

Problem 1.

In your own words describe how RM ANOVA can be used to test for differences between groups, differences between repeated observations, and differences in observations as a function of group membership. These tests have specific names we discussed in lecture.

(a) Groups Hypothesis (Test of Levels)

- (i) Do groups score similarly on the collected set of measures?
- (ii) This test checks if the mean of the groups over all measurement times is the same. This does not mean that all the groups performed the same at each time point, just that their average performance is the same over all time points. Mathematically, this is simple measuring the relative contributions of between-group and within-group contributions to the total sum of squared errors. If the group ‘levels’ are significantly different, then the equal levels null hypothesis is rejected.

(b) Flatness Hypothesis (Test of Flatness)

- (i) Is the dependent variable similar across assessments?
- (ii) This test checks if the difference in the response between assessments (i.e., the slope) is nonzero. The flatness null hypothesis is that the slope between all assessments is zero, and that the results are flat. This is checked for each group, and is effectively a within-subjects test. If the slope is statistically significantly nonzero, then there is a within groups main effect of time.

(c) Parallelism Hypothesis (Test of Parallelism)

- (i) Do different groups have parallel profiles?
- (ii) This test checks if the slope between each set of assessments is the same across all groups. This is equivalent to a one-way ANOVA on the slopes, we are checking that there is no interaction of the within-subjects factor with the between-subjects factor.

Problem 2.

Why do we say that in RM ANOVA time is treated as a categorical variable?

RM ANOVA treats time as a categorical factor. Time is considered a set of discrete, fixed conditions in which each subject contributes one outcome variable per time condition. The ANOVA model for describing mean change over time uses as many parameters as there are discrete occasions. This is also known as a saturated means model—the model for time is saturated by using all possible degrees of freedom for differences across conditions of the time variable. The goal of the ANOVA model is not to predict or summarize the pattern of means, but simply to reproduce the observed mean per occasion using fixed effects equal to the number of occasions minus 1 (given that the fixed intercept is already included).

Problem 3.

What is the difference between polynomial contrasts and comparisons of means between different observation periods?

Polynomial contrast tests can be used to test which level of polynomial (linear, quadratic, cubic) explains the relation under study.

Problem 4.

Researchers were interested in how confidence in students fluctuates from the end of sophomore year through the end of senior year in college. Further, researchers were interested in whether students attended a private (i.e., Harvard, Stanford, and Yale) or public (i.e., UMass—Boston, San Francisco State University, and Southern Connecticut State University) university would influence end-of-year student confidence ratings, and possible trajectories.

Use the data set `confidence.csv` to answer the following questions. For your information, `t1-t3` indicate time of observation 1-3; i.e., end of sophomore year, end of junior year, and end of senior year, respectively. The variable `public` indicates whether the student attended a public (`public = 1`) or private (`public = 0`) university.

1. Convert the wide format data set to a long format data set. Remember that you will need to create an ID variable, and an Observation/Time variable. Show syntax and the header and footer of the long format data set.

```

1 import pandas as pd
2
3 df = pd.read_csv('confidence.csv')
4
5 # Convert to long format
6 r = pd.melt(df.reset_index(),
7             value_vars=['t1', 't2', 't3'],
8             id_vars=['index', 'public'],
9             var_name='Time', value_name='Observation')
10
11 # Some basic sorting
12 r = r.sort(['index', 'Time']).reset_index(drop=True)
13
14 # Rename first column to ID
15 cols = r.columns.tolist()
16 cols[0] = 'ID'
17 r.columns = cols
18
19 # Print head and tail of new dataframe
20 print(r.head())
21 print(r.tail())

```

OUTPUT

	ID	public	Time	Observation
0	0	0	t1	4
1	0	0	t2	6
2	0	0	t3	6

```

5 3 1 0 t1 3
6 4 1 0 t2 5
7
8      ID public Time Observation
9 295 98 1 t2 4
10 296 98 1 t3 4
11 297 99 1 t1 4
12 298 99 1 t2 5
13 299 99 1 t3 4

```

2. Test whether public and private universities differed in their confidence scores. Conduct any pairwise comparison necessary. Report your conclusions.

```
1 print(anova_lm(ols("Observation ~ C(public)", df).fit(), typ=2))
```

OUTPUT

```

1      sum_sq    df      F      PR(>F)
2 C(public)  48.803333    1  32.419994  2.975196e-08
3 Residual  448.593333  298      NaN      NaN

```

A one-way between subjects ANOVA was conducted to compare the effect of different types of universities on student confidence for students at public and private universities. There was a significant effect of amount of sugar on words remembered at the $p < .001$ level for the two conditions [$F(1, 298) = 32.4, p = 2.98e-08$]. Comparisons of the mean indicated that the mean confidence scores for the private university students ($M = 3.77, SD = 1.30$) was significantly different than the public university students ($M = 2.96, SD = 1.15$).

