CLT Simulations

```r
library(tidyverse)
```

Normally Distributed Random Variables

Generate 100 normal random variables from N(0,1) to observe the underlying distribution

```r
# Define distribution parameters
mu = 0
sigma2 = 1
N = 100

# Random sample data frame
set.seed(4)
norm.df = tibble(
  sample = rnorm(N, mu, sigma2),
  id     = "Original Normal Distribution"
)

# Visualize
norm.df %>%
  ggplot(aes(x = sample, y = ..density.., fill = id)) +
  geom_histogram(binwidth = 0.3, colour = "black", alpha = 0.1) +
  geom_density(alpha = 0.4) +
  labs(
    x = sprintf("Xi ~ N(%i, %i)", mu, sigma2),
    y = "Density",
    title = "Histogram and Density Curve for the Original Normal Data") +
  scale_fill_viridis_d("") +
  theme_classic() +
```

```
theme(legend.position = "none")
```

Histogram and Density Curve for the Original Normal Data
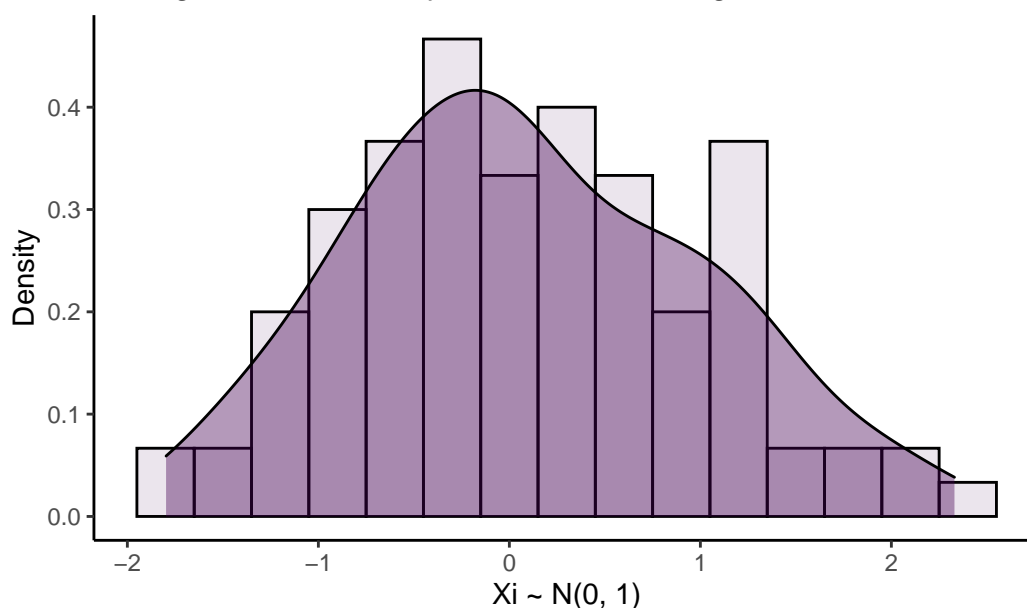
What if we can only observe/sample 10 observations from this population? What if we repeatedly took samples of size 10 from the original population? How would the means of each smaller sample be distributed, compare to the whole sample?

