```
In [ ]:
In [ ]:
In [ ]:
In [ ]:
```

C2M2: Peer Reviewed Assignment

Outline:

The objectives for this assignment:

1. Utilize contrasts 2.
2. Understand power and why it's important to statistical conclusions.

General tips:

1. Read the questions carefully to understand what is being asked.
2. This work will be reviewed by another human, so make sure that you are clear and concise in what your explanations and answers.

Problem 1: Contrasts and Coupons

Consider a hardness testing machine that presses a rod with a pointed tip into a metal specimen with a known force. By measuring the depth of the depression caused by the tip, the hardness of the specimen is determined.

Suppose we wish to determine whether or not four different tips produce different readings on a hardness testing machine. The experimenter has decided to obtain four observations on Rockwell C-scale hardness for each tip. There is only one factor - tip type - and a completely randomized single-factor design would consist of randomly assigning each one of the 4×4=16 runs to an experimental unit, that is, a metal coupon, and observing the hardness reading that results. Thus, 16 different metal test coupons would be required in this experiment, one for each run in the design.

In [1]:

```
tip    <- factor(rep(1:4, each = 4))
coupon <- factor(rep(1:4, times = 4))
y <- c(9.3, 9.4, 9.6, 10,
       9.4, 9.3, 9.8, 9.9,
       9.2, 9.4, 9.5, 9.7,
       9.7, 9.6, 10, 10.2)
hardness <- data.frame(y, tip, coupon)
hardness
```

y	tip	coupon
9.3	1	1
9.4	1	2
9.6	1	3
10.0	1	4
9.4	2	1
9.3	2	2
9.8	2	3
9.9	2	4
9.2	3	1
9.4	3	2
9.5	3	3
9.7	3	4
9.7	4	1
9.6	4	2
10.0	4	3
10.2	4	4

1. (a) Visualize the Groups

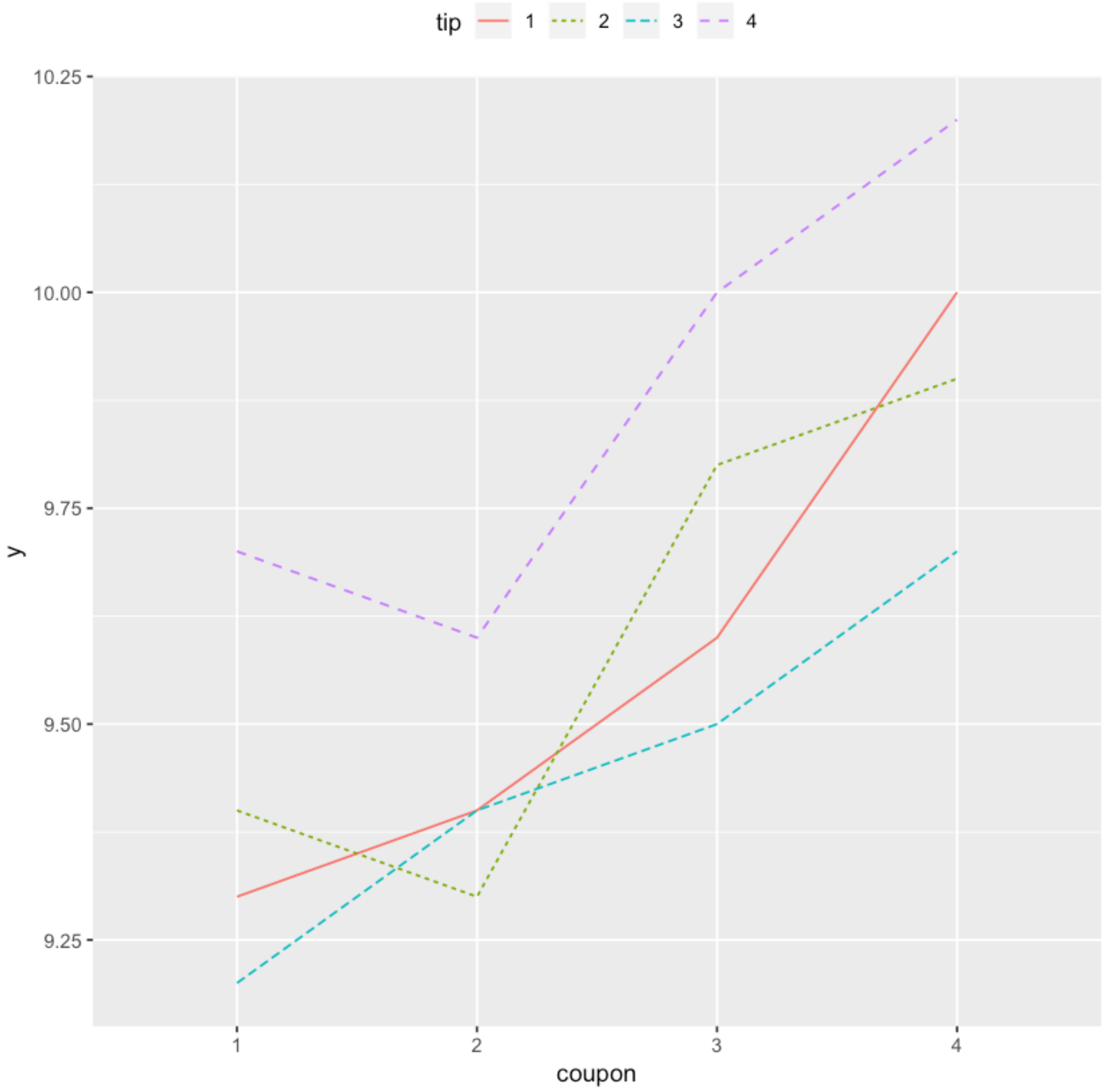
Before we start throwing math at anything, let's visualize our data to get an idea of what to expect from the eventual results.

