MATH 140 **Name** _____ **Date:**

Worksheet: Matched Pairs

Paired Data: Two sets of observations are **paired** if each observation in one set has a special correspondence or connection with exactly one observation in the other data set. To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations, and do inference on these differences.

The Scene: The data set `textbooks` in OpenIntro Stats gives the price of textbooks in the UCLA bookstore against the price of these books on Amazon. Here's a glimpse at the data set:

	dept_abbr	course	isbn	ucla_new	amaz_new	more	diff
1	Am Ind	C170	978-0803272620	27.67	27.95	Y	-0.28
2	Anthro	9	978-0030119194	40.59	31.14	Y	9.45
3	Anthro	135T	978-0300080643	31.68	32.00	Y	-0.32
⋮			⋮				⋮
72	Wom Std	M144	978-1570755637	23.76	18.72	Y	5.04
73	Wom Std	285	978-0822341147	27.70	18.22	Y	9.48

1. We want to test whether there is any difference in prices between the UCLA bookstore and Amazon. State the null and alternative hypotheses for such a test in words.

2. Let μ denote the true mean difference (UCLA price − Amazon price) in textbook prices. Restate the null and alternative hypotheses for our test using the symbol μ.

3. Using RStudio, I found the mean and standard deviation of the difference column to be:

$$\overline{x} = 12.76 \qquad s = 14.26.$$

   Determine a 95% confidence interval for the difference in prices between the UCLA bookstore and Amazon.

4. Based on this confidence interval is there strong evidence that the UCLA bookstore charges more than Amazon? At the $\alpha = 0.05$ level, would we reject the null hypothesis in favor of the alternative?

5. Does this mean students should buy their books from Amazon? State your opinion on this.

In RStudio we have two methods for conducting a test of significance on paired data. First, we can do a standard 1-sample t-test on the difference column. Second, we can use the paired data option on the two original columns in the data set, UCLA prices and Amazon prices. Notice, the outputs are identical. In either case, the p-value of the two-sided hypothesis test is $p = 6.928\text{e}-11$, which, remember, is shorthand for

$$p = 0.00000000006928,$$

which is very, very, very, very close to 0. At the $\alpha = .05$ level, we reject the null hypothesis in favor of the alternative because $p < \alpha$.

```
> t.test(books$diff, alternative="two.sided")

One Sample t-test

data:  books$diff
t = 7.6488, df = 72, p-value = 6.928e-11
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  9.435636 16.087652
sample estimates:
mean of x
 12.76164
```

```
> t.test(books$ucla_new,books$amaz_new,alternative="two.sided",paired=TRUE))

Paired t-test

data:  books$ucla_new and books$amaz_new
t = 7.6488, df = 72, p-value = 6.928e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  9.435636 16.087652
sample estimates:
mean of the differences
               12.76164
```