The phrase *analysis of variance* comes from the idea of analyzing variability in the data to see how much can be attributed to differences in the μ's and how much is due to variability in the individual populations. In Figure 15.1(a), the within-sample variability is small relative to the between-sample variability, whereas in Figure 15.1(b), a great deal more of the total variability is due to variability within each sample. If differences between the sample means can be explained by within-sample variability, there is no compelling reason to reject H_0.

Notation and Assumptions

Notation in single-factor ANOVA is a natural extension of the notation used in Chapter 11 for comparing two population or treatment means.

ANOVA Notation

k = number of populations or treatments being compared

Population or treatment	1	2	...	k
Population or treatment mean	μ_1	μ_2	...	μ_k
Population or treatment variance	σ_1^2	σ_2^2	...	σ_k^2
Sample size	n_1	n_2	...	n_k
Sample mean	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_k$
Sample variance	s_1^2	s_2^2	...	s_k^2

$N = n_1 + n_2 + \cdots + n_k$ (the total number of observations in the data set)

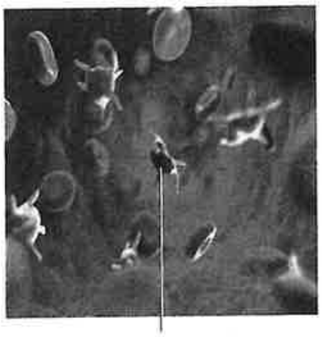
T = grand total = sum of all N observations in the data set $= n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k$

$\bar{\bar{x}}$ = grand mean $= \dfrac{T}{N}$

A decision between H_0 and H_a is based on examining the $\bar{x}$ values to see whether observed differences are small enough to be attributable simply to sampling variability or whether an alternative explanation for the differences is more plausible.

Example 15.1 An Indicator of Heart Attack Risk

Understand the context ➤



Activated platelet

Science Photo Library/Alamy Stock Photo

The article "Could Mean Platelet Volume Be a Predictive Marker for Acute Myocardial Infarction?" (*Medical Science Monitor* [2005]: 387–392) describes an experiment in which four groups of patients seeking treatment for chest pain were compared with respect to mean platelet volume (MPV, measured in fL). The four groups considered were based on the clinical diagnosis and were (1) noncardiac chest pain, (2) stable angina pectoris, (3) unstable angina pectoris, and (4) myocardial infarction (heart attack). The purpose of the study was to determine if the mean MPV differed for the patients in the four groups, and in particular if the mean MPV was different for the heart attack group. If this is the case, MPV could be used as an indicator of heart attack risk.

To carry out this study, patients seen for chest pain were divided into groups according to diagnosis. The researchers then selected a random sample of 35 from each of the resulting $k = 4$ groups. The researchers believed that this sampling process would result in samples that were representative of the four populations of interest and that could be regarded as if they were random samples from these four populations. Table 15.1 presents summary values given in the paper.

With μ_i denoting the true mean MPV for group i ($i = 1, 2, 3, 4$), consider the null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$. Figure 15.2 shows a comparative boxplot for the four samples (based on data consistent with summary values given in the paper). The mean

TABLE 15.1 Summary Values for MPV Data of Example 15.1

Group Number	Group Description	Sample Size	Sample Mean	Sample Standard Deviation
1	Noncardiac chest pain	35	10.89	0.69
2	Stable angina pectoris	35	11.25	0.74
3	Unstable angina pectoris	35	11.37	0.91
4	Myocardial infarction (heart attack)	35	11.75	1.07

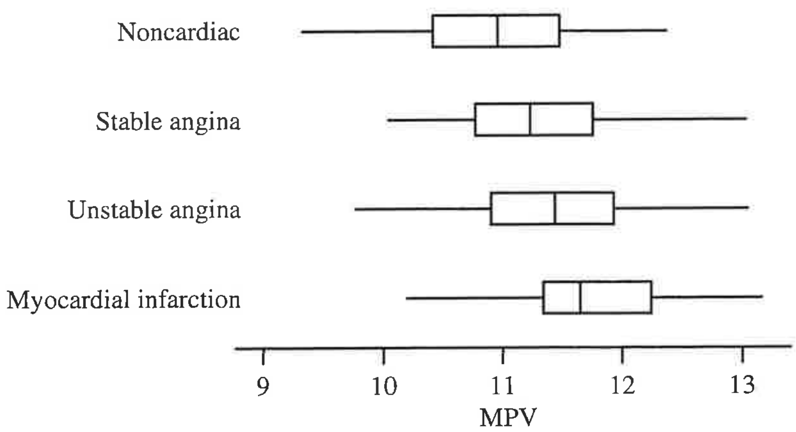


FIGURE 15.2
Boxplots for Example 15.1.

