Lab2

Justin Maynard

2024-01-16

```
rm(list = ls())
```

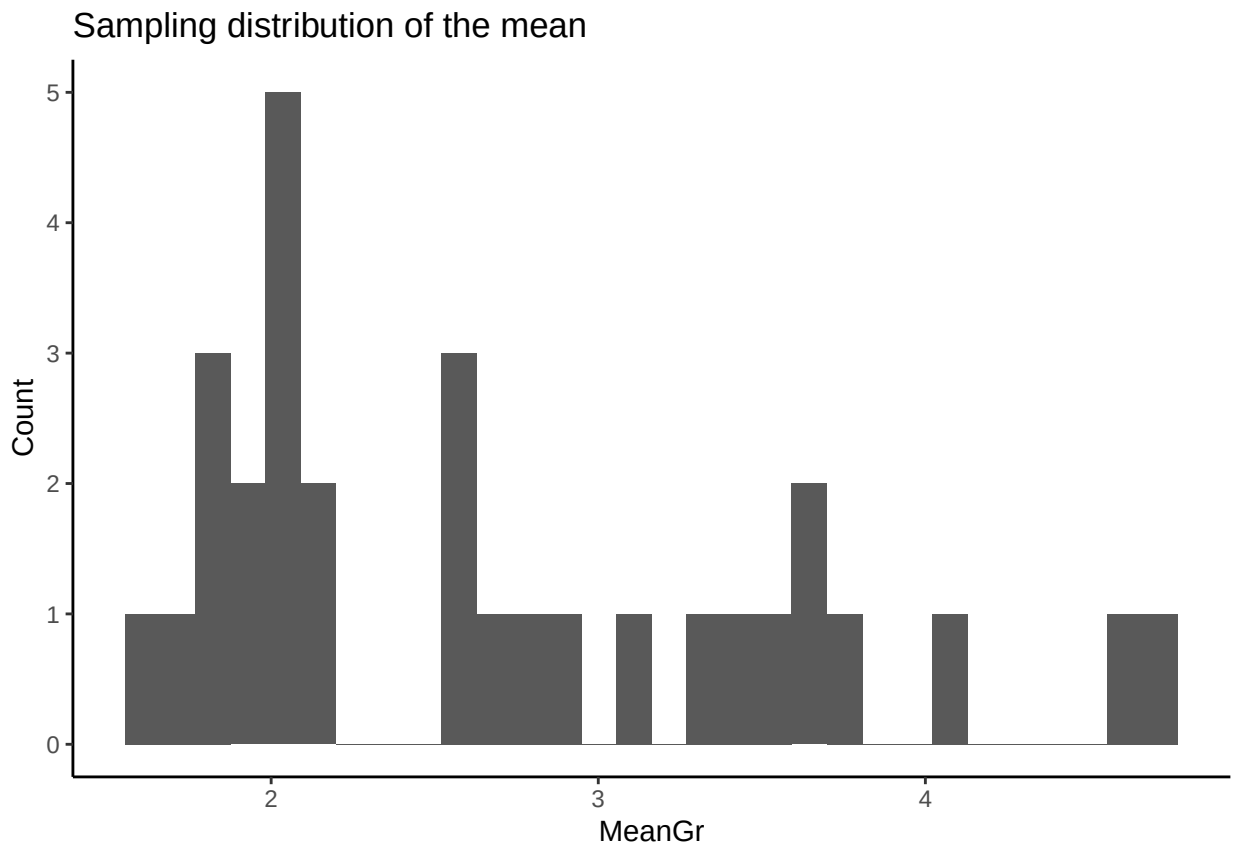
Lab 2

Problem 1 (0.5 points)

Summarize the data for mean growth MeanGr of the plots in afdat. Mean growth is the difference in average DBH for each plot between the first tree census and the second census; thus, there are only data for MeanGr during the second census. Include the following in your analysis: a. Plot a histogram and boxplot of MeanGr. b. Report the mean, median, variance, standard deviation and coefficient of variation of the MeanGr.

```
afdat <- read.csv(here("data/AfrPlots.csv"), header = T)
MeanGr <- (afdat$MeanGr)
nas <- is.na(MeanGr)
afdat_2 <- MeanGr[!nas]


#histogram
MeanGr_histogram <- ggplot(afdat, aes(x = afdat$MeanGr)) +
  geom_histogram() +
  theme_classic() +
  ggtitle("Sampling distribution of the mean") +
  ylab("Count") +
  xlab("MeanGr")
MeanGr_histogram
```

