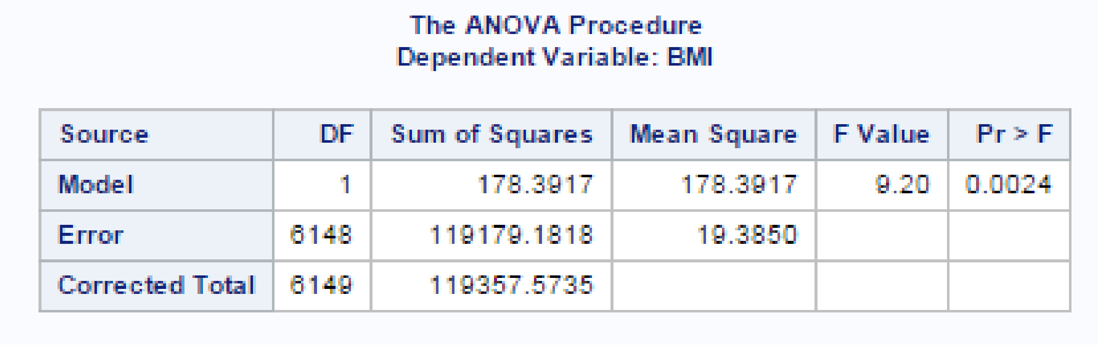
Data analysis tools

Assignment 1

ANOVA

I chose addhealth as my date set. The relationship between gender (categorical, 2 levels) and BMI (quantitative), and between grade (categorical, 6 levels) and BMI are analyzed respectively. For all analysis, α = 0.05.

First, the relationship between gender and BMI is analyzed by ANOVA.

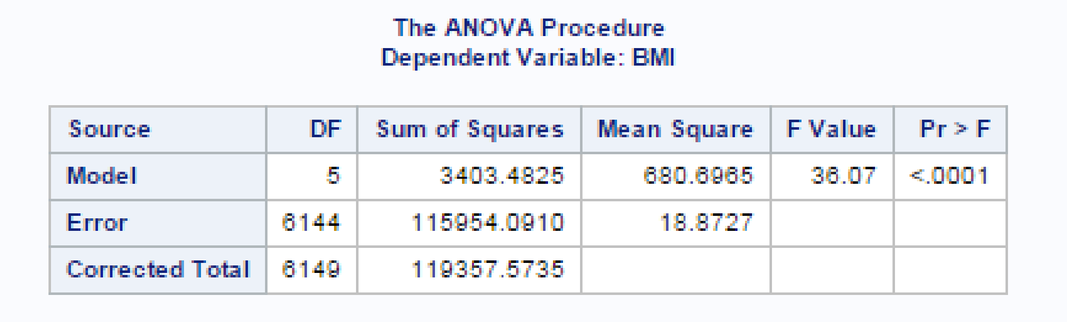


The ANOVA Procedure
Dependent Variable: BMI

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	178.3917	178.3917	9.20	0.0024
Error	6148	119179.1818	19.3850		
Corrected Total	6149	119357.5735			

According to the output, p-value is less than 0.05. So the null hypothesis can be rejected. We think there is significant difference between the BMI of male and the BMI of female.

Second, the relationship between grade and BMI is analyzed by ANOVA.



