# Lesson 17 Companion

**Research question:**

For this document I will use the breastfeeding vs. intelligence data set from the book. The research question here is whether or not breastfeeding has a connection with cognitive functioning at age four.

$H_0$ : There is no association between breastfeeding and general cognitive index (GCI) at age four.
$H_a$ : There is an association between breastfeeding and general cognitive index (GCI) at age four.

Our explanatory variable is the binary categorical variable of whether or not a child was breastfed and the response is that child's GCI at age four. The parameter of interest is the population difference in average CGI between the groups (breastfed/not) ($\mu_{breastfed} - \mu_{not}$), the relevant statistic is the sample difference in average GCI ($\bar{x}_{breastfed} - \bar{x}_{not}$), and we can rewrite the hypothesis as follows:

$H_0 : \mu_{breastfed} - \mu_{not} = 0$
$H_a : \mu_{breastfed} - \mu_{not} \neq 0$

```
library(tidyverse)

GCI = read_table2("http://www.isi-stats.com/isi/data/chap6/BreastFeedIntell.txt")

head(GCI)
```

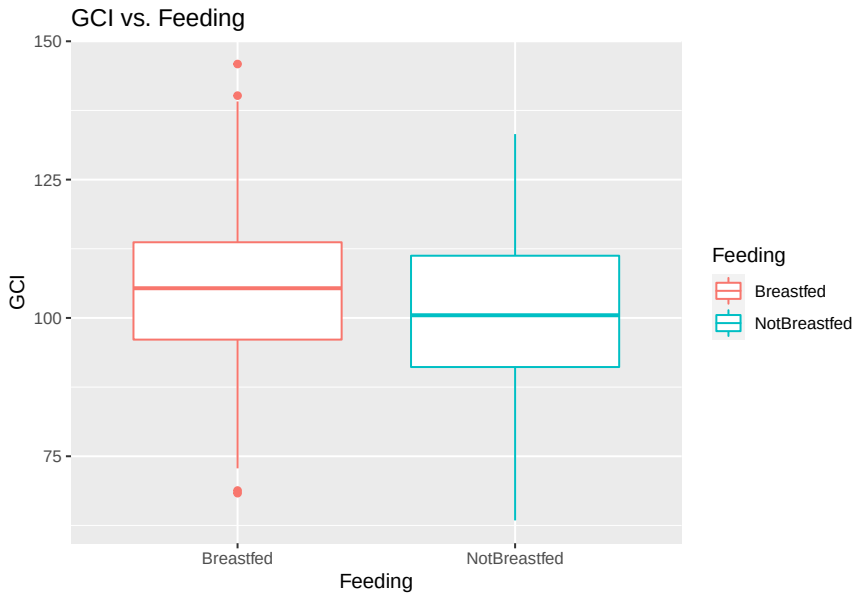
```
## # A tibble: 6 x 2
##   Feeding     GCI
##   <chr>     <dbl>
## 1 Breastfed 127.
## 2 Breastfed 125.
## 3 Breastfed  99.8
## 4 Breastfed 105.
## 5 Breastfed  97.3
## 6 Breastfed 131.
```

```
feeding_means = GCI %>%
  group_by(Feeding) %>%
  summarise(mean = mean(GCI))

sample_stat = feeding_means[[2]][[1]] - feeding_means[[2]][[2]]

null_mean = 0
```

**Data exploration:**

The best choice of visualizations for this data set is a box plot.

```
#Don't need to color, was just too boring without it
GCI %>%
  ggplot(aes(x = Feeding, y = GCI, color = Feeding)) +
  geom_boxplot() +
  labs(x = "Feeding", y = "GCI",
       title = "GCI vs. Feeding")
```




From this box plot, it appears as if the median GCI for breastfed children is slightly higher in the sample. The variability is also slightly lower in this group. There appears to be three outliers in the breastfed group: two on the high end of the GCI and one on the low end.