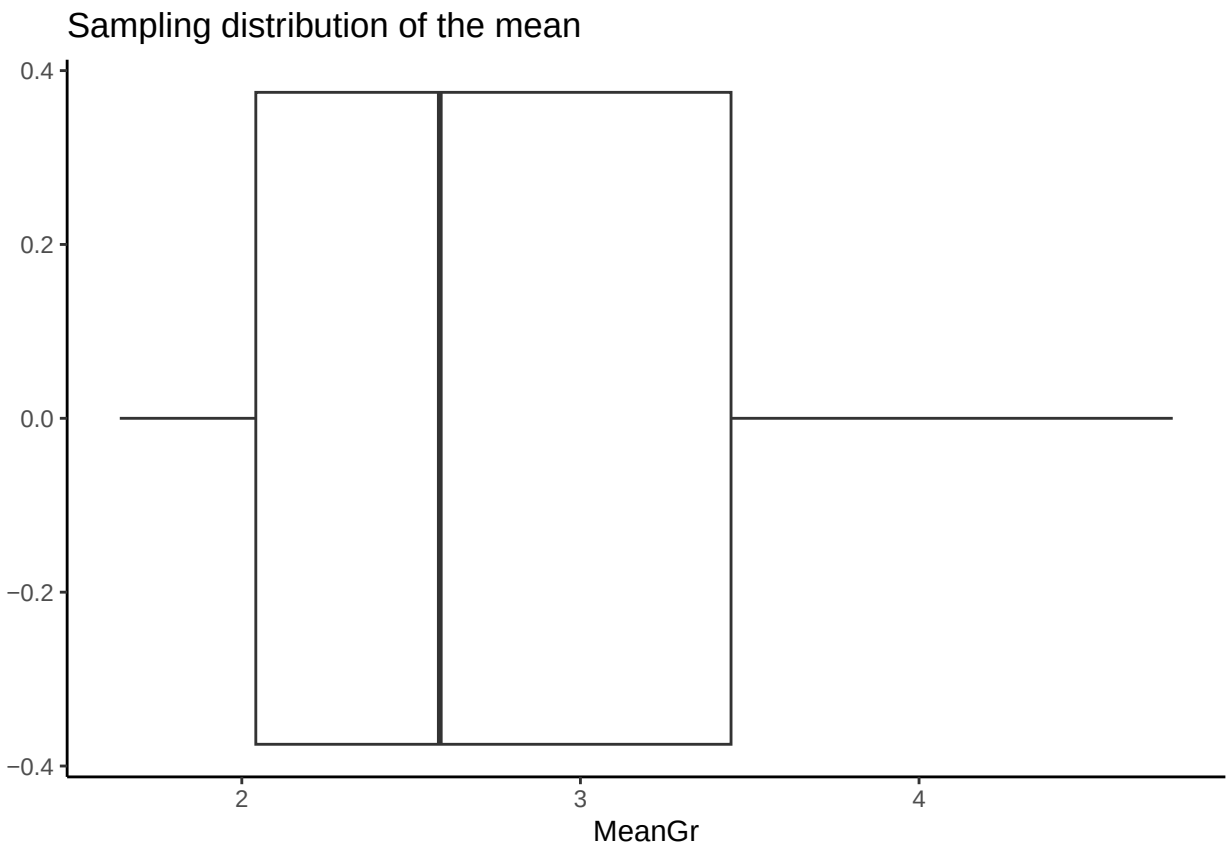
```
## Warning: Use of `afdat$MeanGr` is discouraged.
## i Use `MeanGr` instead.

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 30 rows containing non-finite values (`stat_bin()`).
```




```
MeanGr_boxplot <- ggplot(afdat, aes(x = afdat$MeanGr)) +
  geom_boxplot() +
  theme_classic() +
  xlab("MeanGr") +
  ggtitle("Sampling distribution of the mean")
MeanGr_boxplot
```

```
## Warning: Use of ‘afdat$MeanGr‘ is discouraged.
## i Use ‘MeanGr‘ instead.
```

```
## Warning: Removed 30 rows containing non-finite values (‘stat_boxplot()‘).
```




```
#mean
mean(afdat$MeanGr, na.rm = TRUE)
```

```
## [1] 2.705763
```

```
#variance
var(afdat$MeanGr, na.rm = TRUE)
```

```
## [1] 0.8000796
```

```
#standard deviation
sd(afdat$MeanGr, na.rm = TRUE)
```

```
## [1] 0.8944717
```

```
#coefficient of variation
sd(afdat$MeanGr, na.rm = TRUE) / mean(afdat$MeanGr, na.rm = TRUE)
```

```
## [1] 0.3305802
```

