140.615.HW5.Delahanty.Jeremy

Jeremy Delahanty

2024-03-10

1

1a) Assuming you plan on running ELISAs for the same number of cases and controls, how many subjects do you need to have 80% power to detect a two-fold change in abundance between cases and controls, i. e. delta = 1, if the within-group standard deviation of the log2(abundances) is the same as the differences in means of the log2(abundances), i. e. $\sigma = \delta$?

In this question, we are assuming that σ is the same value as δ. Specifically, we are assuming that:

$\delta = \sigma = 1$

Further, we are seeking a power value of 0.8. We can use the `power.t.test` function in `R` to accomplish this calculation.

```
# Use power.t.test for calculating the n required for our desired power
power.t.test(delta = 1, sd = 1, power = 0.8)
```

```
##
##        Two-sample t test power calculation
##
##              n = 16.71477
##          delta = 1
##             sd = 1
##      sig.level = 0.05
##          power = 0.8
##    alternative = two.sided
##
## NOTE: n is number in *each* group
```

This shows that we require 16.71 individuals, or 17 subjects, per group.

1b) How much power would you have if you only ran ten ELISAs per group?

Here, we can use the same function in `R` but instead ask the program to solve for power.

```
# Use power.t.test again, but this time only supply delta, sd, and n
power.t.test(n = 10, delta = 1, sd = 1)
```

```
## 
##      Two-sample t test power calculation 
## 
##               n = 10
##           delta = 1
##              sd = 1
##       sig.level = 0.05
##           power = 0.5619846
##     alternative = two.sided
## 
## NOTE: n is number in *each* group
```

This shows that we would only have a value of power of 0.5620, or 56.20%, when only taking 10 samples per group.

1c) How much power would you have if you only ran ten ELISAs per group, but knew that in truth the protein your are interested could not be down-regulated in the affected probands?

Given the knowledge that our dataset *cannot* experience a decrease in the values assayed from the ELISA, we do not need to concern our tests with two-sided values. Instead, we can use a one-sided test using the same function in R.

```
# Use power.t.test again with the same arguments as 1b, but specify that the
# alternative is a one sided test
power.t.test(n = 10, delta = 1, sd = 1, alternative = "one.sided")
```

```
## 
##      Two-sample t test power calculation 
## 
##               n = 10
##           delta = 1
##              sd = 1
##       sig.level = 0.05
##           power = 0.6935575
##     alternative = one.sided
## 
## NOTE: n is number in *each* group
```

We see that our power has increased slightly to a value of 0.6936!

2

A physician wants to estimate the average body mass index (BMI) for her adult patients. She decides to draw a sample of clinical records and retrieve this information from them. She wants an estimate with a margin of error of 1.5 units of BMI with 95% confidence (i.e. the confidence interval should be 3 units of BMI in length), and believes that the national population standard deviation of adult BMI of 4.5 also applies to her patients. She knows that BMI is approximately normally distributed for adults. How large a sample does she need to draw?

This question asks us for a sample size calculation that accomplishes a margin of error with 1.5 units of BMI at 95% confidence. This margin of error is equal to:

$MOE = \frac{CI\ Width}{2}$

Since we can assume that BMI is approximately normally distributed, we can apply a normal distribution. This allows us to use the formula:

$MOE\ 0.95 = \frac{Z \times \sigma}{\sqrt{n}}$

Since we have a normal distribution, we can use a Z value of 1.96 for the value of a 95% confidence interval. Our value of sigma σ is stated to be 4.5. We must finally solve for the value of n.

```r
# Define the value for the margin of error, width of confidence interval over 2
moe <- 3/2

# Define value of z for 95% confidence interval
z <- 1.96

# Define value for sigma given in question
sd <- 4.5

# Calculate the sample size using our formula
sample_size <- ((z * sd)/moe)^2

# Display the value
cat("Required Sample Size:", sample_size)
```

```
## Required Sample Size: 34.5744
```

We can see that the required sample size given these parameters is 34.5744, or approximately 35 individuals per group.

3

Consider data on the treatment response of 12 mice from strain A and 9 mice from strain B. Assume that the true within-group standard deviations are the same.

3a) Test the hypothesis $H_O \mu_A = \mu_B$ against the alternative $H_A \mu_A \neq \mu_B$ with the t.test() function.

We can simply use the t.test() function in R to solve this question.