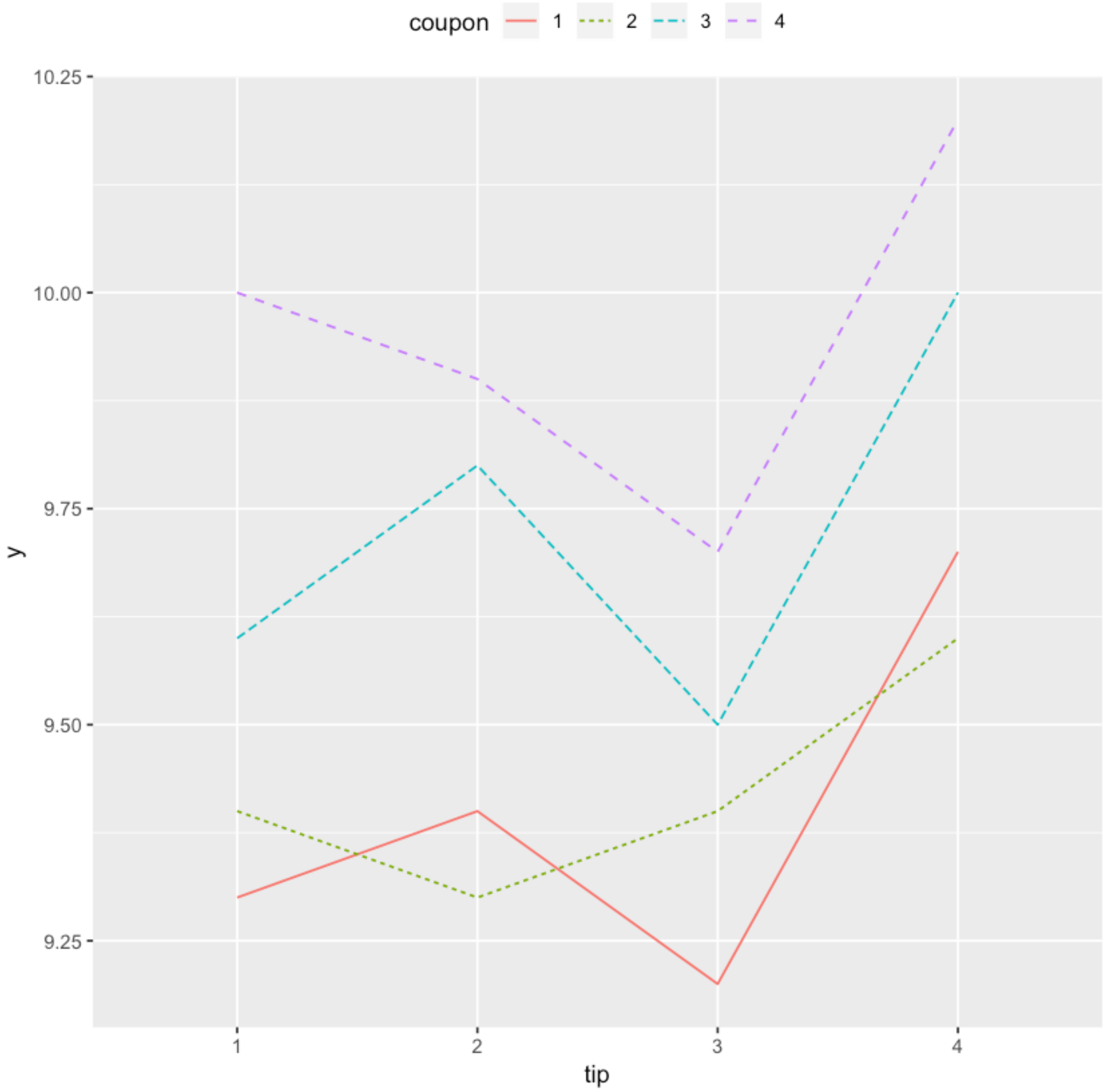
Construct interaction plots for `tip` and `coupon` using ggplot(). Be sure to explain what you can from the plots.

