# Practical 5

**Note: I worked with Sophie Paradis**

You are studying a fish phenotypic trait, "T," which you hypothesize is dominant over the alternative phenotype "t." In classical Mendelian genetics, the offspring of two heterozygous parents (Tt) should exhibit the dominant and recessive traits in a 3:1 ratio (three individuals with the dominant phenotype for every one individual with the recessive phenotype).

In a tank containing only heterozygous parents (Tt), you inspect 350 juveniles and observe that 254 display the dominant trait (T) and 96 display the recessive trait (t). You aim to use simulation to test whether there's a statistically significant difference between the observed numbers of dominant and recessive traits (254:96) and what you would expect if the trait T is truly dominant in a 3:1 ratio (approximately 263 dominant: 87 recessive, given the sample size of 350).

In other words, imagine a scenario where you have a large number of jars. Each jar contains an immense quantity of marbles that have an exact 3:1 ratio of black (representing the dominant trait) to white (indicative of the recessive trait) marbles. From each jar, you randomly select a sample of 350 marbles. Under the most typical circumstances, given the 3:1 ratio, you would expect to retrieve approximately 263 black and 87 white marbles from each jar.

What you want to do here is to assess the probability of encountering a deviation from this anticipated outcome — specifically, how plausible it is to draw a sample comprising 254 black and 96 white marbles as was the case in your fish tank? How plausible it is to draw a distribution that diverges more substantially from the expected ratio, such as 200 black and 150 white marbles, from a jar. This evaluation helps determine whether the observed variations are within the realm of normal statistical fluctuations or if they signify an unusual event that defies the established 3:1 genetic dominance principle.

Recall that the steps to carry out this analysis are as follows:

1. Compute a test statistic to describe the observed difference between the expected and observed values. Hint: this was covred in the `pdf`

```
#observed(black = 254, white = 96) - expected(black = 263, white = 87)
abs(254/350 - 263/350)
```

```
## [1] 0.02571429
```

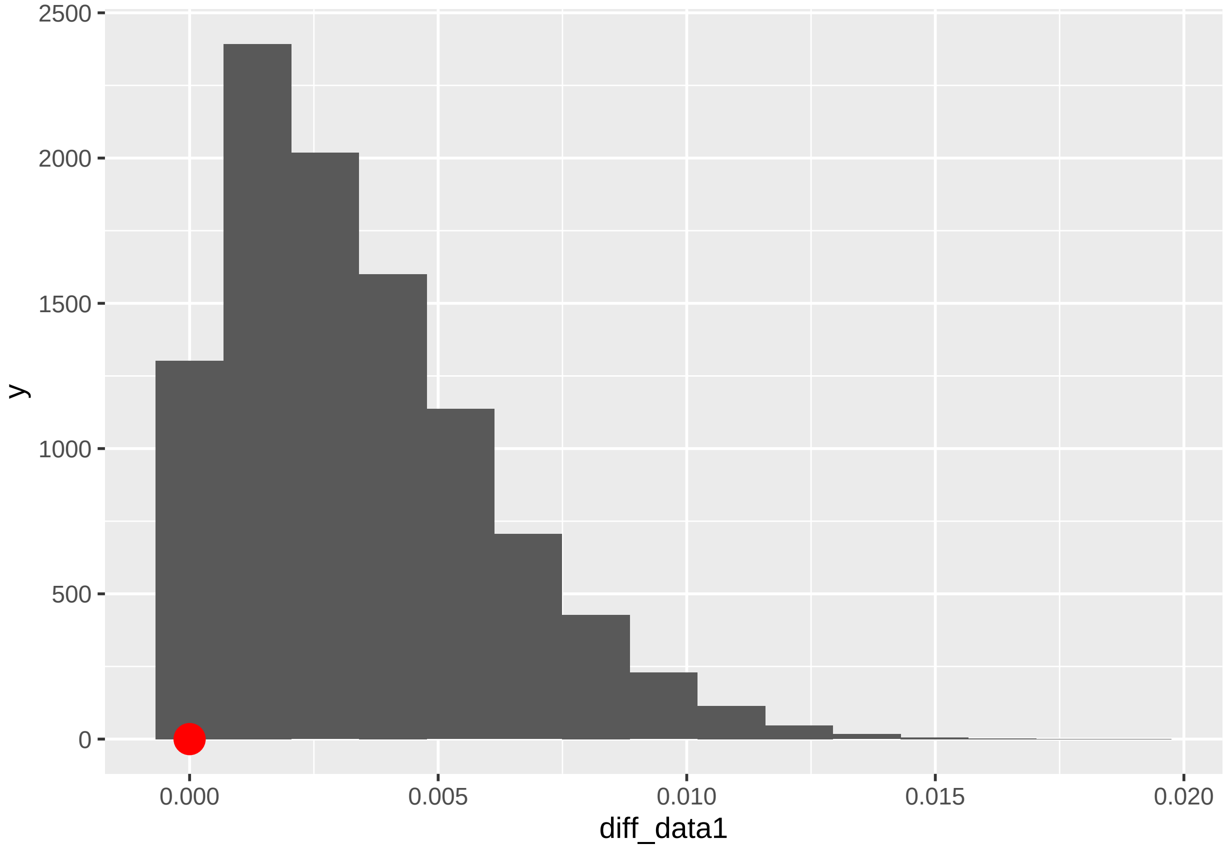
```
#observed(black = 200, white = 150) - expected(black = 263, white = 87)
abs(200/350 - 263/350)
```

```
## [1] 0.18
```