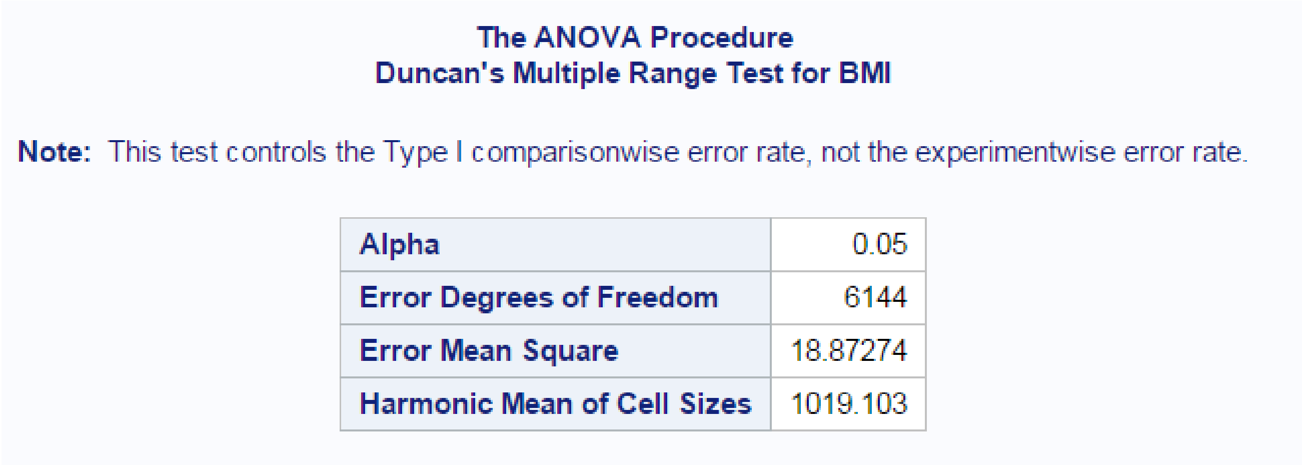
The ANOVA Procedure
Dependent Variable: BMI

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	3403.4825	680.6965	36.07	<.0001
Error	6144	115954.0910	18.8727		
Corrected Total	6149	119357.5735			

According to the output, p-value is less than 0.05. So the null hypothesis can be rejected. We think the BMIs of respondents in different grades are not all equal.

To find out which grade is different from others, two host hoc tests are conducted.

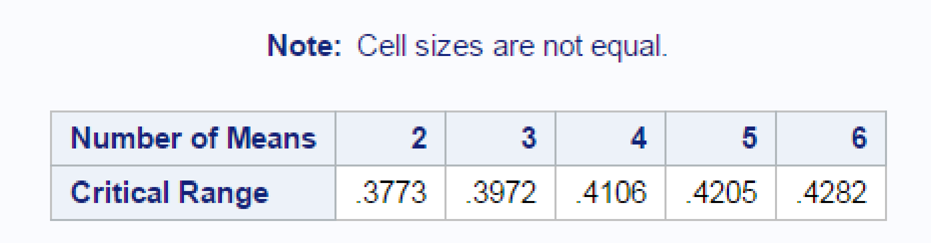
The first is **Duncan's Multiple Range test**.



The ANOVA Procedure
Duncan's Multiple Range Test for BMI

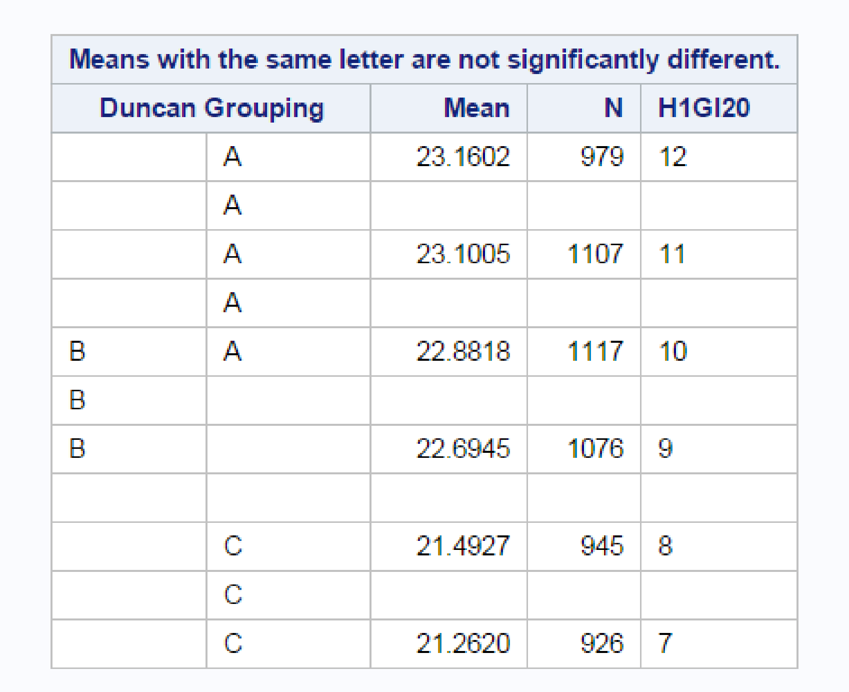
Note: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	6144
Error Mean Square	18.87274
Harmonic Mean of Cell Sizes	1019.103



Note: Cell sizes are not equal.