3

```
# Define data vectors
strain_a <- c(55.2,58.1,41.7,44.9,44.8,48.9,47.5,48.1,48.4,51.6,40.6,48.0)
strain_b <- c(48.7,52.6,65.2,70.4,44.2,54.7,44.0,66.5,56.8)

# Use the t.test function for strain a against strain b
a_vs_b <- t.test(strain_a, strain_b, var.equal = TRUE)

# Display complete t.test result
a_vs_b
```

```
##
##  Two Sample t-test
##
## data:  strain_a and strain_b
## t = -2.3817, df = 19, p-value = 0.02784
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -14.5607027  -0.9392973
## sample estimates:
## mean of x mean of y
##     48.15     55.90
```

```
# Display the value of the p value
cat("P Value for strain A and B means being different:", a_vs_b$p.val)
```

```
## P Value for strain A and B means being different: 0.02784457
```

3b) Use the wilcox.test() function to perform a rank-sum test on these data.

In R we can do this simply on our previously defined data.

```
# All our data was defined in the previous cell
# Use the Wilcox Test function
wilcox.test(strain_a, strain_b, var.equal = TRUE)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  strain_a and strain_b
## W = 29, p-value = 0.08154
## alternative hypothesis: true location shift is not equal to 0
```

3c) Give an interpretation of your results.

While it appears that our T-Test shows a statistically significant difference between the two means, it assumes that the data being sampled comes from a normal distribution. We should first confirm whether or not the data does indeed appear to be normally distributed through plotting.

First, we can try using a strip chart to inspect the data.

```
# Use strip chart to display our data as described in lecture and lab
stripchart(list(strain_a, strain_b), vertical = T, pch = 21, xlim = c(00.5, 2.5))
segments(0.9, mean(strain_a), 1.1, mean(strain_a), lwd = 2, col = "red")
segments(1.9, mean(strain_b), 2.1, mean(strain_b), lwd = 2, col = "blue")
```
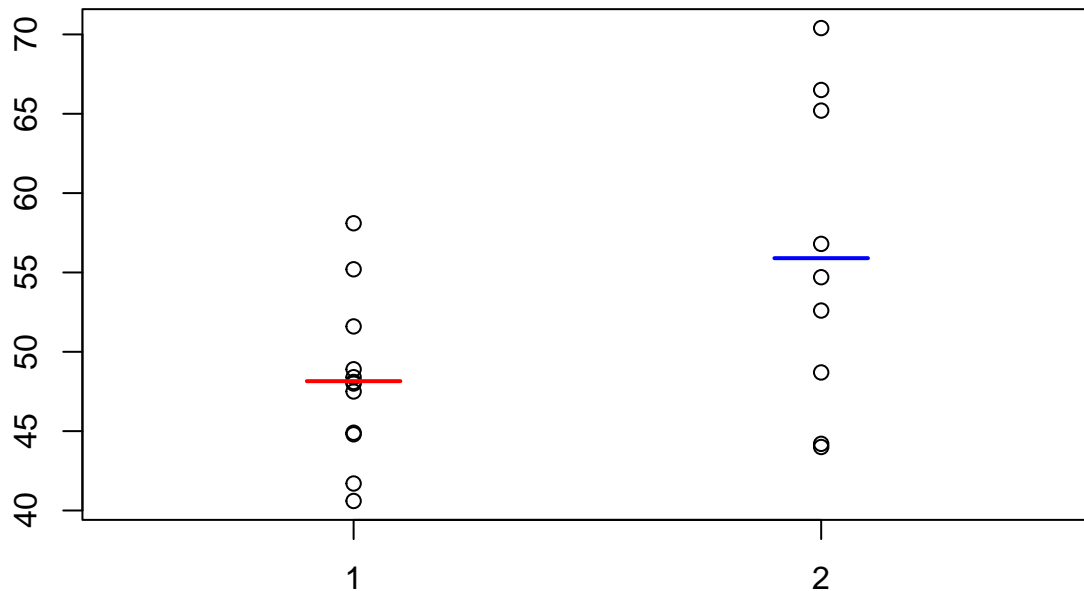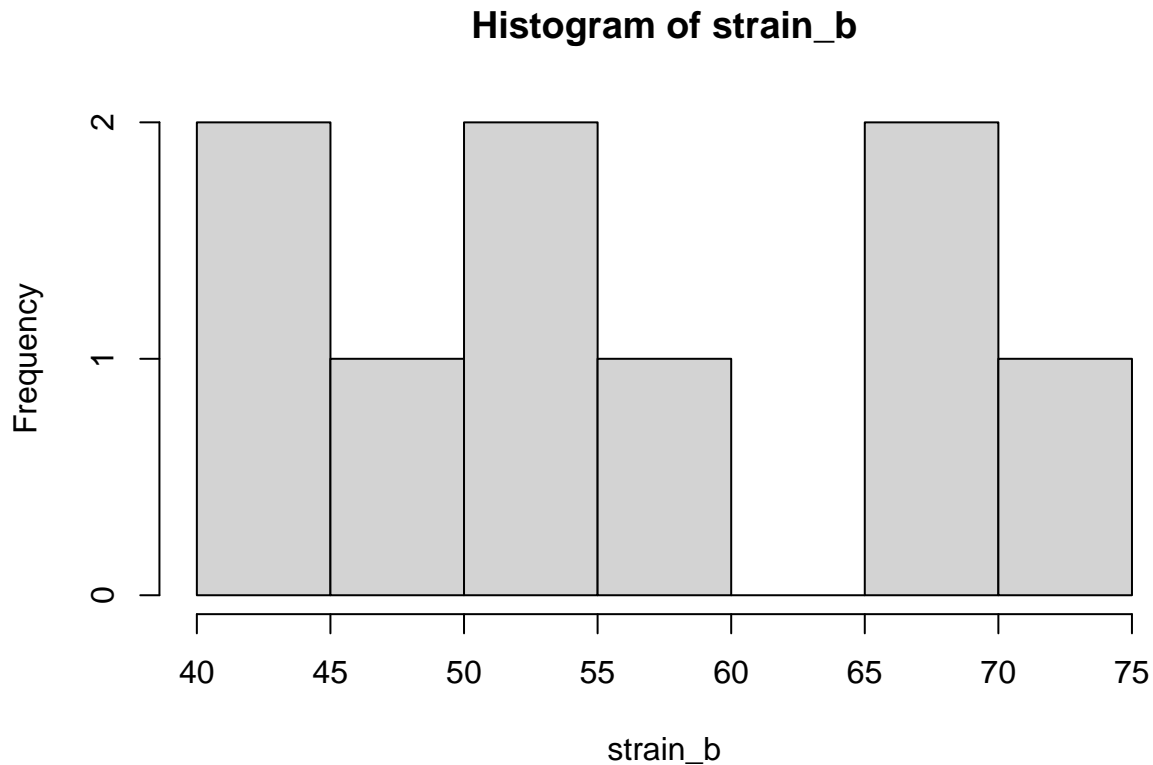