2. Quantify what is considered a normal sampling variation. In other words, use simulation to determine occurrnces resulting from normal statistical fluctuations. This involves simulating many instances of drawing 350 marbles from jars with a 3:1 ratio and seeing, using the test statistic above, the values that expects due to the randomness inherent to sampling alone.

```
library(ggplot2)
library(patchwork)

#computing the absolute difference between reference proportion and observed proportion
computDiff = function(black_prob, white_prob, sample_size, reference_prob){
model_proportions = c(black_prob, white_prob)
data = sample(c("black", "white"), 10000, replace=TRUE, prob=model_proportions)
abs((sum(data=="black")/length(data)) - reference_prob)
}

#create distributions of the test statistics
#reference proportion = 263/350; observed proportion = 263/350
```

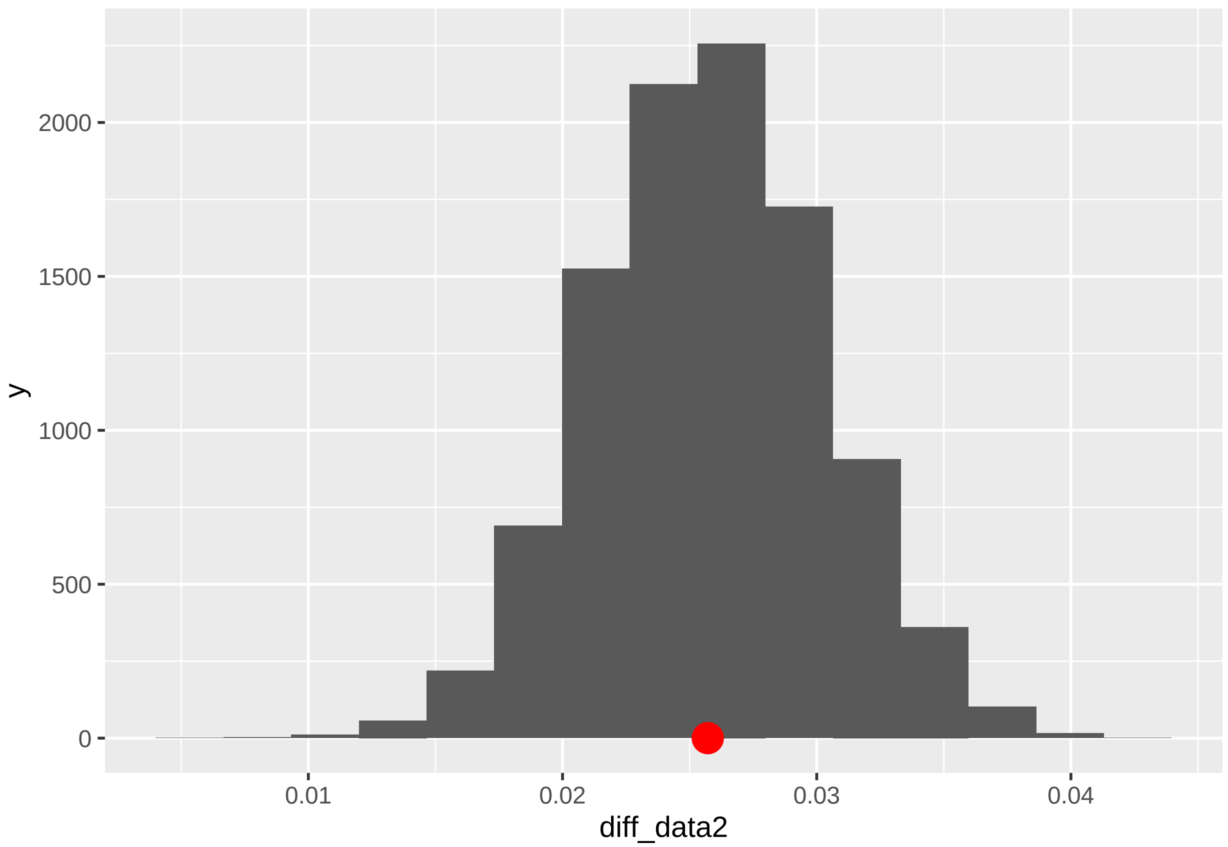
```
diff_data1 = replicate(10000, computDiff(263/350, 87/350, 10000, 263/350))
ggplot() +
geom_histogram(aes(diff_data1), bins=15) +
geom_point(aes(abs(263/350 - 263/350), 0), size=5, color = "red")
```



```
var(diff_data1)
```

```
## [1] 6.875915e-06
```