In [2]:

```
library(ggplot2)
p = ggplot(hardness, aes(x = coupon, y = y, group = tip, linetype = tip))
p=p + geom_line(aes(color = tip)) + theme(legend.position = "top", legend.direction = "horizontal")
p
p = ggplot(hardness, aes(x = tip, y = y, group = coupon, linetype = coupon))
p=p + geom_line(aes(color = coupon)) + theme(legend.position = "top", legend.direction = "horizontal")
p
```

```
Registered S3 methods overwritten by 'ggplot2':
  method         from
  [.quosures     rlang
  c.quosures     rlang
  print.quosures rlang
```





These plots are somewhat ambiguous, but neither provides strong evidence of interactions. Let's focus the interpretation on the first plot. As we move from coupon 1 to 4, the distance (in units of the response) between tip 2 and tip 4 stays relatively constant, suggesting no interaction between these levels.

However, as we move from coupon 1 to 4, the distance (in units of the response) between tip 2 and tip 3 swaps, suggesting an interaction between these levels. More specifically, for coupon 1, tip 2 makes a deeper impression than tip 3; for coupon 2, tip 3 makes a deeper impression than tip 2, i.e., the effect is reversed. Then, as we move to coupon 3 and 4, they reverse again. All of this suggests an interaction. These plots show similar interactions for other levels.

Importantly, though, notice that there is no *replication* within blocks: for example, tip 1 is tested on coupon 1 only once. So, with 16 total data points, we won't have enough data to conduct any statistical tests of we include interaction terms. And further, the plots don't provide strong evidence about what would happen on average if we had replication. So, we should draw conclusions with much caution!

1. (b) Interactions

Should we test for interactions between `tip` and `coupon` ? Maybe there is an interaction between the different metals that goes beyond our current scientific understanding!

Fit a linear model to the data with predictors `tip` and `coupon` , and an interaction between the two. Display the summary and explain why (or why not) an interaction term makes sense for this data.