Problem 2 (0.5 points)

Say you flip a fair coin 20 times. What is the probability of obtaining 5 or fewer heads or more than 5 heads? Show how you would answer this question using the dbinom function and calculating it with your binomial equation from above.

```r
n <- 10
p <- 1/2
x <- 0:5

#Calculating with dbinom 0:5
pr <- dbinom(x, size = n, prob = p)
sum(pr[1:6])
```

```
## [1] 0.6230469
```

```r
#Calcuating with dbinom 0:5
sum(dbinom(0:5, size = n, prob = p))
```

```
## [1] 0.6230469
```

```r
#Calculating with dbiom 5:10
sum(dbinom(5:10, size = n, prob = p))
```

```
## [1] 0.6230469
```

```r
#Calculating with bionomial equation
( choose(n,0) * (p^0) * ((1-p)^(n-0))  ) +
( choose(n,1) * (p^1) * ((1-p)^(n-1))  ) +
( choose(n,2) * (p^2) * ((1-p)^(n-2))  ) +
( choose(n,3) * (p^3) * ((1-p)^(n-3))  ) +
( choose(n,4) * (p^4) * ((1-p)^(n-4))  ) +
( choose(n,5) * (p^5) * ((1-p)^(n-5))  )
```

```
## [1] 0.6230469
```

$$P(X) = \sum_{k=0}^{5} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

Problem 3 (0.5 points)

Suppose there are 20 multiple-choice questions on a Stats quiz. Each question has four possible answers, only one of which is correct. Find the probability of answering 17 or more answers correctly if you attempt to answer them at random.

```r
sum(dbinom(17:20, size = 20, prob = .25))
```

```
## [1] 2.960496e-08
```

Problem 4 (0.5 points)

If there are 4 butterflies feeding at a flower per hour on average, find the probability of having 9 or more butterflies feeding at the flower in a particular hour. Although the possible maximum count of butterflies could be much greater, let's limit the maximum number of butterflies to be 13. Do this by (1) writing out the Poisson formula in R, and (2) by using dpois(). Please also graph the probability distribution.

