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

Worksheet: Matched Pairs or 2-sample t test?

1. **Friday the 13th.** In 1993 researchers published an article examining whether people tend to stay home on Friday the 13th. They counted the number of cars passing a certain junction on a highway in England for consecutive Fridays (the 6th and 13th) for five different time periods:

	year	month	cars6th	cars13th
1	1990	July	134012	132908
2	1991	September	133732	131843
3	1991	December	121139	118723
4	1992	March	124631	120249
5	1992	November	117584	117263

Based on the data, is there evidence that more people stay home on average on Friday the 13th? Here are two possible analyses of the data.

First analysis: Test $H_o : \mu_{6th} = \mu_{13th}$ vs. $H_a : \mu_{6th} > \mu_{13th}$

```
> t.test(data$cars6th,data$cars13th, alternative="greater")

Welch Two Sample t-test

data:  data$cars6th and data$cars13th
t = 0.42729, df = 7.9975, p-value = 0.3402
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -6779.355       Inf
sample estimates:
mean of x mean of y
 126219.6  124197.2
```

Second analysis: Matched Pairs Test $H_o : \mu_{\text{diff}} = 0$ vs. $\mu_{\text{diff}} > 0$.

```
> t.test(data$cars6th,data$cars13th, alternative="greater", paired=TRUE)

Paired t-test

data:  data$cars6th and data$cars13th
t = 2.9377, df = 4, p-value = 0.02124
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 554.778     Inf
sample estimates:
mean of the differences
                 2022.4
```

(a) Which of the tests is appropriate for these data? Explain.

(b) Using the test you selected, state your conclusion.

2. **Do actresses face age discrimination issues in Hollywood?** To help investigate this question we might compare the average age of Best Actor winners at the Academy Awards against the average age of Best Actress winners. Load the data `oscars.csv` (on our course resource page) into RStudio.

   A peek at the data:

date_of_award	name	award	age_at_award
16-May-29	Emil Jannings	best actor	44.80
16-May-29	Janet Gaynor	best actress	22.61
3-Apr-30	Warner Baxter	best actor	41.00
3-Apr-30	Mary Pickford	best actress	37.98
⋮	⋮	⋮	⋮
9-Feb-20	Joaquin Phoenix	best actor	45.30
9-Feb-20	Renee Zellweger	best actress	50.79
25-Apr-21	Anthony Hopkins	best actor	83.30
25-Apr-21	Frances McDormand	best actress	63.84

   Here are two possible analyses of the data.

   I. First analysis: 2-sample t Test $H_o : \mu_{\text{actor}} = \mu_{\text{actress}}$ vs. $H_a : \mu_{\text{actor}} \neq \mu_{\text{actress}}$ which, we can conduct in RStudio with the code:

```
t.test(age_at_award ~ award, data=oscars, alternative="two.sided")
```

   II. Second analysis: Matched Pairs Test $H_o : \mu_{\text{diff}} = 0$ vs. $\mu_{\text{diff}} \neq 0.$, which we conduct in RStudio with the code:

```
t.test(age_at_award ~ award, data=oscars, alternative="two.sided", paired=TRUE)
```

   (a) There have been $n = 94$ best actor awards given and $n = 94$ best actress awards given. Are these samples independent, or are they paired? Which test is appropriate for these data, 2-sample t or matched pairs?

   (b) In RStudio, run the test you selected as appropriate, and state your conclusion. Does there appear to be a statistically significant difference in the age of best actor winners vs the age of best actress winners?

   (c) Are the assumptions and conditions for inference met? To address this question, consider the sample size as well as whether the data is extremely skewed. To check skew, make a histogram of the 94 differences (age of Best Actor winner − age of Best Actress winner) in RStudio. You can use this code to find and plot these differences:

```
library(tidyverse)
oscars <- read.csv("https://mphitchman.com/stats/data/oscars.csv")
actor <- oscars %>% filter(award=="best actor")
actress <- oscars %>% filter(award=="best actress")
diff.age <- actor$age_at_award - actress$age_at_award
hist(diff.age)
```

```
> t.test(age_at_award~award,oscars,alternative="two.sided")

Welch Two Sample t-test

data:  age_at_award by award
t = 4.66, df = 176.57, p-value = 6.208e-06
alternative hypothesis: true difference in means between group best actor
and group best actress is not equal to 0
95 percent confidence interval:
  4.302924 10.624736
sample estimates:
  mean in group best actor mean in group best actress
                  44.88511                   37.42128
```

```
> t.test(age_at_award~award,oscars,alternative="two.sided",paired=TRUE)

Paired t-test

data:  age_at_award by award
t = 5.082, df = 93, p-value = 1.922e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  4.54735 10.38031
sample estimates:
mean of the differences
                7.46383
```