Number of Means	2	3	4	5	6
Critical Range	.3773	.3972	.4106	.4205	.4282

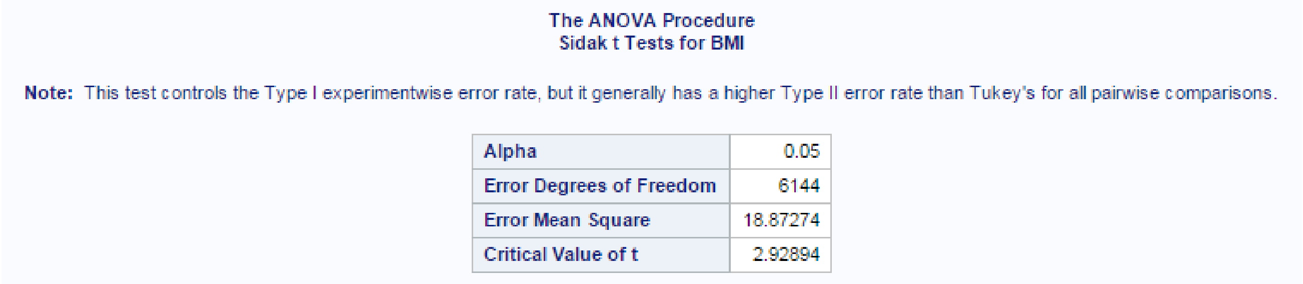


Means with the same letter are not significantly different.

Duncan Grouping		Mean	N	H1GI20
	A	23.1602	979	12
	A			
	A	23.1005	1107	11
	A			
B	A	22.8818	1117	10
B				
B		22.6945	1076	9
	C	21.4927	945	8
	C			
	C	21.2620	926	7

According to the output, grade 7 and 8 are similar to each other. Grade 9 and 10 are similar to each other. Grade 11 and 12 are similar to each other. In addition to this, there is significant difference between other pairs.

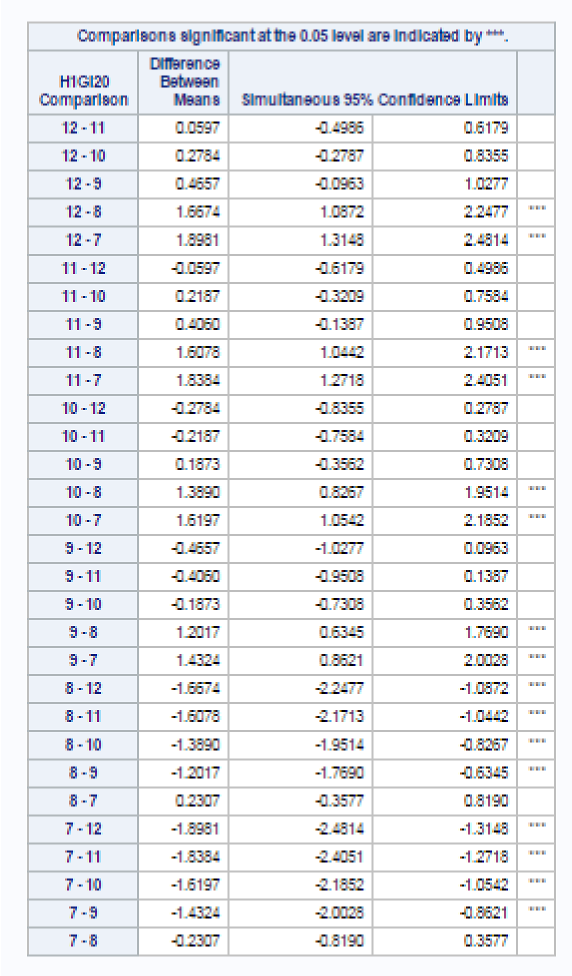
The second one is **Sidak t test**.