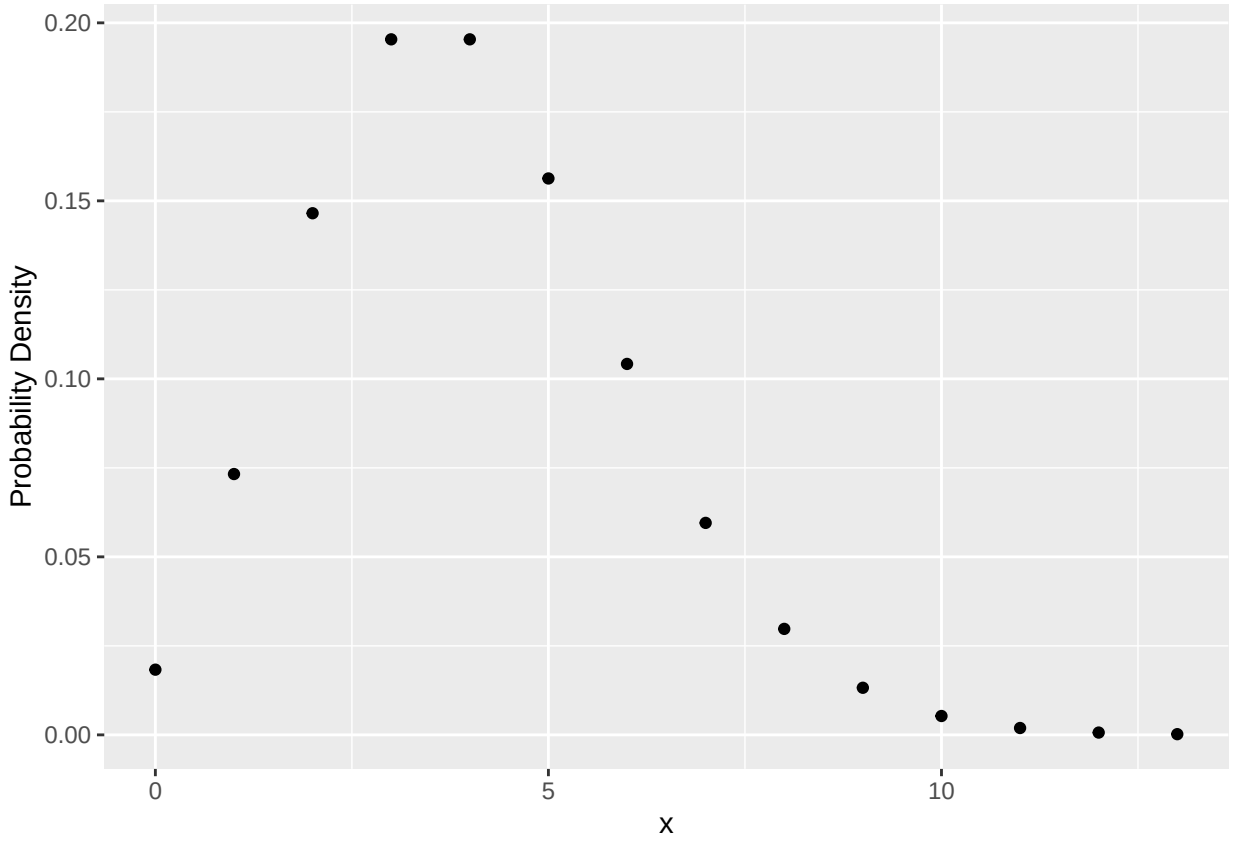
```
#Calculating using written out formula
((4^9) / (factorial(9)))* exp(1)^(-4) +
  ((4^10) / (factorial(10)))* exp(1)^(-4) +
  ((4^11) / (factorial(11)))* exp(1)^(-4) +
  ((4^12) / (factorial(12)))* exp(1)^(-4) +
  ((4^13) / (factorial(13)))* exp(1)^(-4)
```

```
## [1] 0.02128711
```

```
#Calculating using dpois()
sum(dpois(x = 9:13, lambda =4))
```

```
## [1] 0.02128711
```

```
q4_plot <- ggplot(data.frame(x=c(0:13)), aes(x)) +
    geom_point(aes(y=dpois(x, 4)), colour="black") +
  ylab("Probability Density")
q4_plot
```