Generate and plot the means of 100 random samples of size 10 (with repalcement) from the original Normal distribution. (This will form the sampling distribution of sample means.)

```
# define parameters
n.sample = 100
n = 10

# Sample indices
sample.indices = list()
for (i in 1:n.sample) {
  sample.indices[[i]] <- sample(1:N, n, replace = TRUE)
}

# Take the mean of each sample and attach to original data frame
norm.df = bind_rows(norm.df,
                    tibble(
```
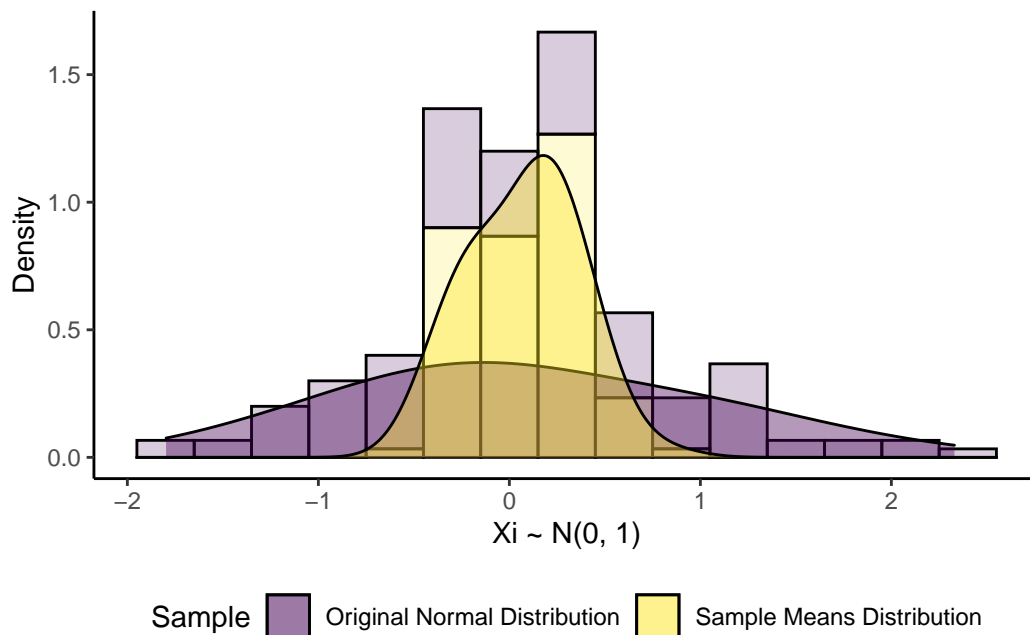
```
                      sample = map_dbl(.x = sample.indices, ~ mean(norm.df$sample[.x])),
                      id = "Sample Means Distribution")
                  )

# Visualize
norm.df %>%
  ggplot(aes(x = sample, y = ..density.., fill = id)) +
  geom_histogram(binwidth = 0.3, colour = "black", alpha = 0.2) +
  geom_density(alpha = 0.4, adjust = 1.6) +
  labs(
    x = sprintf("Xi ~ N(%i, %i)", mu, sigma2),
    y = "Density") +
  scale_fill_viridis_d("Sample") +
  theme_classic() +
  theme(legend.position = "bottom")
```