The data could be plausibly normal when it looks like this, but it would be better to confirm with the a histogram of each of the data.

```
# Plot a histogram for strain a
hist(strain_a)
```

Histogram of strain_a
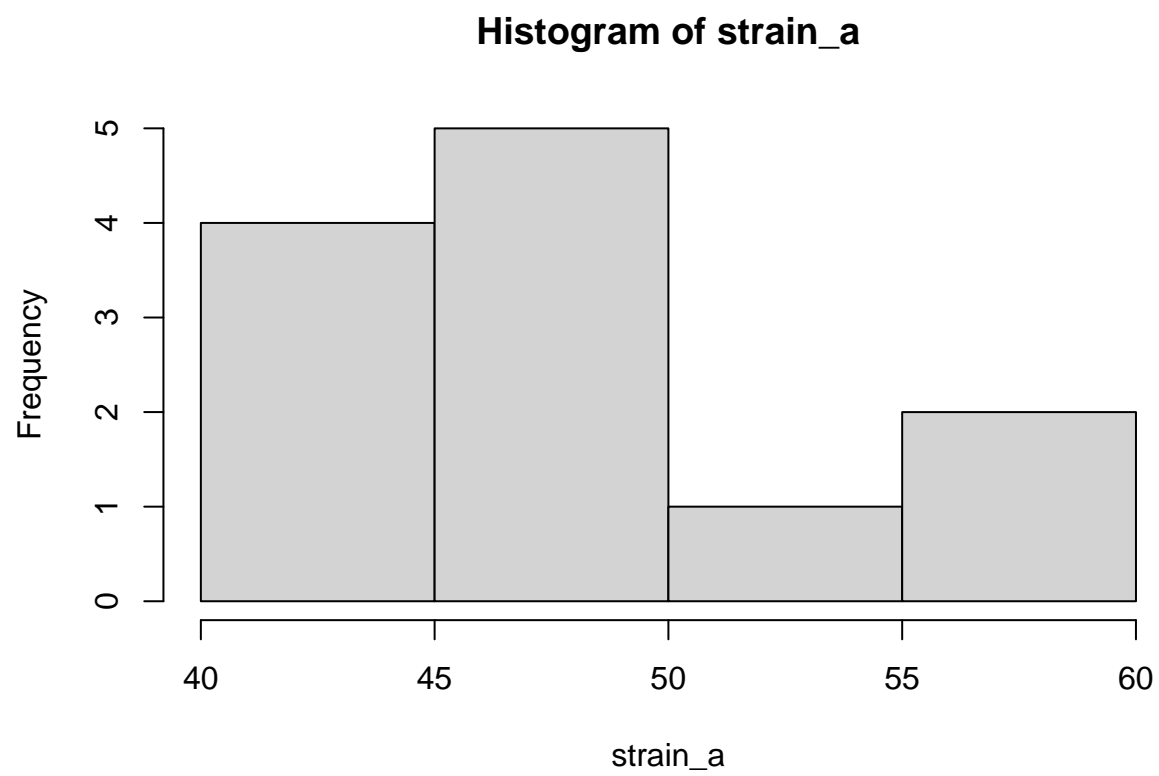
strain_a

```r
# Plot a histogram for strain b
hist(strain_b)
```

Histogram of strain_b

We can see that our data does not actually appear to be normally distributed. This can be further confirmed with the use of the Shapiro-Wilk Normality test for each strain.

```
# perform the shapiro-wilk test for each dataset
shapiro.test(strain_a)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  strain_a
## W = 0.95194, p-value = 0.6656
```

```
shapiro.test(strain_b)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  strain_b
## W = 0.929, p-value = 0.4719
```

We can see that this data does not demonstrate normality.

Given that the data does not show normality, we should be very cautious in using a parametric test like the t-test and instead use the analysis provided by the Rank Sum test. Since the value of that test is above our defined significance level of $\alpha = 0.05$, we would fail to reject the null hypothesis.

4

4a) Do the values of the react data set (notice that this is a single vector, not a data frame) look reasonably normally distributed? Does the mean differ significantly from zero according to a t-test?

