Math 1070-Chapter 27 Notes

One-Way Analysis of Variance(ANOVA):

Comparing Several Means by using ANOVA F test

1 Comparing Ch 21. with Ch 27.

	Ch 21	Ch 27
The number of comparing populations(or groups)	2	3 (or more)
H_0	$\mu_1 = \mu_2$	$\mu_1 = \mu_2 = \mu_3$
H_a	$\mu_1 \neq \mu_2$ $\mu_1 < \mu_2$ $\mu_1 > \mu_2$	H_0 is not true. = Not all of μ_1, μ_2, and μ_3 are equal.
Significance Test	t test	ANOVA F test
Test statistic	$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$F = \frac{\text{MSG}}{\text{MSE}}$

2 Example

Table 27.1 gives length measurements (in millimeters) for samples of three varieties of Heliconia, each fertilized by a different species of hummingbird. Do the three varieties display distinct distributions of length? In particular, are the mean lengths of their flowers different at $\alpha = 0.05$?

TABLE 27.1 FLOWER LENGTHS (MILLIMETERS) FOR THREE *HELICONIA* VARIETIES

H. BIHAI							
47.12	46.75	46.81	47.12	46.67	47.43	46.44	46.64
48.07	48.34	48.15	50.26	50.12	46.34	46.94	48.36
***H. CARIBAEA* RED**							
41.90	42.01	41.93	43.09	41.47	41.69	39.78	40.57
39.63	42.18	40.66	37.87	39.16	37.40	38.20	38.07
38.10	37.97	38.79	38.23	38.87	37.78	38.01	
***H. CARIBAEA* YELLOW**							
36.78	37.02	36.52	36.11	36.03	35.45	38.13	37.10
35.17	36.82	36.66	35.68	36.03	34.57	34.63	

2.1 Preliminary works

2.1.1 The summary of Table 27.1

Sample	Variety	Sample size	Mean length	Standard deviation
1	*bihai*	16	47.60	1.213
2	red	23	39.71	1.799
3	yellow	15	36.18	0.975

2.1.2 Conditions for using ANOVA inference

CONDITIONS FOR ANOVA INFERENCE

- We have **I independent SRSs,** one from each of I populations. We measure the same response variable for each sample.
- The ith population has a **Normal distribution** with unknown mean μ_i. One-way ANOVA tests the null hypothesis that all the population means are the same.
- All the populations have the **same standard deviation** σ, whose value is unknown.

In the second condition, ANOVA becomes safe if the sample sizes is large even though there is not normality assumption , because of the CLT. The third condition is annoying: It is not easy to check the condition that the populations have equal standard deviations. Here is a rule of thumb that is safe in almost all situations.

$$\frac{\text{largest sample sd}}{\text{smallest sample sd}} < 2$$

$\Longrightarrow$ We can safely use ANOVA F test.

Let's check if standard deviations above satisfy rule of thumb. If yes then we can safely use ANOVA.

2.1.3 F distribution and F Table

Table entry for p is the critical value F^* with probability p lying to its right.

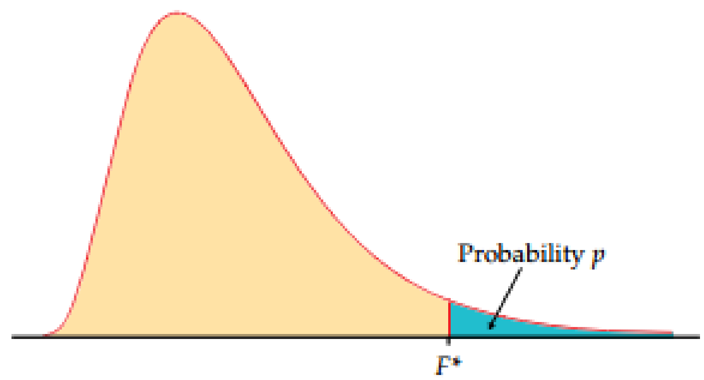

TABLE E

F critical values

Degrees of freedom in the denominator	p	Degrees of freedom in the numerator: 1	2	3	4	5	6	7	8	9
1	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
	.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
	.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28
	.010	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
	.001	405284	500000	540379	562500	576405	585937	592873	598144	602284
2	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
	.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
	.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39
3	.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
	.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
	.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
	.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86
4	.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
	.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90
	.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
	.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47
5	.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
	.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
	.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
	.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24

Our brief notation will be F(df1, df2) for the F distribution with df1 degrees of freedom in the numerator and df2 in the denominator.

2.2 The ANOVA F test

2.2.1 Step 1) Make hypotheses

$H_0 : \mu_1 = \mu_2 = \mu_3$ vs. H_a : Not all of μ_1, μ_2, and μ_3 are equal

2.2.2 Step2) Critical value of F

- degree of freedoms in F(df1, df2)
  - df1: degrees of freedom in the numerator = $I - 1$,
  - df2: degrees of freedom in the denominator = $N - I$
  - ,where I is the number of populations and N is the total number of observations.
  - In the example above, $I = 3$, $N = 16 + 23 + 15 = 54$.