Non-normally Distributed Random Variables

Poisson Distribution

Generate 100 random variables from Pois(2) distribution.

```
# define parameters
N = 100
lambda = 2

# Random sample data frame
set.seed(2)
pois.df = tibble(
  sample = rpois(N, lambda),
  id     = "Orig Poisson Dist"
)

# Visualize
pois.df %>%
  ggplot(aes(x = sample, y = ..density.., fill = id)) +
  geom_histogram(bins = 30, colour = "black", alpha = 0.1) +
  geom_density(alpha = 0.4) +
  labs(
    x = sprintf("Xi ~ Pois(2)"),
    y = "Density",
    title = "Histogram and Density Curve for the Original Poisson Data") +
  scale_fill_viridis_d("") +
  theme_classic() +
  theme(legend.position = "none")
```
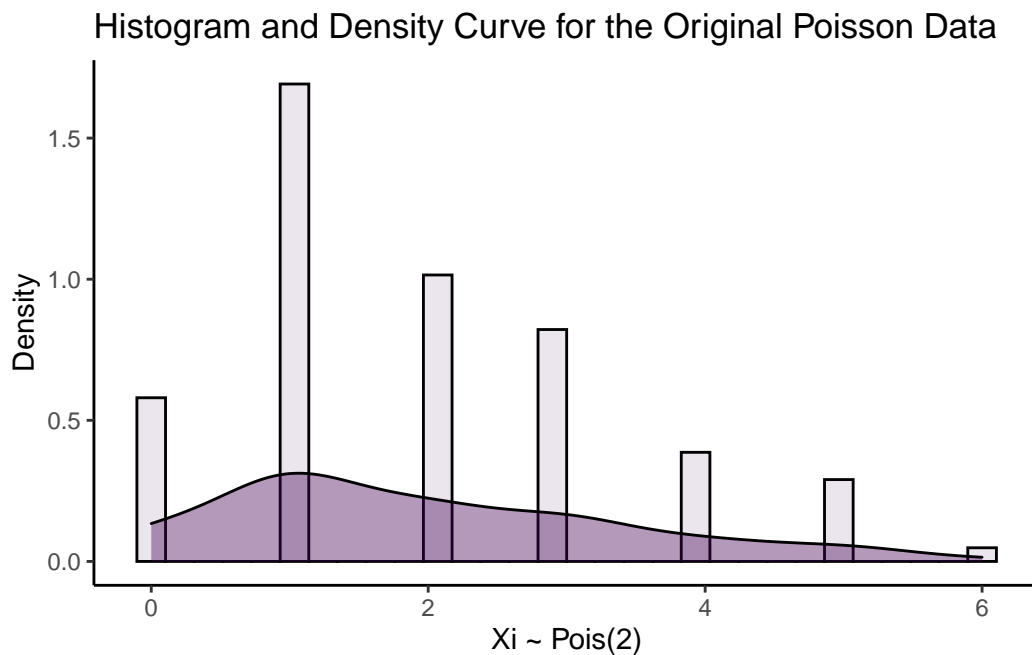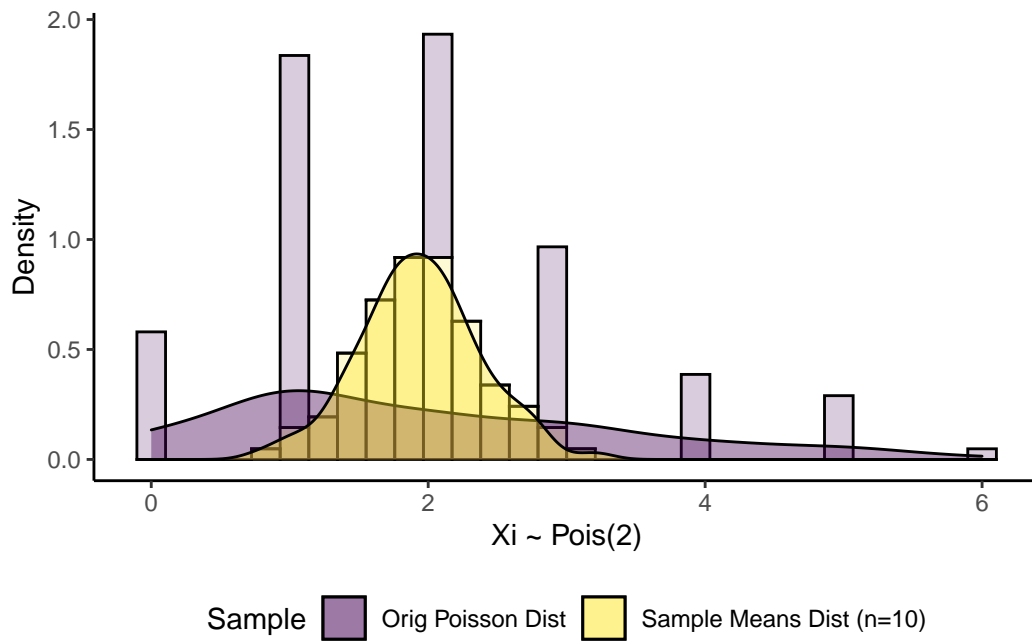
Histogram and Density Curve for the Original Poisson Data