To see if the datasest looks normally distributed, we can look at both a stripchart as well as a histogram of the dataset first.
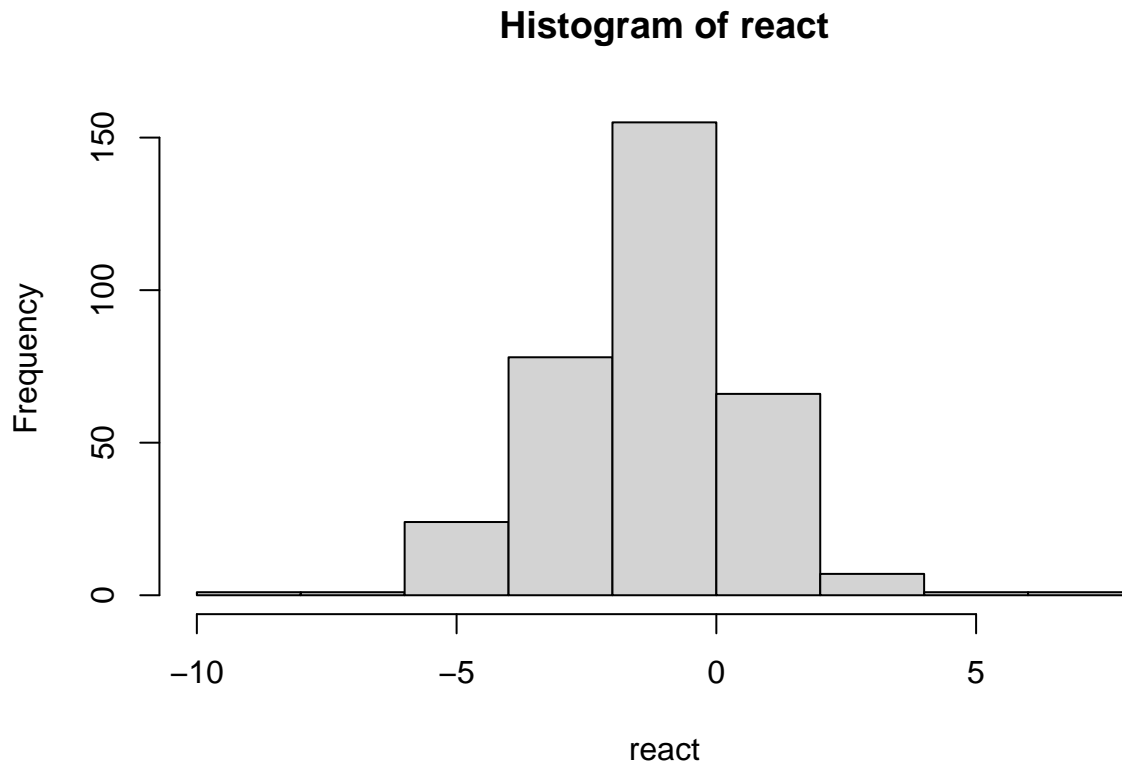
```r
# Use the ISwR package for the dataset
library(ISwR)
```

```
## Warning: package 'ISwR' was built under R version 4.3.3
```

```r
# Plot the react data first as a strip chart
stripchart(react)
```
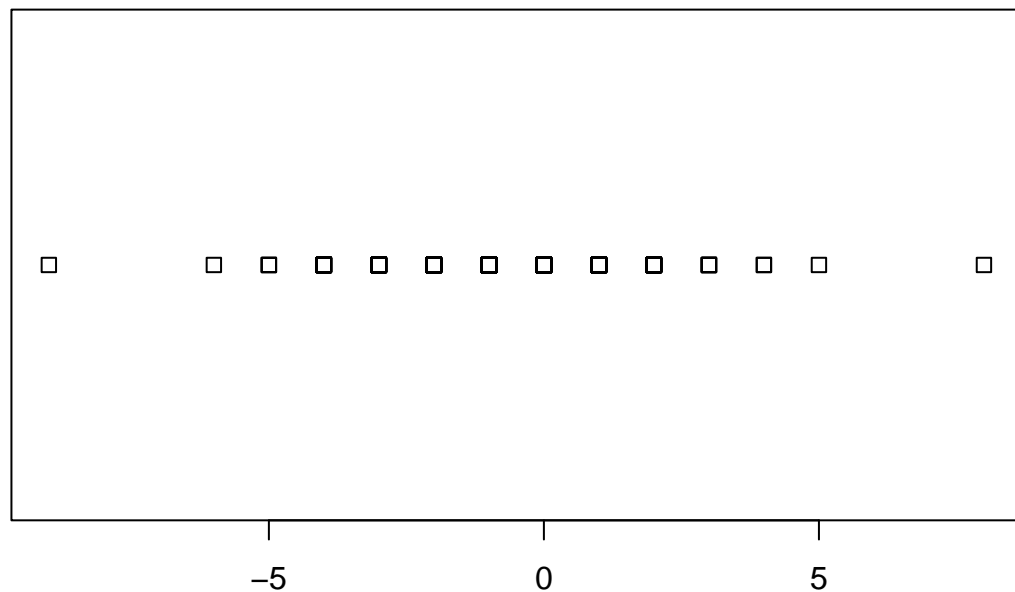


```r
# Now plot a histogram of the data
hist(react)
```

Histogram of react

Looking at both of these plots appears to confirm that the data appears normally distributed. With this satisfied, we can perform the simple `t.test()` function in R.

```
# Perform a t test on the react dataset where it compares if the mean is 0
t.test(react, mu = 0)
```

```
##
##  One Sample t-test
##
## data:  react
## t = -7.7512, df = 333, p-value = 1.115e-13
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.9985214 -0.5942930
## sample estimates:
##  mean of x
## -0.7964072
```

We can see that our p-value is extremely small at 1.115e-13 with a confidence interval between -0.99 ad -0.59, not including the value of 0. Thus, we can conclude that the mean is statistically significantly different than 0 and reject the null hypothesis.

4b)) In the data set vitcap, use a t-test to compare the vital capacity for the two groups. Can we conclude that the population means differ? Calculate a 99% confidence interval for the difference. The result of this comparison may be misleading. Why?

