd=2)

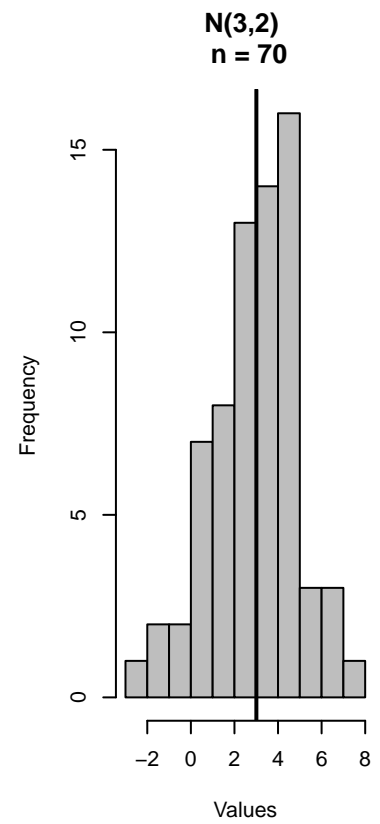
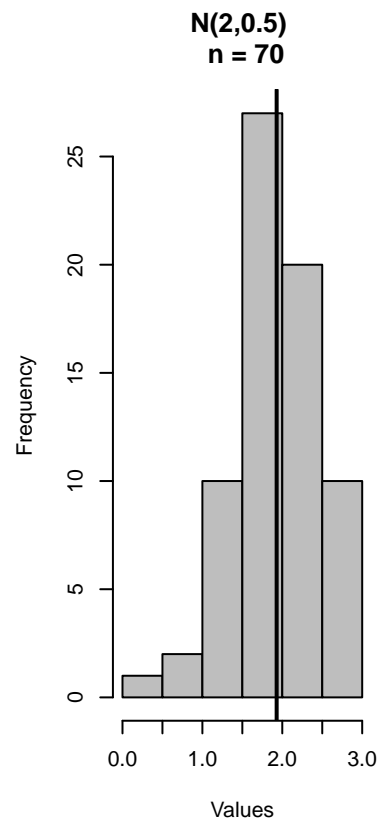
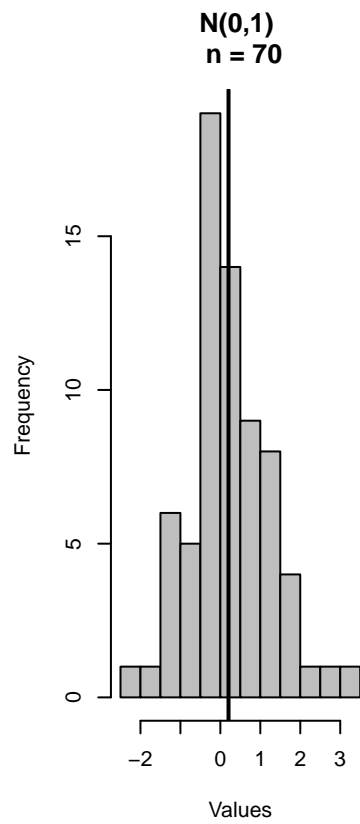
  points(1:3, sapply(groups, `[`, "mu"),
        pch=4, cex=1.5, lwd=2)
}

```

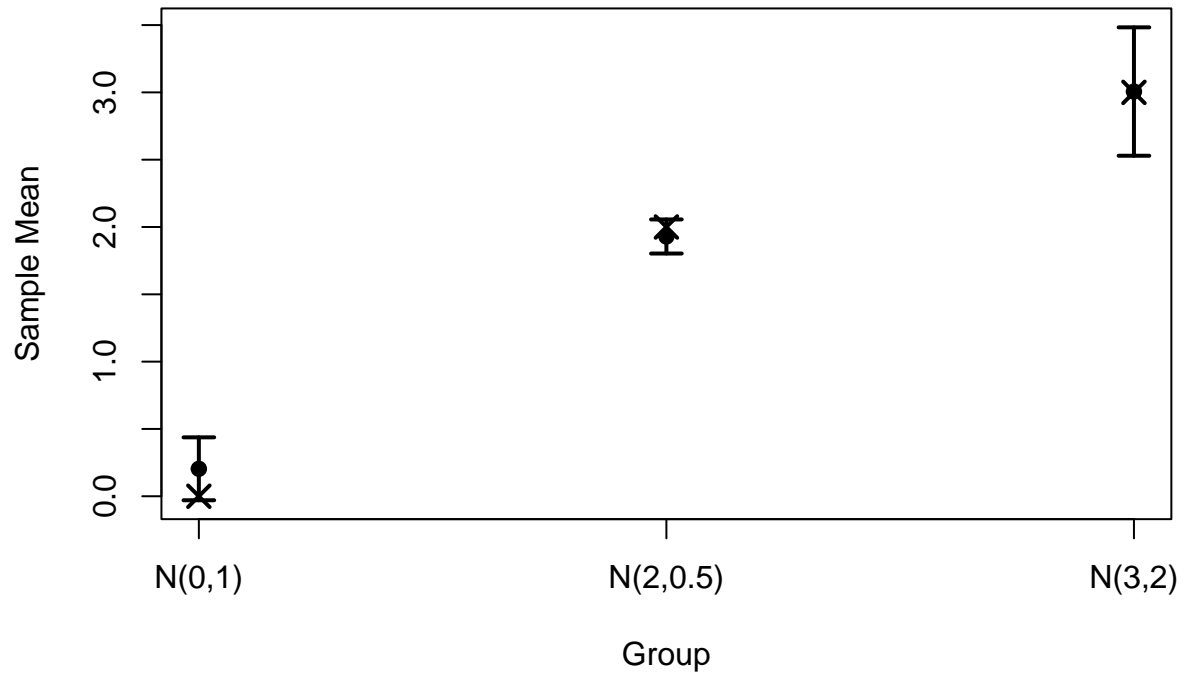