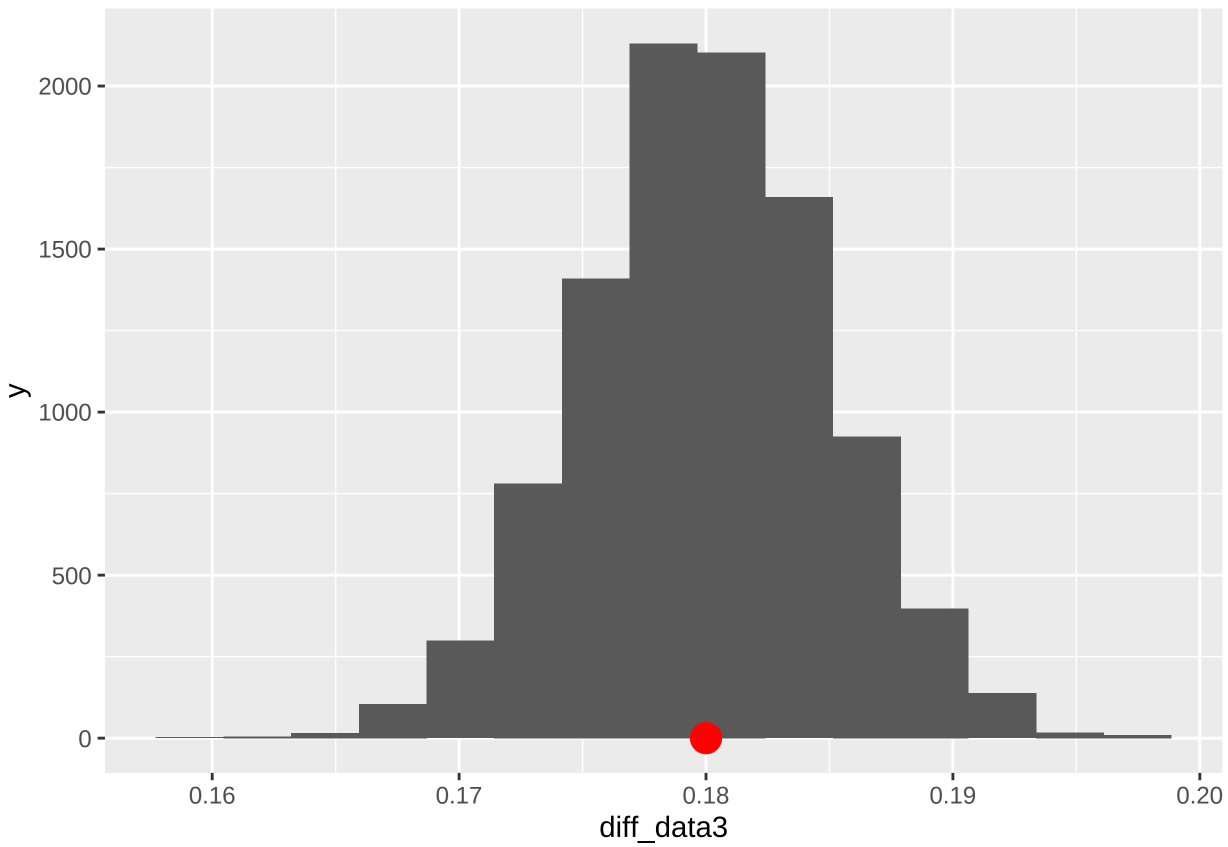
```
#reference proportion = 263/350; observed proportion = 254/350
diff_data2 = replicate(10000, computDiff(254/350, 96/350, 10000, 263/350))
ggplot() +
geom_histogram(aes(diff_data2), bins=15) +
geom_point(aes(abs(254/350 - 263/350), 0), size=5, color = "red")
```



```r
var(diff_data2)
```

```
## [1] 2.025635e-05
```

```r
#reference proportion = 263/350; observed proportion = 200/350
diff_data3 = replicate(10000, computDiff(200/350, 150/350, 10000, 263/350))
ggplot() +
geom_histogram(aes(diff_data3), bins=15) +
geom_point(aes(abs(200/350 - 263/350), 0), size=5, color = "red")
```