3. Test whether confidence differs over time. Do this by comparing all observations to each other. Also do this by testing the maximum allowable number of polynomial contrasts. Report your conclusions for both tests.

```

1 from patsy import dmatrix
2
3 # This is equivalent to R's contr.poly
4 p = dmatrix("C(df.Time, Poly())", df)
5 poly = pd.DataFrame(p, columns=['Intercept', 'Linear', 'Quadratic'])
6 df = pd.concat((df, poly), axis=1)
7 print(anova_lm(ols("Observation ~ (Time)", data=df).fit()))
8 print(anova_lm(ols("Observation ~ Linear + Quadratic", data=df).fit(), typ=2))

```

OUTPUT

```

1      df    sum_sq  mean_sq      F      PR(>F)
2 C(Time)    2  13.406667  6.703333  4.113494  0.017291
3 Residual  297  483.990000  1.629596      NaN      NaN
4
5      sum_sq    df      F      PR(>F)
6 Linear    7.605000    1  4.666801  0.031551
7 Quadratic    5.801667    1  3.560187  0.060156
8 Residual  483.990000  297      NaN      NaN

```

4. Test for a possible interaction between variables tested in question 2 (university type) and 3 (mean differences between observations and significant trends in confidence). Conduct any

follow-up analyses necessary if there is a significant interactions. Report your conclusions for all tests.

```

1 print(anova_lm(ols("Observation ~ C(Time)*C(public)", data=df).fit(), typ=2))
2
3 # Test for simple effects
4 print(anova_lm(ols("Observation ~ C(Time)", data=df.query('public == 0')).fit(), typ=2))
5 print(anova_lm(ols("Observation ~ C(Time)", data=df.query('public == 1')).fit(), typ=2))
6 print(anova_lm(ols("Observation ~ C(public)", data=df.query('Time == 1')).fit(), typ=2))
7 print(anova_lm(ols("Observation ~ C(public)", data=df.query('Time == 2')).fit(), typ=2))
8 print(anova_lm(ols("Observation ~ C(public)", data=df.query('Time == 3')).fit(), typ=2))

```

OUTPUT

```

1              sum_sq    df          F          PR(>F)
2 C(Time)          13.406667    2    4.995134  7.356729e-03
3 C(public)        48.803333    1   36.366858  4.897455e-09
4 C(Time):C(public) 40.646667    2   15.144371  5.496327e-07
5 Residual        394.540000  294          NaN          NaN
6
7              sum_sq    df          F          PR(>F)
8 C(Time)        37.213333    2   12.685187  0.000008
9 Residual      215.620000  147          NaN          NaN
10
11             sum_sq    df          F          PR(>F)
12 C(Time)        16.84    2    6.91784  0.001345
13 Residual      178.92  147          NaN          NaN
14
15             sum_sq    df          F          PR(>F)
16 C(public)        0.81    1    1.107113  0.295298
17 Residual        71.70   98          NaN          NaN
18
19             sum_sq    df          F          PR(>F)
20 C(public)         4.00    1    2.911468  0.091119
21 Residual       134.64   98          NaN          NaN
22
23             sum_sq    df          F          PR(>F)
24 C(public)        84.64    1   44.073964  1.766331e-09
25 Residual       188.20   98          NaN          NaN

```

Source	SS	df	F	$PR(> F)$
Time	13.40	2	5.00	< .001
Public	48.80	1	36.37	< .001
Interaction	40.64	2	15.14	< .001
Residual	394.54	294		

Table 1: Factorial ANOVA Results

5. Create a plot of the data that represents the trajectory for public and private students. Make sure the trajectories have some indication of variability around the average trajectory. Sufficiently label and describe your figure.

Source	<i>SS</i>	<i>df</i>	<i>F</i>	<i>PR(> F)</i>
<i>Public</i>				
Private	37.21	2	12.7	< .001
Public	16.84	2	6.9	.001
<i>Time</i>				
1	0.81	1	1.1	.001
2	4.00	1	2.9	.295
3	84.64	1	44.1	< .001

Table 2: Results of Simple Effects Analysis