Generate and plot the means of 100 random samples of size 10 (with repalcement) from the
original Poisson distribution. (This will form the sampling distribution of sample means.)

```
# define n's
n.sample = 100
n = 10

# Sample indices
sample.indices = list()
for (i in 1:n.sample) {
  sample.indices[[i]] <- sample(1:N, n, replace = TRUE)
}

# Take the mean of each sample and attach to original data frame
pois.df = bind_rows(pois.df,
                    tibble(
                    sample = map_dbl(.x = sample.indices, ~ mean(pois.df$sample[.x])),
                      id = "Sample Means Dist (n=10)")
                    )

# Visualize
pois.df %>%
  ggplot(aes(x = sample, y = ..density.., fill = id)) +
  geom_histogram(bins=30, colour = "black", alpha = 0.2) +
  geom_density(alpha = 0.4) +
  labs(
    x = sprintf("Xi ~ Pois(2)"),
    y = "Density") +
  scale_fill_viridis_d("Sample") +
  theme_classic() +
  theme(legend.position = "bottom")
```

Generate and plot the means of 100 random samples of size 30 (with repalcement) from the
original Poisson distribution. (This will form the sampling distribution of sample means.)

```r
# define n's
n.sample = 100
n = 30

# Sample indices
sample.indices = list()
for (i in 1:n.sample) {

  sample.indices[[i]] <- sample(1:N, n, replace = TRUE)

}

# Take the mean of each sample and attach to original data frame
pois.df = bind_rows(pois.df,
                    tibble(
                    sample = map_dbl(.x = sample.indices, ~ mean(pois.df$sample[.x])),
                      id = "Sample Means Dist (n=30)")
                    )

poisson_cols <- c(
```

```
    "Orig Poisson Dist" = "#440154FF",
    "Sample Means Dist (n=10)" = "#FDE725FF",
    "Sample Means Dist (n=30)" = "#21908CFF"
)

# Visualize
pois.df %>%
  ggplot(aes(x = sample, y = ..density.., fill = id)) +
  geom_histogram(bins=30, colour = "black", alpha = 0.2) +
  geom_density(alpha = 0.4, adjust = 1.6) +
  labs(
    x = sprintf("Xi ~ Pois(2)"),
    y = "Density") +
  scale_fill_manual(values = poisson_cols, name = "Sample") +
  theme_classic() +
  theme(legend.position = "bottom",
        legend.text = element_text(size = 7))
```
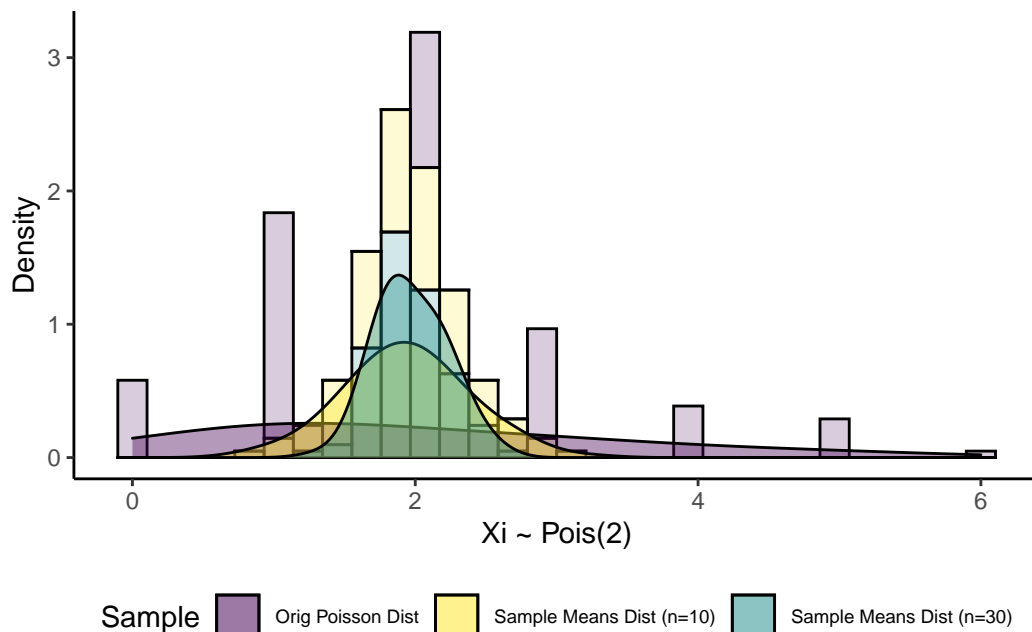