In [3]:

```
lmodInt = lm(y ~ coupon + tip + tip:coupon, data = hardness)
summary(lmodInt)
```

```
Call:
lm(formula = y ~ coupon + tip + tip:coupon, data = hardness)

Residuals:
ALL 16 residuals are 0: no residual degrees of freedom!

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.300e+00         NA      NA       NA
coupon2       1.000e-01         NA      NA       NA
coupon3       3.000e-01         NA      NA       NA
coupon4       7.000e-01         NA      NA       NA
tip2          1.000e-01         NA      NA       NA
tip3         -1.000e-01         NA      NA       NA
tip4          4.000e-01         NA      NA       NA
coupon2:tip2 -2.000e-01         NA      NA       NA
coupon3:tip2  1.000e-01         NA      NA       NA
coupon4:tip2 -2.000e-01         NA      NA       NA
coupon2:tip3  1.000e-01         NA      NA       NA
coupon3:tip3  3.351e-15         NA      NA       NA
coupon4:tip3 -2.000e-01         NA      NA       NA
coupon2:tip4 -2.000e-01         NA      NA       NA
coupon3:tip4  3.365e-15         NA      NA       NA
coupon4:tip4 -2.000e-01         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:    NaN 
F-statistic:   NaN on 15 and 0 DF,  p-value: NA
```

Notice that this model, which includes an interaction term, does not provide any statistical information (e.g., standard errors, t-tests). That is because, without replication, there are as many parameters as data points (16), and thus, only enough data points to estimate all the necessary parameters, and none left over to estimate standard errors. So, it does not make sense to include interaction terms for this experiment.

1. (c) Contrasts

Let's take a look at the use of contrasts. Recall that a contrast takes the form

$$\sum_{i=1}^{t} c_i \mu_i = 0,$$

where $\mathbf{c} = (c_1, \ldots, c_t)$ is a constant vector and $\mu = (\mu_1, \ldots, \mu_t)$ is a parameter vector (e.g., μ_1 is the mean of the i^{th} group).

We can note that $\mathbf{c} = (1, -1, 0, 0)$ corresponds to the null hypothesis $H_0 : \mu_2 - \mu_1 = 0$, where μ_1 is the mean associated with tip1 and μ_2 is the mean associated with tip2. The code below tests this hypothesis.

Repeat this test for the hypothesis $H_0 : \mu_4 - \mu_3 = 0$. Interpret the results. What are your conclusions?

In [23]:

```
library(multcomp)
lmod = lm(y~tip+coupon, data=hardness)
summary(lmod)
fit.gh2 = glht(lmod, linfct = mcp(tip = c(-1,1,0,0)))
summary(fit.gh2)

#estimate of mu_2 - mu_1
with(hardness, sum(y[tip == 2])/length(y[tip == 2]) -
     sum(y[tip == 1])/length(y[tip == 1]))
```

```
Call:
lm(formula = y ~ tip + coupon, data = hardness)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.10000 -0.05625 -0.01250  0.03125  0.15000 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.35000    0.06236 149.934  < 2e-16 ***
tip2         0.02500    0.06667   0.375 0.716345    
tip3        -0.12500    0.06667  -1.875 0.093550 .  
tip4         0.30000    0.06667   4.500 0.001489 ** 
coupon2      0.02500    0.06667   0.375 0.716345    
coupon3      0.32500    0.06667   4.875 0.000877 ***
coupon4      0.55000    0.06667   8.250 1.73e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09428 on 9 degrees of freedom
Multiple R-squared:  0.938,	Adjusted R-squared:  0.8966 
F-statistic: 22.69 on 6 and 9 DF,  p-value: 5.933e-05
```

```
	 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts


Fit: lm(formula = y ~ tip + coupon, data = hardness)

Linear Hypotheses:
       Estimate Std. Error t value Pr(>|t|)
1 == 0  0.02500    0.06667   0.375    0.716
(Adjusted p values reported -- single-step method)
```