We can also produce a five number summary to get the same information in numerical form:

```
GCI %>%
  group_by(Feeding) %>%
  summarise(n = n(),
            min = fivenum(GCI)[1],
            Q1 = fivenum(GCI)[2],
            median = fivenum(GCI)[3],
            Q3 = fivenum(GCI)[4],
            max = fivenum(GCI)[5])
```

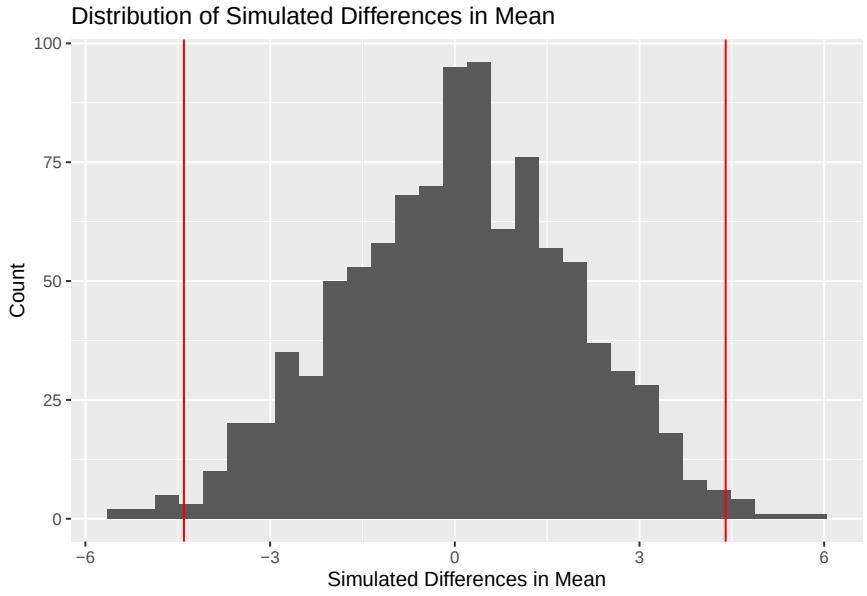
```
## # A tibble: 2 x 7
##   Feeding          n   min    Q1 median    Q3   max
##   <chr>        <int> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 Breastfed      237  68.3  96.1   105.  114.  146.
## 2 NotBreastfed    85  63.4  91.1   100.  111.  133.
```

**Simulation-based approach:**

The first step is to build a null distribution which represents the GCI differences between the groups we'd expect to see if there was no association between breastfeeding and GCI. I hope by this point you have completely inculcated what's going on with the null distribution and why we build it.

```
replications_dataframe = NULL

num_reps = 1000

for (i in 1:num_reps){

  sim_GCI = GCI %>%
    mutate(new_Feeding = sample(Feeding, size = n(), replace = FALSE))
```

```
  feeding_means = sim_GCI %>%
    group_by(new_Feeding) %>%
    summarise(mean = mean(GCI))

  trial_stat = feeding_means[[2]][[1]] - feeding_means[[2]][[2]]

  replications_dataframe = rbind(replications_dataframe, data.frame(trial_stat))

}

replications_dataframe %>%
  ggplot(aes(x = trial_stat)) +
  geom_histogram() +
  labs(x = "Simulated Differences in Mean", y = "Count",
       title = "Distribution of Simulated Differences in Mean") +
  geom_vline(xintercept = -sample_stat, color = "red") +
  geom_vline(xintercept = sample_stat, color = "red")
```




```
replications_dataframe %>%
  summarise(pvalue = sum(abs(trial_stat) >= abs(sample_stat)) / n())
```

```
##   pvalue
## 1  0.017
```

The p-value of 0.017 suggests that our sample provides strong evidence against the null hypothesis which suggests there is a difference GCI at age four between breastfed and non-breastfed children.