The ANOVA Procedure
Sidak t Tests for BMI

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than Tukey's for all pairwise comparisons.

Alpha	0.05
Error Degrees of Freedom	6144
Error Mean Square	18.87274
Critical Value of t	2.92894

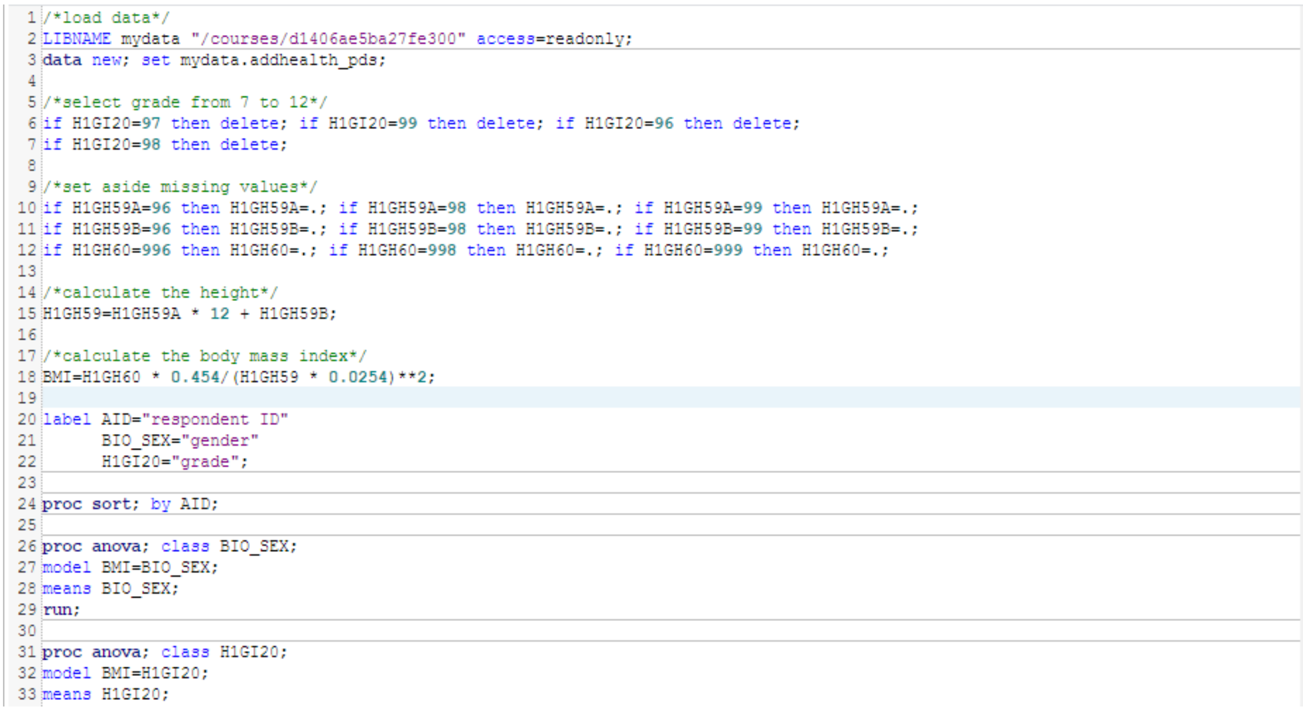


Comparisons significant at the 0.05 level are indicated by ***.