    Therefore, df1 = 3-1 = 2, and df2 = 54-3 = 51 .
- Since F Table does not have df2 = 51, we take df2 = 50 close to 51 and $\alpha = 0.05$, $F^* = 3.18$. The right area of 3.18 is the rejection region.

	.100	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79
	.050	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
40	.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45
	.010	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
	.001	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02
	.100	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76
	.050	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07
50	.025	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38
	.010	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78
	.001	12.22	7.96	6.34	5.46	4.90	4.51	4.22	4.00	3.82
	.100	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74
	.050	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
60	.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33
	.010	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
	.001	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69

2.2.3 Step 3) ANOVA F statistic

THE ANOVA *F* TEST

Draw an independent SRS from each of I Normal populations that have a common standard deviation but may have different means. The sample from the ith population has size n_i, sample mean $\overline{x}_i$, and sample standard deviation s_i.

To test the null hypothesis that all I populations have the same mean against the alternative hypothesis that not all the means are equal, calculate the **ANOVA *F* statistic**

$$F = \frac{\text{MSG}}{\text{MSE}}$$

The numerator of F is the **mean square for groups**

$$\text{MSG} = \frac{n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2 + \cdots + n_I(\overline{x}_I - \overline{x})^2}{I - 1}$$

The denominator of F is the **mean square for error**

$$\text{MSE} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_I - 1)s_I^2}{N - I}$$

When H_0 is true, F has the ***F* distribution** with $I - 1$ and $N - I$ degrees of freedom.

Overall mean $\overline{x}$ is the mean of all N observations together. You can find $\overline{x}$ from the I sample means by

$$\overline{x} = \frac{\text{sum of all observations}}{N} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2 + \cdots + n_I\overline{x}_I}{N}$$

- The basic ANOVA table

Source of variation	df	SS	MS	*F* statistic
Variation among samples	2	1082.87	MSG = 541.44	259.12
Variation within samples	51	106.57	MSE = 2.09	

- How to calculate

Sample	Variety	Sample size	Mean length	Standard deviation
1	*bihai*	16	47.60	1.213
2	red	23	39.71	1.799
3	yellow	15	36.18	0.975

  - The overall mean $\overline{x}$

    $$\overline{x} = \frac{n_1\overline{x}_1+n_2\overline{x}_2+n_3\overline{x}_3}{N} = \frac{(16)(47.60)+(23)(39.71)+(15)(36.18)}{54} = \frac{2217.63}{54} = 41.067$$

  - The mean squre for group (MSG)

    $$\text{MSG} = \frac{n_1(\overline{x}_1-\overline{x})^2+n_2(\overline{x}_2-\overline{x})^2+n_3(\overline{x}_3-\overline{x})^2}{I-1}$$
    $$= \frac{(16)(47.60-41.067)^2+(23)(39.71-41.067)^2+(15)(36.18-41.067)^2}{3-1}$$
    $$= \frac{1083.476286}{2} = 541.74$$

  - The mean square for error (MSE)

    $$\text{MSE} = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2+(n_3-1)s_3^2}{N-I}$$
    $$= \frac{(16-1)(1.213^2)+(23-1)(1.799^2)+(15-1)(0.975^2)}{54-3}$$
    $$= \frac{106.580}{51} = 2.09$$

  - The ANOVA F statistic

    $$F = \frac{\text{MSG}}{\text{MSE}} = \frac{541.74}{2.09} = 259.20$$

- In Excel

Microsoft Excel - ta25-01.dat

	A	B	C	D	E	F	G
1	Anova: Single Factor						
2							
3	SUMMARY						
4	*Groups*	*Count*	*Sum*	*Average*	*Variance*		
5	bihai	16	761.56	47.5975	1.471073		
6	red	23	913.36	39.7113	3.235548		
7	yellow	15	542.7	36.18	0.951257		
8							
9							
10	ANOVA						
11	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
12	Between Groups	1082.872	2	541.4362	259.1193	1.92E-27	3.178799
13	Within Groups	106.5658	51	2.089525			
14							
15	Total	1189.438	53				

Sheet4 ta25-01

Our work differs slightly from the output in Excel because of roundoff error.

2.2.4 Step 4) Conclusion

- Method 1)

  $F^* < F$ implies that F is in the rejection region. Hence, we reject H_0.

- Method 2)

  Since $P-$value = $P(F(2,51) > 259.20) \approx 0$, which means that $\alpha > P-$value, we reject H_0.

- Not all of μ_1, μ_2, and μ_3 are equal.

3 Exercise

Do high school students from 3 different racial/ethnic groups have different attitudes toward mathematics? Measure the level of interest in mathematics on a 5-point scale for a national random sample of students. Here are summaries for students who were taking math at the time of the survey assuming that all three samples are from normal distribution populations respectively.

Racial/ethnic group	n	$\overline{x}$	s
African American	10	2.57	1.40
White	14	2.32	1.36
Asian/Pacific Islander	6	2.63	1.32

(a) Can we safely use ANOVA F test to compare the mean for the three populations?

(b) Conduct the ANOVA F test at $\alpha = 0.05$.