MPV for the heart attack sample is larger than for the other three samples, and the boxplot for the heart attack sample appears to be shifted a bit higher than the boxplots for the other three samples. However, because the four boxplots show substantial overlap, it is not obvious whether H_0 is plausible or should be rejected. In situations like this, a formal test procedure is helpful.

As with the inferential methods of previous chapters, the validity of the ANOVA test for H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$ requires some assumptions.

Assumptions for ANOVA

1. Each of the k population or treatment response distributions is normal.
2. $\sigma_1 = \sigma_2 = \cdots = \sigma_k$ (The k normal distributions have equal standard deviations.)
3. The observations in the sample from any particular one of the k populations or treatments are independent of one another.
4. When comparing population means, the samples are independent random samples. When comparing treatment means, experimental units are assigned at random to treatments.

In practice, the test based on these assumptions works well as long as the assumptions are not too badly violated. If the sample sizes are reasonably large, normal probability plots or boxplots of the data in each sample are helpful for checking the assumption of normality. Often, however, sample sizes are so small that a separate normal probability plot or boxplot for each sample is of little value in checking normality. In this case, a single combined plot can be constructed by first subtracting $\bar{x}_1$ from each observation in the first sample, $\bar{x}_2$ from each value in the second sample, and so on and then constructing a normal probability or boxplot of all N deviations from their respective means. The plot should be reasonably straight. Figure 15.3 shows such a normal probability plot for the data of Example 15.1.

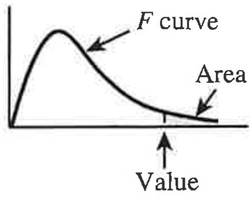


TABLE 6 Values That Capture Specified Upper-Tail F Curve Areas

df_2	Area	df_1 1	2	3	4	5	6	7	8	9	10
1	.10	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19
	.05	161.40	199.50	215.70	224.60	230.20	234.00	236.80	238.90	240.50	241.90
	.01	4052.00	5000.00	5403.00	5625.00	5764.00	5859.00	5928.00	5981.00	6022.00	6056.00
2	.10	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39
	.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
	.01	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
	.001	998.50	999.00	999.20	999.20	999.30	999.30	999.40	999.40	999.40	999.40
3	.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23
	.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
	.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
	.001	167.00	148.50	141.10	137.10	134.60	132.80	131.60	130.60	129.90	129.20
4	.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92
	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
	.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
	.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47	48.05
5	.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30
	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
	.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
	.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24	26.92
6	.10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94
	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
	.01	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
	.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69	18.41
7	.10	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70
	.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
	.01	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
	.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33	14.08
8	.10	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54
	.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
	.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
	.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77	11.54
9	.10	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42
	.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
	.01	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
	.001	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11	9.89

(Continued)

TABLE 6 Values That Capture Specified Upper-Tail *F* Curve Areas *(Continued)*

		df_1									
df_2	Area	1	2	3	4	5	6	7	8	9	10
30	**.10**	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
	.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
	.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
	.001	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39	4.24
40	**.10**	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76
	.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
	.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
	.001	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02	3.87
60	**.10**	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
	.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
	.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
	.001	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69	3.54
90	**.10**	2.76	2.36	2.15	2.01	1.91	1.84	1.78	1.74	1.70	1.67
	.05	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
	.01	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52
	.001	11.57	7.47	5.91	5.06	4.53	4.15	3.87	3.65	3.48	3.34
120	**.10**	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65
	.05	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
	.01	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
	.001	11.38	7.32	5.78	4.95	4.42	4.04	3.77	3.55	3.38	3.24
240	**.10**	2.73	2.32	2.10	1.97	1.87	1.80	1.74	1.70	1.65	1.63
	.05	3.88	3.03	2.64	2.41	2.25	2.14	2.04	1.98	1.92	1.87
	.01	6.74	4.69	3.86	3.40	3.09	2.88	2.71	2.59	2.48	2.40
	.001	11.10	7.11	5.60	4.78	4.25	3.89	3.62	3.41	3.24	3.09
∞	**.10**	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60
	.05	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83
	.01	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32
	.001	10.83	6.91	5.42	4.62	4.10	3.74	3.47	3.27	3.10	2.96