0.0250000000000021

In [24]:

```
library(multcomp)
fit.gh2 = glht(lmod, linfct = mcp(tip = c(0,0,-1,1)))
summary(fit.gh2)

#estimate of mu_4 - mu_3
with(hardness, sum(y[tip == 4])/length(y[tip == 4]) -
     sum(y[tip == 3])/length(y[tip == 3]))
```

```
	 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts


Fit: lm(formula = y ~ tip + coupon, data = hardness)

Linear Hypotheses:
       Estimate Std. Error t value Pr(>|t|)    
1 == 0  0.42500    0.06667   6.375 0.000129 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

0.425000000000001

From the summary of `fit.gh2` , we see that our estimate for $\mu_4 - \mu_3$ is 0.425, and that the t-test associated with this difference of parameters is statistically significant. So, we have statistical evidence that there is a different between $\mu_4 - \mu_3$.

1. (d) All Pairwise Comparisons

What if we want to test all possible pairwise comparisons between treatments. This can be done by setting the treatment factor (`tip`) to "Tukey". Notice that the p-values are adjusted (because we are conducting multiple hypotheses!).

Perform all possible Tukey Pairwise tests. What are your conclusions?

In [25]:

```
fit.gh = glht(lmod, linfct = mcp(tip = c(0,-1,1,0)))
summary(fit.gh)

fit.gh = glht(lmod, linfct = mcp(tip = "Tukey"))
summary(fit.gh)
```

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts


Fit: lm(formula = y ~ tip + coupon, data = hardness)

Linear Hypotheses:
       Estimate Std. Error t value Pr(>|t|)  
1 == 0 -0.15000    0.06667   -2.25    0.051 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = y ~ tip + coupon, data = hardness)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)    
2 - 1 == 0  0.02500    0.06667   0.375  0.98091    
3 - 1 == 0 -0.12500    0.06667  -1.875  0.30244    
4 - 1 == 0  0.30000    0.06667   4.500  0.00655 ** 
3 - 2 == 0 -0.15000    0.06667  -2.250  0.18177    
4 - 2 == 0  0.27500    0.06667   4.125  0.01112 *  
4 - 3 == 0  0.42500    0.06667   6.375  < 0.001 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

In the above code, we've conducted 6 tests simultaneously. So, we've adjusted our type I error rate and can draw some conclusions about differences. We see that, at significance levele $\alpha = 0.05$, there is statistical evidence that there are true differences between tip 4 and tip 1, tip 4 and tip 2, and tip 4 and tip 3.

Problem 2: Ethics in my Math Class!

In your own words, answer the following questions:

- What is power, in the statistical context?
- Why is power important?
- What are potential ethical/societal consequences of ignoring/not including power calculations in statistical analyses?

- Power, in the statistical context, is the probability of a statistical test to find a result if there is truly a result to find. For example, in the experiment in problem 1, the power of the hypothesis test concerning $H_0 : \mu_4 - \mu_3$ has to do with the probability of finding a true difference between μ_4 and μ_3 with the test given that there really is a true difference.
- Power is very important! For a fixed α, a higher power test means that there's a lower probability of making a type II error, i.e., an error of not finding a result when there is one...something that we want to avoid!
- If a test has a low power, then it is unlikely to find a a relationship when there actually is one. This can have ethical consequences. A poorly designed, low-powered test used to find a vaccine that could save many lives, for example, has the consequence of making it likely that a truly effective vaccine won't be discovered to be effective by the test. This can clearly cost lives.

Problem 3: Post-Hoc Tests

There's so many different post-hoc tests (e.g., Tukey, Bonferroni)! Let's try to understand them better. Answer the following questsions in the markdown cell:

- Why are there multiple post-hoc tests?
- When would we choose to use Tukey's Method over the Bonferroni correction, and vice versa?
- Do some outside research on other post-hoc tests. Explain what the method is and when it would be used.

- There are many different types of multiple comparison/post-hoc testing methods because there's no single "correct" way to adjust for multiple comparisons. Each adjusts the overall level of type I error in their own way (some more conservatively than others, which impacts power!).
- A conservative method, like the Bonferroni method, will guard against type I errors at the expense of type II errors. Less conservative methods, like the Tukey method will err on the side of larger type I errors.
- Answers will vary!

```
In [ ]:
```