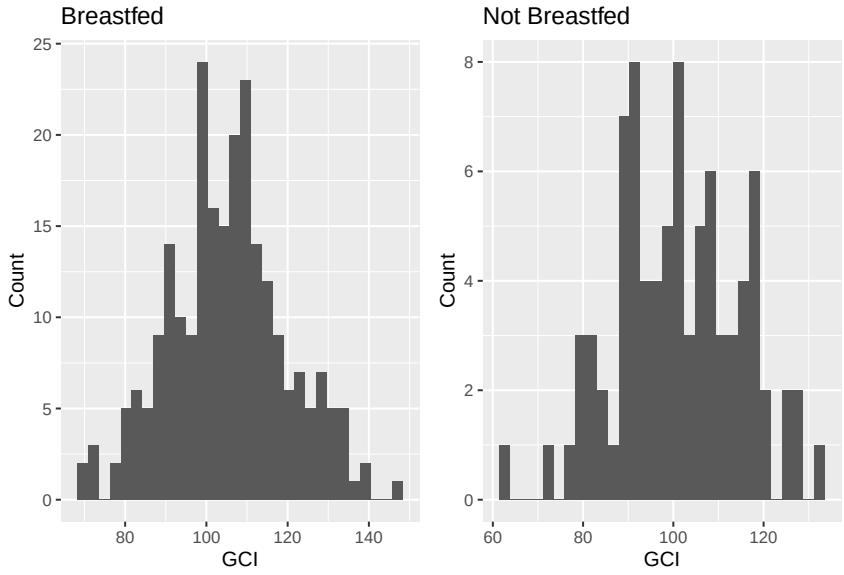
**Two-sample t-test:**

The validity conditions for the theory-based approach are:

- The quantitative variable (GCI in this case) should a symmetric distribution in both groups **OR**
- There should be at least 20 observations in each group and the sample distributions should not be strongly skewed.

No matter what we are doing to have to take a look at the distributions of GCI in both groups.

```r
library(patchwork)

p1 = GCI %>%
  filter(Feeding == "Breastfed") %>%
  ggplot(aes(x = GCI)) +
  geom_histogram() +
  labs(x = "GCI", y = "Count",
       title = "Breastfed")

p2 = GCI %>%
  filter(Feeding == "NotBreastfed") %>%
  ggplot(aes(x = GCI)) +
  geom_histogram() +
  labs(x = "GCI", y = "Count",
       title = "Not Breastfed")

p1 + p2
```




These histograms looks pretty symmetrical to me so we actually meet both conditions (even though we really one need to meet one).

Because we meet the validity conditions, we can now use our theory-based approach for statistical inference. For these types of problems we want to use the unpaired two-sample t-test (more to follow on "unpaired"). I will demonstrate two methods to perform this test. The first is not really the preferred method (you'll see why) but it brings up an important point:

```r
#Sample sizes and standard deviations for each group
group_stats = GCI %>%
  group_by(Feeding) %>%
  summarise(sd = sd(GCI),
            n = n())

sd_breastfed = group_stats[[2]][[1]]

n_breastfed = group_stats[[3]][[1]]
```

```
sd_not = group_stats[[2]][[2]]

n_not = group_stats[[3]][[2]]
```

From here it would be a matter of calculating the standard error of the statistic and the t-statistic and using **pt()** but you may recall that we need a *degrees of freedom* parameter for that function. How we calculate these last few values (including the degrees of freedom value) depends on the type of two-sample t-test we are performing: Student's t-test or Welch's t-test.

Student's t-test relies on equal variances of the quantitative variables between groups. In other words, the GCI values would need approximately equal variance in both the breastfed children and the non-breastfed children. Let's take a look at how we would test whether the variances were approximately equal.

**Fisher's F-test**

We can use Fisher's F-test to calculate a p-value to demonstrate whether there is evidence to believe that the variances are approximately equal. Our hypotheses for this test would be:

$H_0 : \sigma^2_{breastfed} = \sigma^2_{not}$
$H_a : \sigma^2_{breastfed} \neq \sigma^2_{not}$

```
#"Degrees of freedom" values for the F-test
df_breastfed = n_breastfed - 1
df_not = n_not - 1

F_stat = (sd_breastfed^2) / (sd_not^2)

2 * (1 - pf(F_stat, df1 = df_breastfed, df2 = df_not))
```

```
## [1] 0.7194021
```

A p-value of 0.719 means our sample offers little to no evidence against the null hypothesis which suggests we could make the conclusion that there is no significant difference between our variances. Coincidentally, *R* has a shorter method of doing this (of course):

```
var.test(GCI$GCI ~ GCI$Feeding)
```

```
##
##  F test to compare two variances
##
## data:  GCI$GCI by GCI$Feeding
## F = 1.0727, num df = 236, denom df = 84, p-value = 0.7194
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7416205 1.5055995
## sample estimates:
## ratio of variances
##           1.072705
```

As we have evidence to suggest the variances are not significant different, we could use Student's t-test for this situation. However, here is a piece of advice (and prepared to be disappointed): **just use Welch's t-test**. All is not lost though because this is the second time you've interacted with the F-statistic and we're not even to that section of the book yet!