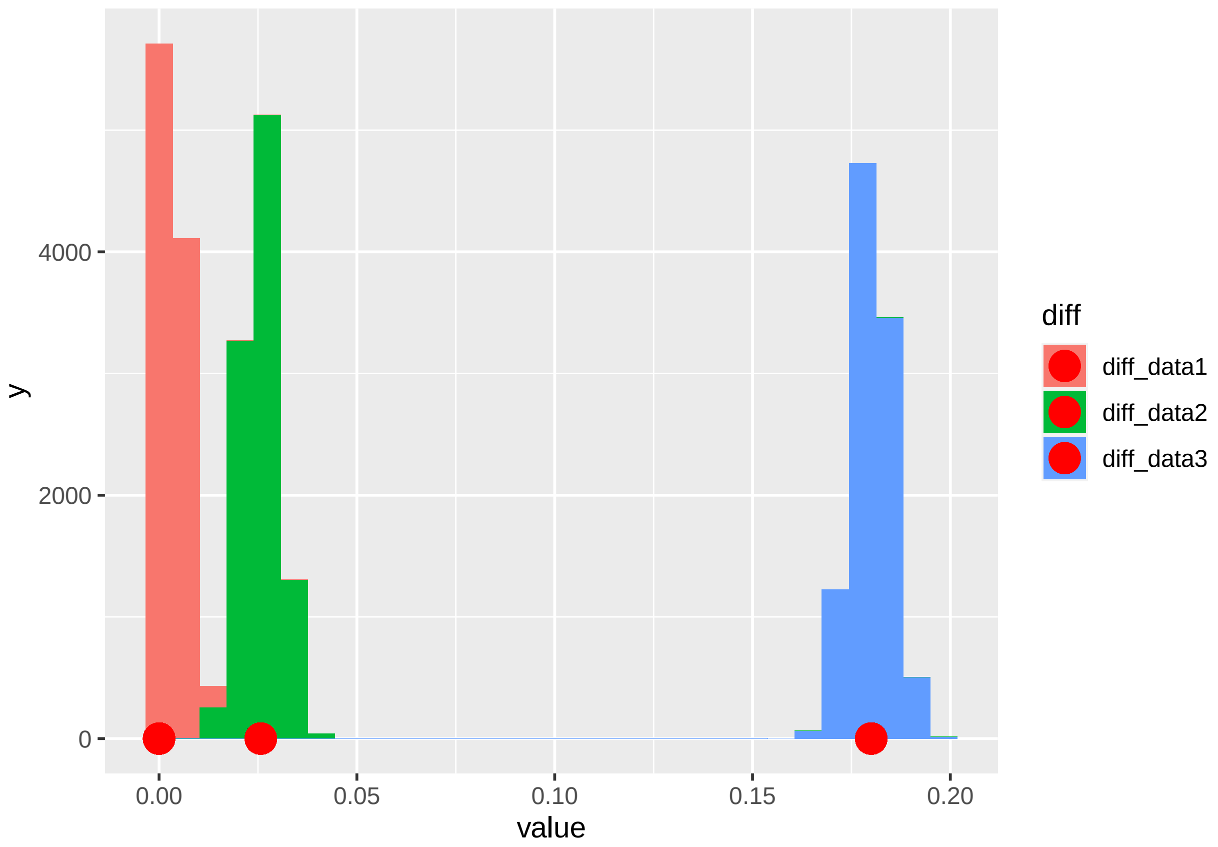


```
var(diff_data3)
```

```
## [1] 2.484783e-05
```

```
#combining plots
diff_data1 <- data.frame(diff_data1)
diff_data1$diff <- "diff_data1"
colnames(diff_data1) <- c("value", "diff")

diff_data2 <- data.frame(diff_data2)
diff_data2$diff <- "diff_data2"
colnames(diff_data2) <- c("value", "diff")

diff_data3 <- data.frame(diff_data3)
diff_data3$diff <- "diff_data3"
colnames(diff_data3) <- c("value", "diff")

diff_df <- rbind(diff_data1, diff_data2, diff_data3)

ggplot(data = diff_df, aes(x = value, fill = diff)) +
geom_histogram() +
geom_point(aes(abs(263/350 - 263/350), 0), size=5, color = "red") +
geom_point(aes(abs(254/350 - 263/350), 0), size=5, color = "red") +
geom_point(aes(abs(200/350 - 263/350), 0), size=5, color = "red")
```

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



According to the variance of the distributions of the test statistics, a normal sampling variation may include 6.76745e-06, 2.052444e-05, and 2.426072e-05.

3. Compute an empirical p-value and explain your findings.