$$P(X) = \sum_{x=9}^{13} \frac{\lambda^x}{x!} e^{-\lambda}$$

Question 5

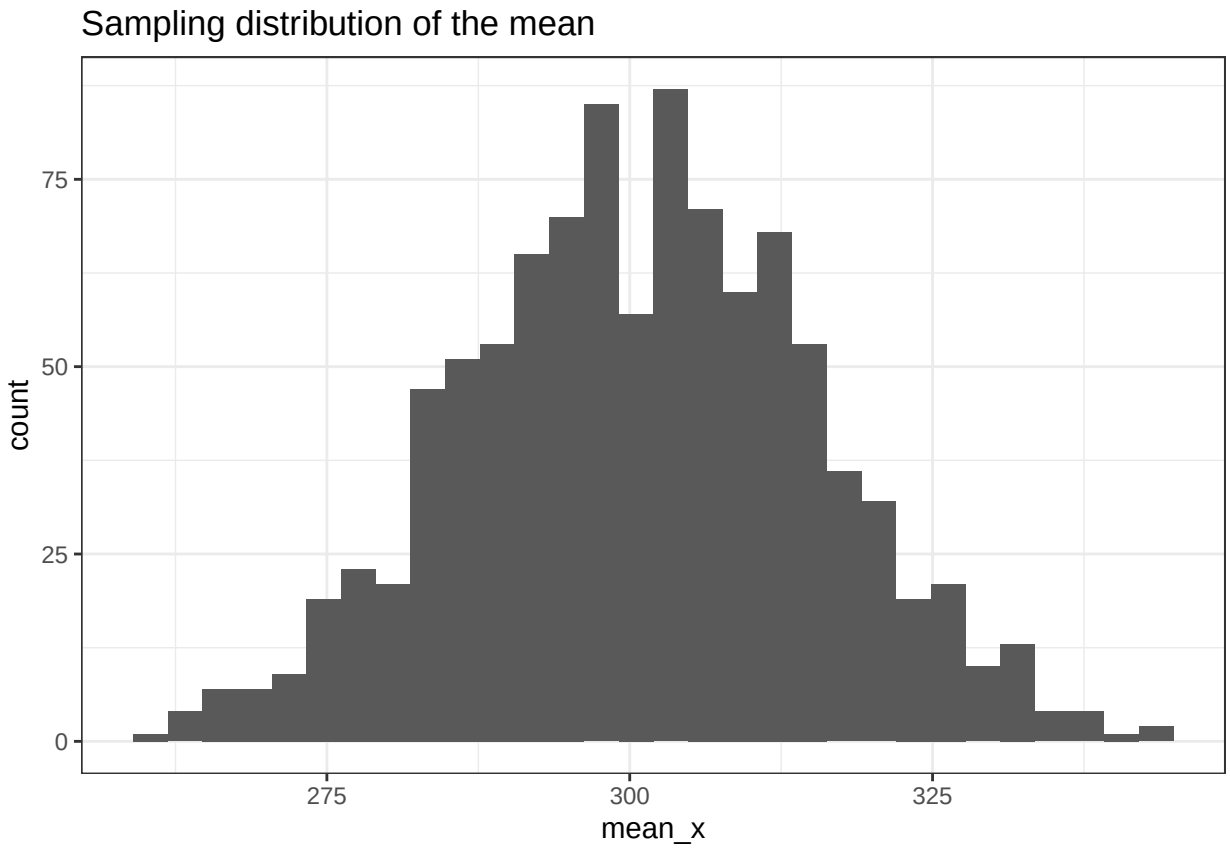
You are going to conduct many experiments similar to the one described above under the 'Sampling Distribution' section.

a. Suppose you are interested in knowing the spatial pattern of a species' distribution (i.e. presence or absence) across an area that is 100km x 100km. You decide to divide the study area into 1km x 1km square grid cells. The 10,000 resulting cells are far too many to realistically sample given the budget that you have to conduct your study, but you have the ability to sample 10% of them. Further, from a prior census, you know that the probability that the species is present is p = 0.30. Conduct 1,000 experiments in which you randomly generate a data point that indicates whether or not the species is present in a given grid cell, and estimate the proportion of cells where the species is present. Store the resulting estimates and plot them as seen in the example above. Comment on the resulting plot. What does it look like? What distribution does it appear to follow?

```
set.seed(123)
n_samples <- 1000
prob <- .30


sample_means <- matrix(NA,nrow=n_samples)

for(i in 1:n_samples){
```

```
  x <- rbinom(n_samples, n_samples, prob)
  sample_means[i] <- mean(x)
  }

x <- data.frame(mean_x = x)


q5a_plot <- ggplot(x,aes(mean_x))+
  geom_histogram()+
  theme_bw()+
  ggtitle("Sampling distribution of the mean")
q5a_plot
```