Binomial Distribution

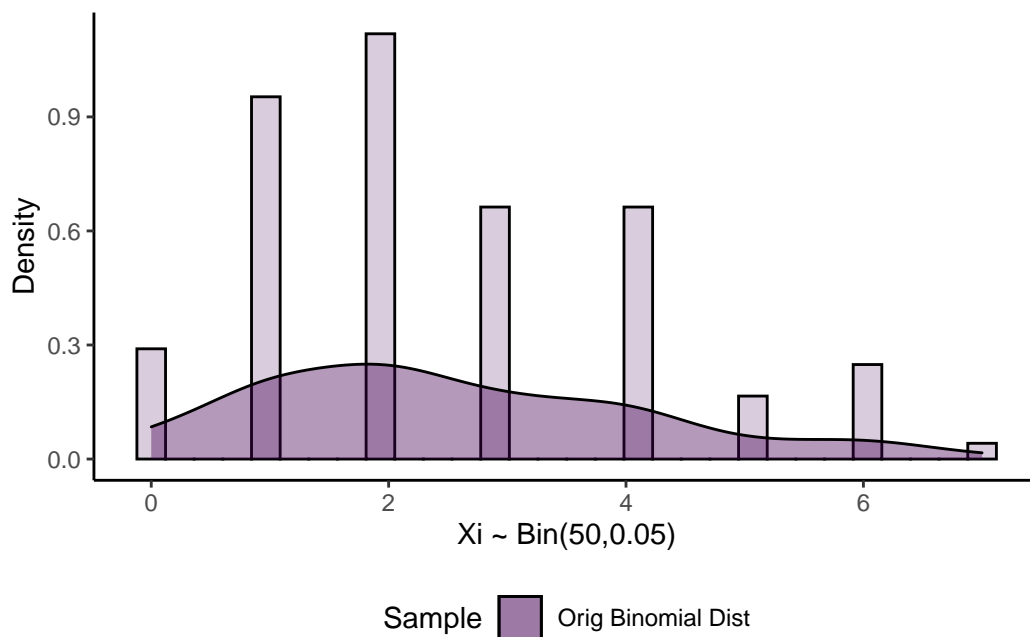Generate 100 random variables from Bin(50,0.30) distribution.

```r
# define parameters
N = 100
nbin = 50
p = 0.05

# Random sample data frame
set.seed(2)
bin.df = tibble(
  sample = rbinom(N, nbin, p),
  id = "Orig Binomial Dist"
)

bin.df %>%
  ggplot(aes(x = sample, y = ..density.., fill = id)) +
  geom_histogram(bins=30, colour = "black", alpha = 0.2) +
  geom_density(alpha = 0.4) +
  labs(
    x = sprintf("Xi ~ Bin(50,0.05)"),
    y = "Density") +
  scale_fill_viridis_d("Sample") +
  theme_classic() +
  theme(legend.position = "bottom")
```