Note that the approach described above is similar to the methodology discussed during our class exercise. However, unlike the procedure we followed in class, where we employed permutations as part of simulating a t-test-like process, this example doesn't necessitate permutations.

```
#p-value: black marbles observed = 254/350 vs expected = 263/350
sum2 <- sum(diff_data1$value > (abs(254/350 - 263/350)))
p_value1 <-  sum2/length(diff_data1)
p_value1
```

```
## [1] 0
```

```
#checking p-value calculation
p_value2 <- t.test(diff_data1$value, mu = abs(254/350 - 263/350), alternative = "two.sided")
p_value2$p.value
```

```
## [1] 0
```

```
#p-value: black marbles observed = 200/350 vs expected = 263/350
sum3 <- sum(diff_data1$value > (abs(200/350 - 263/350)))
p_value1 <-  sum3/length(diff_data1)
p_value1
```

```
## [1] 0
```

```r
#checking p-value calculation
p_value2 <- t.test(diff_data1$value, mu = abs(200/350 - 263/350), alternative = "two.sided")
p_value2$p.value
```

```
## [1] 0
```

There is a significant difference between the observed proportion of black marbles (254/350) and the expected proportion of black marbles (263/350) (p-value = 2.2e-16). There is a significant difference between the observed proportion of black marbles (200/350) and the expected proportion of black marbles (263/350) (p-value = 2.2e-16).