So we've decided that we are always going to use the Welch's t-test but, for completeness, I'm going to include the code to the remaining values for Student's t-test as well.

```
#Student's t-test
#Don't run this!
pooled_sd = sqrt((((n_breastfed - 1) * sd_breastfed^2) + ((n_not - 1) * sd_not^2)) /
                   ((n_breastfed - 1) + (n_not - 1)))

stand_error = pooled_sd * sqrt((1 / n_breastfed) +
                                 (1 / n_not))
df = n_breastfed + n_not - 2

t_stat = (sample_stat - null_mean) / stand_error

####################

#Welch's t-test
#These are the one's you want!

#Standard error of the sample statistic
stand_error = sqrt((sd_breastfed^2 / n_breastfed) +
                     (sd_not^2 / n_not))

t_stat = (sample_stat - null_mean) / stand_error

#This is really bad, see below for a better depiction of it.
df = ((sd_breastfed^2 / n_breastfed) + (sd_not^2 / n_not))^2 /
  ((sd_breastfed^4 / (n_breastfed^2 * (n_breastfed - 1))) +
     (sd_not^4 / (n_not^2 * (n_not - 1))))
```

Welch's t-test Degrees of Freedom Equation:

$$\frac{(\frac{s_a^2}{n_a}+\frac{s_b^2}{n_a})^2}{\frac{s_a^4}{n_a^2\times(n_a-1)}+\frac{s_b^4}{n_b^2\times(n_b-1)}}$$

Here the **a** group is *breastfed* and the **b** group is *not*.

```
#Find the p-value
2 * (1 - pt(abs(t_stat), df = df))
```

```
## [1] 0.01491057
```

Our p-value from the Welch's t-test is a little different than the one we got from our simulation-based approach but not enough to change our conclusion. Let's now take a look at the easy method using **t.test()**.

```
breastfed_data = GCI %>%
  filter(Feeding == "Breastfed")

not_data = GCI %>%
  filter(Feeding == "NotBreastfed")

t.test(x = breastfed_data$GCI, y = not_data$GCI)
```

```
##
##  Welch Two Sample t-test
##
## data:  breastfed_data$GCI and not_data$GCI
## t = 2.4624, df = 153.01, p-value = 0.01491
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  0.8698749 7.9302245
## sample estimates:
## mean of x mean of y
##     105.3     100.9
```

Like most of the time, I wait to show you the easy way at the end. You can see that we need to separate our data frame into two data frames to run this test. You can also see that **t.test()** defaults to the Welch's t-test. Because I know you are interested, you can force it to do a Student's t-test using this command:

```
t.test(x = breastfed_data$GCI, y = not_data$GCI, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  breastfed_data$GCI and not_data$GCI
## t = 2.4218, df = 320, p-value = 0.016
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.8256165 7.9744830
## sample estimates:
## mean of x mean of y
##     105.3     100.9
```

**Confidence intervals:**

We can use **t.test()** to get a confidence interval but this wouldn't be MA256 if I don't show you the long way as well.

```
lower = sample_stat + qt(0.025, df = df) * stand_error

upper = sample_stat - qt(0.025, df = df) * stand_error

paste("(", lower, ",", upper, ")")
```

```
## [1] "( 0.869874947233419 , 7.93022453154544 )"
```

To interpret this we can say that we are 95% confident that breastfed children score between 0.870 and 7.93 points higher in GCI score than non-breastfed babies.

**Practical importance:**

It is clear that we our sample provides statistically significant evidence of a different in GCI scores between breastfed and non-breastfed babies. The 95% confidence interval places the true difference between 0.870 and 7.93 points. Is this a practical difference? I believe that this is a topic on which you can have a defensible position on either side. The real tasks if building a defense for whatever position you take.

Whenever I consider question like this the phrase "opportunity cost" comes to mind. I think the main opportunity cost here is the time spent breastfeeding the child versus not breastfeeding them. There a multitude of other costs that you could consider (formula prices, mother's medical issues, etc.) but, for me, the opportunity cost of (maybe) the extra time spent breastfeeding doesn't outweigh the potential benefits received. This is a tough one because irrational decisions come into play when children are involved, but I stand by my position.