We can calculate the 99% confidence interval of the data using the `t.test` function in `R`.

```r
# Extract the datasets of the two groups by selecting the relevant columns
group_a <- vitcap[1:12, "vital.capacity"]
group_b <- vitcap[13:24, "vital.capacity"]

# Use the built in t.test function for determining if the differences in means are
# different at a confidence level of 99%
t.test(group_a, group_b, conf.level = 0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  group_a and group_b
## t = -2.9228, df = 19.019, p-value = 0.008724
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -2.06447665 -0.02219002
## sample estimates:
## mean of x mean of y
##  3.949167  4.992500
```

We can see that our p-value is lower than $\alpha = 0.01$ and, at first glance, conclude that our confidence interval between -2.0644 and -0.02219 demonstrates that the means are indeed statistically significantly different from zero. We could reject the null hypothesis given this dataset.

However, in order to use the t-test responsibly, we should first demonstrate that the data is approximately normal through plotting and the Shapiro-Wilk test for normality.

```r
# First plot the data for each group
hist(group_a)
```

Histogram of group_a


```r
hist(group_b)
```
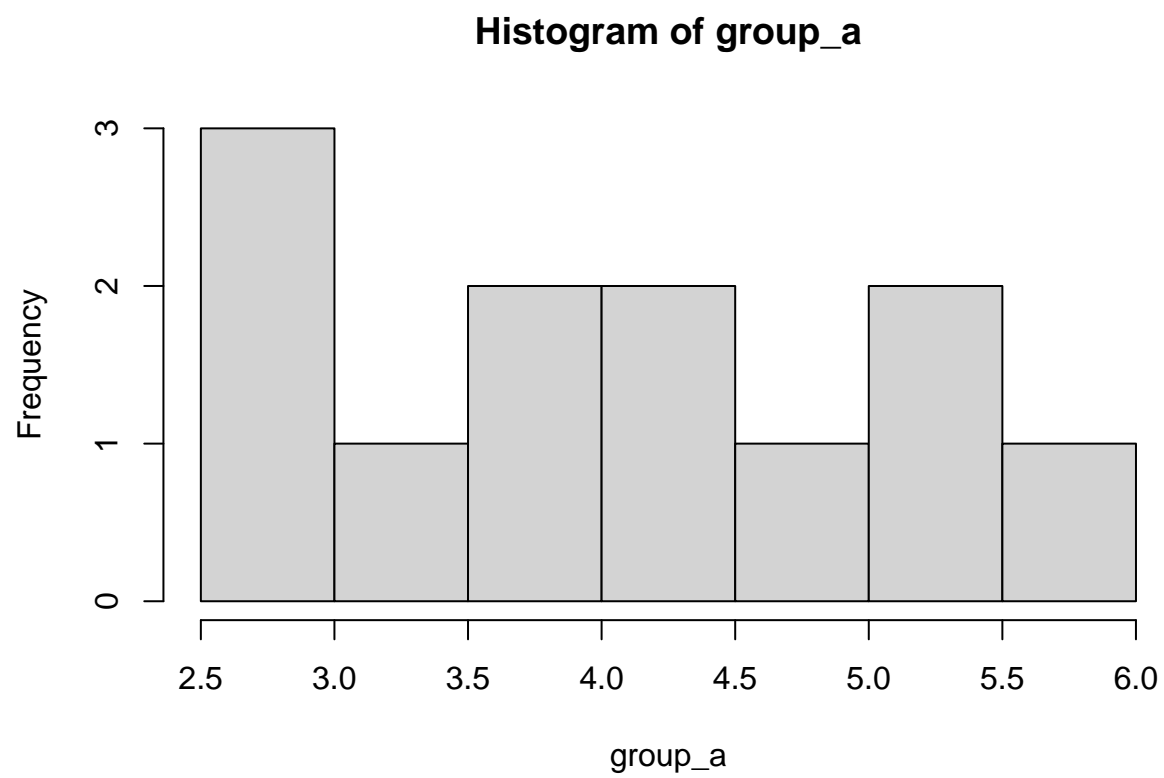
Histogram of group_b
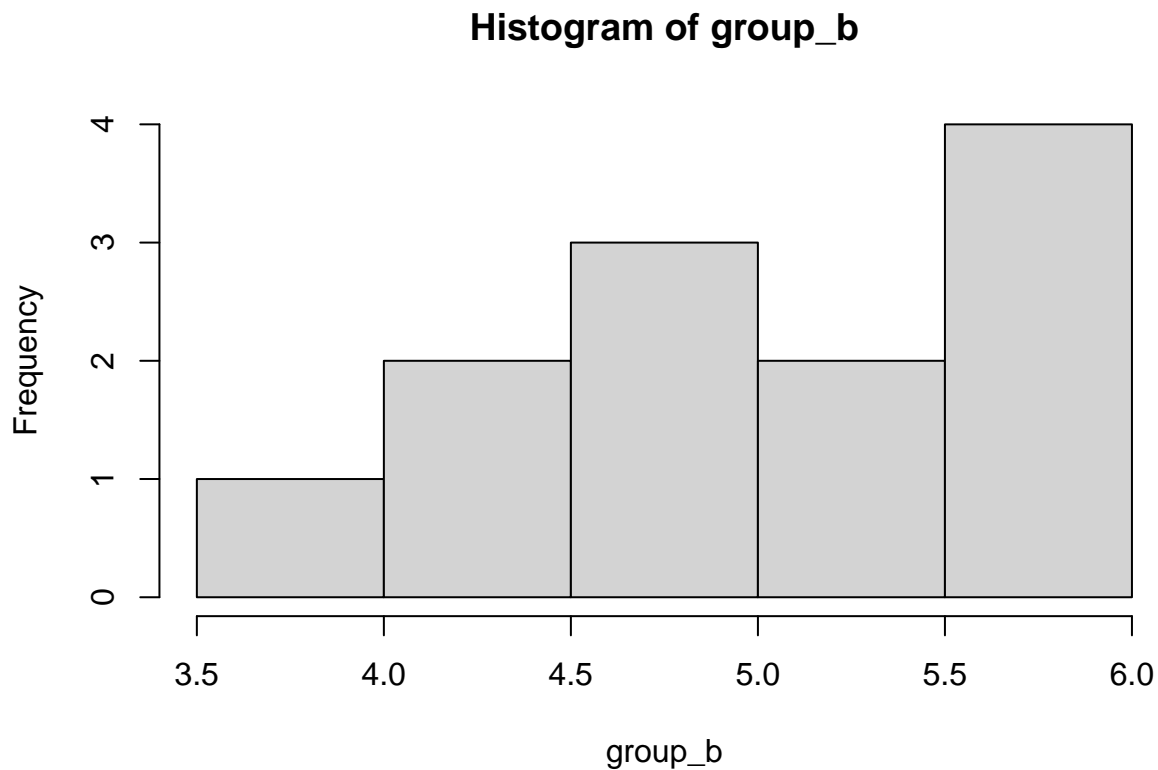
We see here that the data does not appear normally distributed! We can confirm with the Shapiro-Wilk test for normality.

```
# Perform test for normality for each dataset
shapiro.test(group_a)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  group_a
## W = 0.91633, p-value = 0.257
```

```
shapiro.test(group_b)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  group_b
## W = 0.93421, p-value = 0.4269
```