Generate and plot the means of 100 random samples of size 10 (with repalcement) from the original Binomial distribution. (This will form the sampling distribution of sample means.)
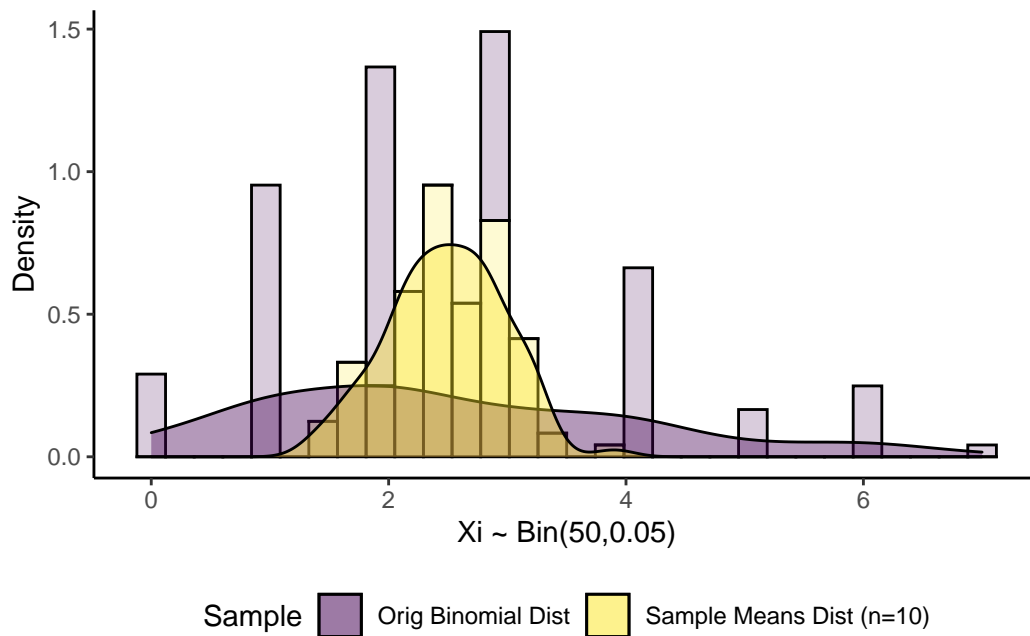
```
# define n's
n.sample = 100
n = 10

# Sample indices
sample.indices = list()
for (i in 1:n.sample) {

  sample.indices[[i]] <- sample(1:N, n, replace = TRUE)

}

# Take the mean of each sample and attach to original data frame
bin.df = bind_rows(bin.df,
                   tibble(
                   sample = map_dbl(.x = sample.indices, ~ mean(bin.df$sample[.x])),
                     id = "Sample Means Dist (n=10)")
                 )

# Visualize
bin.df %>%
  ggplot(aes(x = sample, y = ..density.., fill = id)) +
  geom_histogram(bins=30, colour = "black", alpha = 0.2) +
  geom_density(alpha = 0.4) +
  labs(
    x = sprintf("Xi ~ Bin(50,0.05)"),
    y = "Density") +
  scale_fill_viridis_d("Sample") +
  theme_classic() +
  theme(legend.position = "bottom")
```

Generate and plot the means of 100 random samples of size 30 (with repalcement) from the original Binomial distribution. (This will form the sampling distribution of sample means.)

```r
# define n's
n.sample = 100
n = 30

# Sample indices
sample.indices = list()
for (i in 1:n.sample) {

  sample.indices[[i]] <- sample(1:N, n, replace = TRUE)

}

# Take the mean of each sample and attach to original data frame
bin.df = bind_rows(bin.df,
                   tibble(
                   sample = map_dbl(.x = sample.indices, ~ mean(bin.df$sample[.x])),
                     id = "Sample Means Dist (n=30)")
                   )
```

```r
binomial_cols <- c(
  "Orig Binomial Dist" = "#440154FF",
  "Sample Means Dist (n=10)" = "#FDE725FF",
  "Sample Means Dist (n=30)" = "#21908CFF"
)


# Visualize
bin.df %>%
  ggplot(aes(x = sample, y = ..density.., fill = id)) +
  geom_histogram(bins=30, colour = "black", alpha = 0.2) +
  geom_density(alpha = 0.4, adjust = 1.2) +
  labs(
    x = sprintf("Xi ~ Bin(50,0.05)"),
    y = "Density") +
  scale_fill_manual(values = binomial_cols, name = "Sample") +
  theme_classic() +
  theme(legend.position = "bottom",
        legend.text = element_text(size = 7))
```