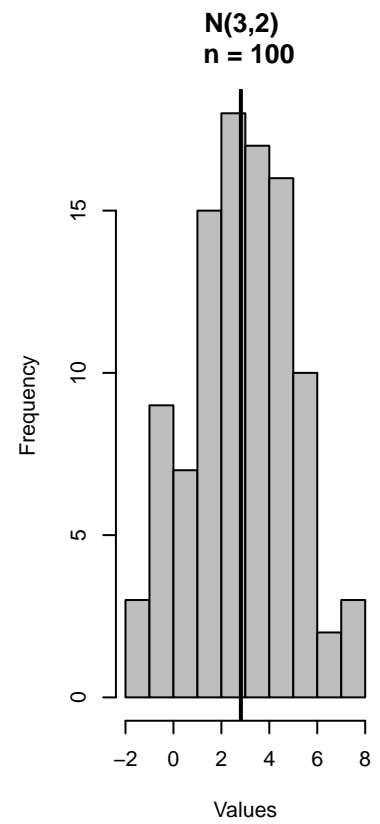
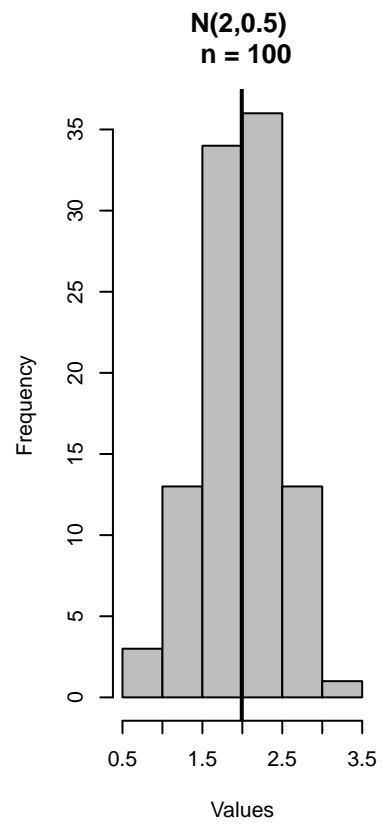
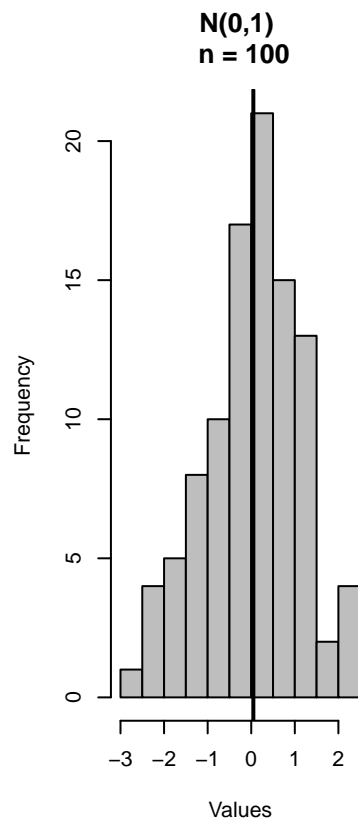
95% CI for Means (n = 30)





95% CI for Means (n = 70)





95% CI for Means (n = 100)