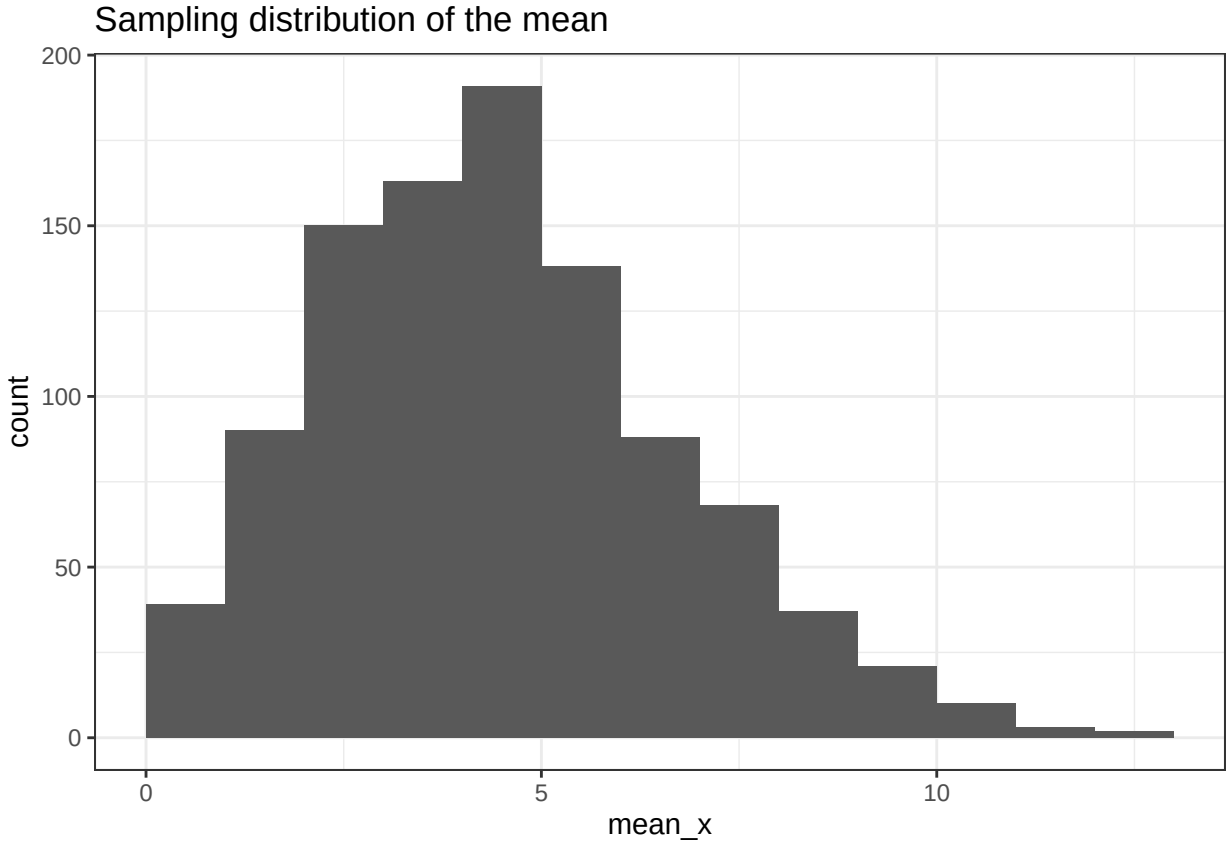
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#This appears to follow the normal distribution
```

b. Now suppose that, across the same area, you are interested in understanding the number of trees in each grid cell. Again, previous research has indicated that the true expected number of trees in each grid cell is 5. Conduct a similar experiment to the one above. Here, you are interested in the average number of trees per grid cell. Store the resulting estimates and plot them as seen in the example above. Again, what does the distribution look like? What distribution does it appear to follow?

```r
n_samples <- 1000
lambda <- 5

sample_means <- matrix(NA,nrow=n_samples)

for(i in 1:n_samples){
  x <- rpois(n = n_samples, lambda = lambda)
  sample_means[i] <- mean(x)
  }

x <- data.frame(mean_x = x)