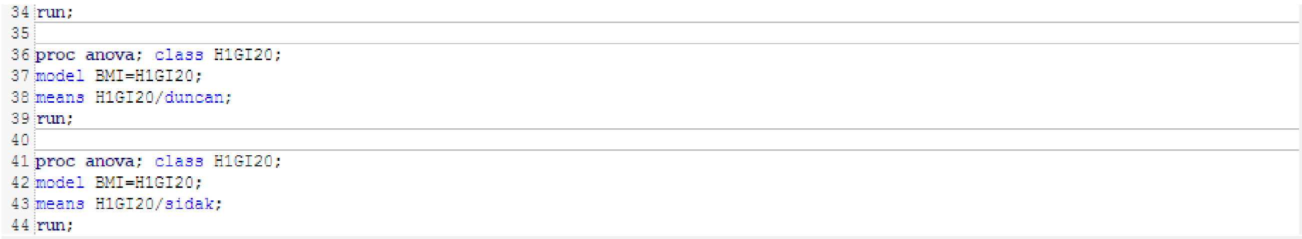
H1GI20 Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
12 - 11	0.0597	-0.4986	0.6179	
12 - 10	0.2784	-0.2787	0.8355	
12 - 9	0.4657	-0.0963	1.0277	
12 - 8	1.6674	1.0872	2.2477	***
12 - 7	1.8981	1.3148	2.4814	***
11 - 12	-0.0597	-0.6179	0.4986	
11 - 10	0.2187	-0.3209	0.7584	
11 - 9	0.4060	-0.1387	0.9508	
11 - 8	1.6078	1.0442	2.1713	***
11 - 7	1.8384	1.2718	2.4051	***
10 - 12	-0.2784	-0.8355	0.2787	
10 - 11	-0.2187	-0.7584	0.3209	
10 - 9	0.1873	-0.3562	0.7308	
10 - 8	1.3890	0.8267	1.9514	***
10 - 7	1.6197	1.0542	2.1852	***
9 - 12	-0.4657	-1.0277	0.0963	
9 - 11	-0.4060	-0.9508	0.1387	
9 - 10	-0.1873	-0.7308	0.3562	
9 - 8	1.2017	0.6345	1.7690	***
9 - 7	1.4324	0.8621	2.0028	***
8 - 12	-1.6674	-2.2477	-1.0872	***
8 - 11	-1.6078	-2.1713	-1.0442	***
8 - 10	-1.3890	-1.9514	-0.8267	***
8 - 9	-1.2017	-1.7690	-0.6345	***
8 - 7	0.2307	-0.3577	0.8190	
7 - 12	-1.8981	-2.4814	-1.3148	***
7 - 11	-1.8384	-2.4051	-1.2718	***
7 - 10	-1.6197	-2.1852	-1.0542	***
7 - 9	-1.4324	-2.0028	-0.8621	***
7 - 8	-0.2307	-0.8190	0.3577	

From the results, we can see there is significant difference between grade 8 and 12, 7 and 12, 8 and 11, 7 and 11, 8 and 10, 7 and 10, 9 and 9, 9 and 7 respectively. The results are similar to Duncan's Multiple Range test.

My code:



```
/*load data*/
LIBNAME mydata "/courses/d1406ae5ba27fe300" access=readonly;
data new; set mydata.addhealth_pds;

/*select grade from 7 to 12*/
if H1GI20=97 then delete; if H1GI20=99 then delete; if H1GI20=96 then delete;
if H1GI20=98 then delete;

/*set aside missing values*/
if H1GH59A=96 then H1GH59A=.; if H1GH59A=98 then H1GH59A=.; if H1GH59A=99 then H1GH59A=.;
if H1GH59B=96 then H1GH59B=.; if H1GH59B=98 then H1GH59B=.; if H1GH59B=99 then H1GH59B=.;
if H1GH60=996 then H1GH60=.; if H1GH60=998 then H1GH60=.; if H1GH60=999 then H1GH60=.;

/*calculate the height*/
H1GH59=H1GH59A * 12 + H1GH59B;

/*calculate the body mass index*/
BMI=H1GH60 * 0.454/(H1GH59 * 0.0254)**2;

label AID="respondent ID"
      BIO_SEX="gender"
      H1GI20="grade";

proc sort; by AID;

proc anova; class BIO_SEX;
model BMI=BIO_SEX;
means BIO_SEX;
run;

proc anova; class H1GI20;
model BMI=H1GI20;
means H1GI20;
```



```
run;

proc anova; class H1GI20;
model BMI=H1GI20;
means H1GI20/duncan;
run;

proc anova; class H1GI20;
model BMI=H1GI20;
means H1GI20/sidak;
run;
```