

# STAT 440 Homework 8

Charlie Lu (Cxl5159)

September 2022

## 1 A

```
data_filtered <- subset(data, medium == "Photograph" | medium == "watercolor")
df <- subset(data_filtered, select = c(medium,height,width))

photo <- subset(df, medium == "Photograph")
watercolor <- subset(df, medium == "watercolor")

photo$area <- photo$height*photo$width
watercolor$area <- watercolor$height*watercolor$width

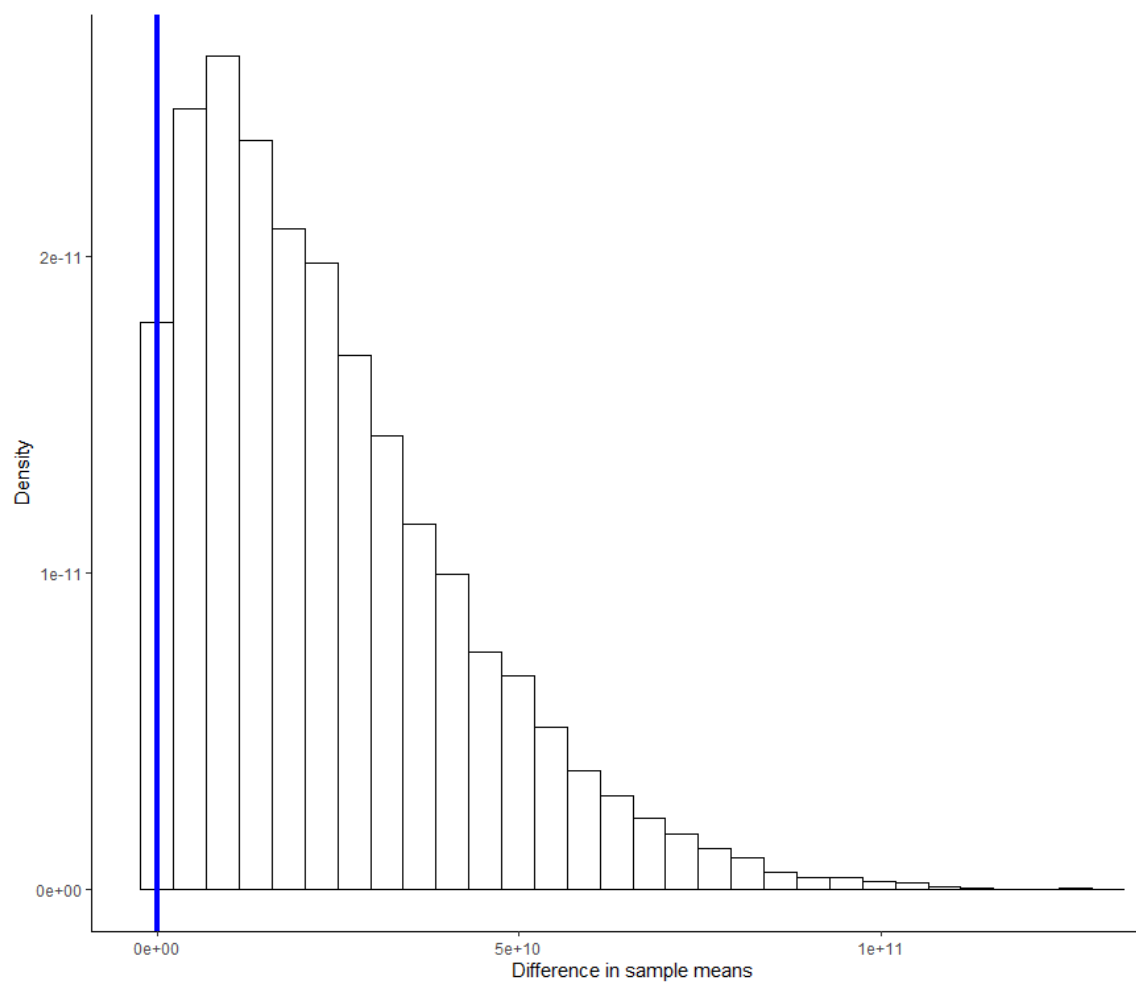
same_dist_perm_test = function(n_perms, xs, ys, test_statistic) {
  #'
  #' perform a generic permutation test
  #' @param n_perms number of permutations to generate
  #' @param xs vector of samples from distribution x
  #' @param yx vector of samples from distribution y
  #' @param test_statistic function that calculates the test statistic

  # calculate the number of samples in x and y
  n = length(xs)
  m = length(ys)
  # define labels (1 = x samples, 0 = y samples)
  labels = c(rep(1, n), rep(0, m))
  all_data = c(xs, ys)

  # for every permutation replication
  replicate(
    n_perms, {
      # permute label orders
      permuted_labels = sample(labels)


      # generate new test statistic under permutation
      test_statistic(all_data[permuted_labels == 1],
                     all_data[permuted_labels == 0])
    }
  )
}

x_obs <- photo$area
y_obs <- watercolor$area
obs_perms <- same_dist_perm_test(10000,photo$area,watercolor$area,function(a, b) {(mean(a)-mean(b))^2})
ggplot(data=data.frame(x=obs_perms)) +
  geom_histogram(aes(x=x, y=..density..), bins=30, color="black", fill="white") +
  geom_vline(xintercept=mean(x_obs) - mean(y_obs), color="blue", size=1.5) +
  xlab("Difference in sample means") + ylab("Density") + theme_classic()
```



## 2 B

```
46 x_obs <- photo$area
47 y_obs <- watercolor$area
48 obs_perms <- same_dist_perm_test(10000, photo$area, watercolor$area,
49                                 function(a, b){(mean(a)-mean(b))^2})
50 ggplot(data=data.frame(x=obs_perms)) +
51   geom_histogram(aes(x=x, y=..density..), bins=30, color="black", fill="white") +
52   geom_vline(xintercept=mean(x_obs) - mean(y_obs), color="blue", size=1.5) +
53   xlab("Difference in sample means") + ylab("Density") + theme_classic()
54
55 p_value = mean(obs_perms > (mean(x_obs) - mean(y_obs))^2)
56 p_value
```

32:3  same\_dist\_perm\_test(n\_perms, xs, ys, test\_statistic) ↕

Console	Terminal ×	Background Jobs ×
R 4.2.1 · C:/Users/Charlie Lu/Desktop/ ↗		
<pre>&gt; p_value = mean(obs_perms &gt; (mean(x_obs) - mean(y_obs))^2) &gt; p_value [1] 0.9863 &gt;</pre>		

This P-Value that we get is really large (0.9863) which indicates that we should accept the null hypothesis of the mean of watercolors and mean of photographs being equal. In this context it makes sense because the blue bar from the graph represents the mean of the observed data, and it's almost directly on top of zero, making our resulting p-value make sense. As for the two-sided aspect of the question, the difference here is squared, so that will naturally include both sides within a single tail.