q5b_plot <- ggplot(x,aes(mean_x))+
  geom_histogram(binwidth = 1, boundary = 0)+
  theme_bw()+
  ggtitle("Sampling distribution of the mean")
q5b_plot
```



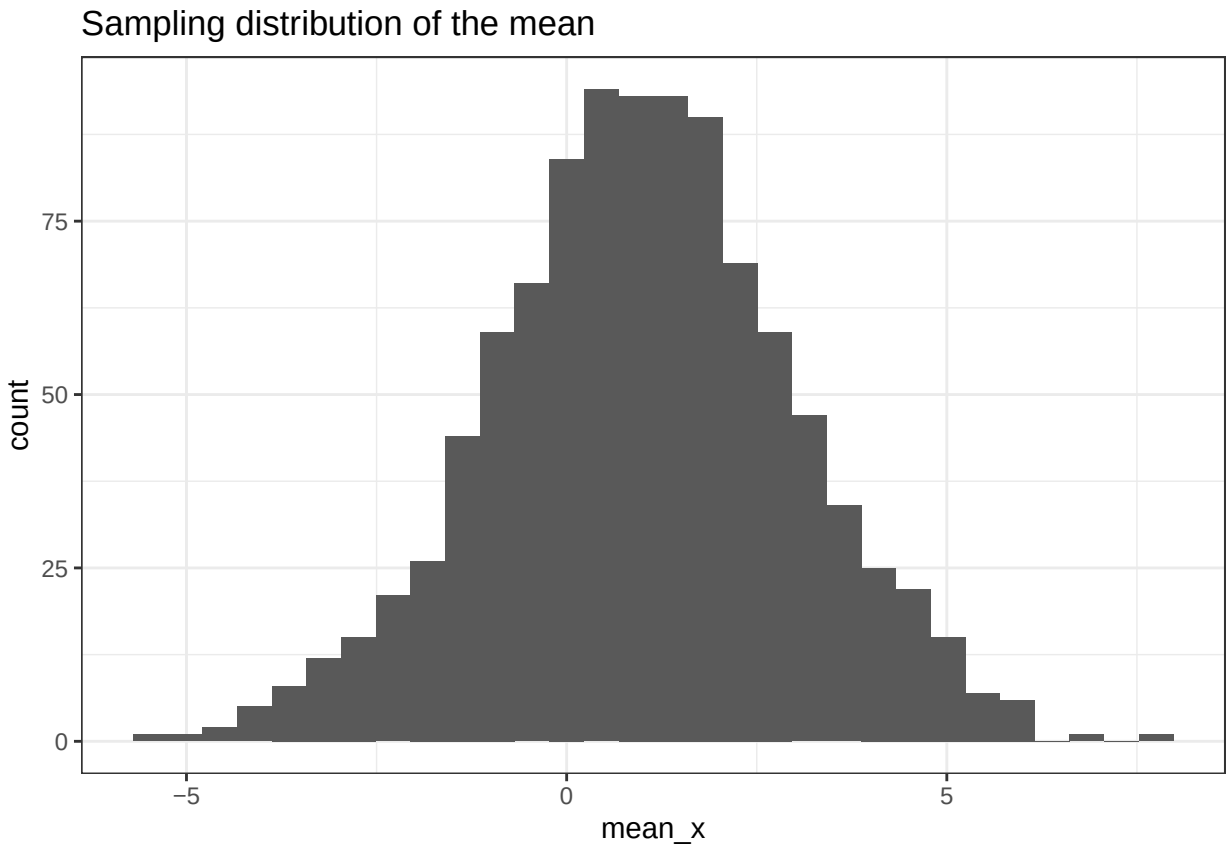

```r
#The distribution looks like a bell curve with a slight tail on the right
#This appears to follow the normal distribution but is skewed right slightly
```

c. Now suppose that you are interested in understanding the diameter at breast height (DBH) in feet of the tree population in this area. Again, suppose you know from prior research that the mean diameter is 1 foot, with a standard deviation of 2 inches. Repeat the procedures of a and b (adapted for this

experiment), store the estimates, and plot them as before. What does the distribution look like? What distribution does it appear to follow?

```
sample_means <- matrix(NA,nrow=n_samples)

for(i in 1:n_samples){
  x <- rnorm(n = 1000, mean = 1, sd = 2)
  sample_means[i] <- mean(x)
  }

x <- data.frame(mean_x = x)


q5c_plot <- ggplot(x,aes(mean_x))+
  geom_histogram()+
  theme_bw()+
  ggtitle("Sampling distribution of the mean")
q5c_plot
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#The distribution looks like a bell curve centered around a mean of 1
#This appears to follow the normal distribution
```