We see here that neither dataset successfully meets the criteria for normality! Thus it can be misleading to reject the null hypothesis when using the T-Test function.

4c) Perform the analyses of the react and vitcap data using nonparametric techniques.

In this case, we can use the non-parametric Wilcox Rank Sum test for each dataset. Remember, our data in the react dataset is appears normal while our vitcap dataset does not appear to be normally distributed.

First, we can use the Wilcox test for the react dataset in `R`.

```
# Perform the wilcox test on the react dataset and test if the mean is equal to
# zero
wilcox.test(react, mu = 0)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  react
## V = 9283.5, p-value = 2.075e-13
## alternative hypothesis: true location is not equal to 0
```

We can see that our p-value is very small at 2.075e-13 and, using this test, can reject the null hypothesis that the react dataset is equal to zero.

Now, lets use this test for testing the difference in means between our two groups in the vitcap dataset. We will use the same value for alpha at 99% confidence.

```
# Perform the wilcox test for the difference in means between two groups
wilcox.test(group_a, group_b, conf.level = 0.99)
```

```
## Warning in wilcox.test.default(group_a, group_b, conf.level = 0.99): cannot
## compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  group_a and group_b
## W = 30.5, p-value = 0.01783
## alternative hypothesis: true location shift is not equal to 0
```

We see that by using the Wilcox rank sum test, we cannot reject the null hypothesis any longer! This demonstrates the difference between the non-parametric Wilcox test and the parmetric t-test.