Analysis of Variance(ANOVA)

Definition of Experiment

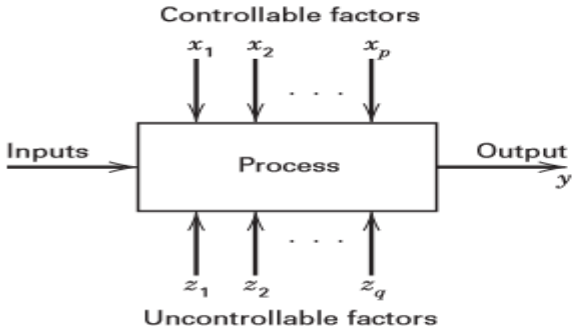
To really understand cause-and-effect relationships in a system you must deliberately change the input variables to the system and observe the changes in the system output that these changes to the inputs produce.

In other words, you need to conduct experiments on the system.

Each experimental run is a **test**

We can define an experiment as a test or series of runs in which purposeful changes are made to the input variables of a process or system so that we may observe and identify the reasons for changes that may be observed in the output response.

Process or System



■ **FIGURE 1.1** General model of a process or system

In general, experiments are used to study the performance of processes and systems.

We can usually visualize the process as a combination of operations, machines, methods, people, and other resources that transforms some input (often a material) into an output that has one or more observable response variables.

Some of the process variables and material properties $x_1, x_2, \ldots, x_p$ are controllable, whereas other variables $z_1, z_2, \ldots, z_q$ are uncontrollable (although they may be controllable for purposes of a test).

Objectives of Experiment

The objectives of the experiment may include the following:

Determining which variables are most influential on the response y

Determining where to set the influential x's so that y is almost always near the desired nominal value

Determining where to set the influential x's so that variability in y is small

Determining where to set the influential x's so that the effects of the uncontrollable variables z_1, z_2, . . . , z_q are minimized.

Strategies of Experimentation

The general approach to planning and conducting the experiment is called the strategy of experimentation

Consider that someone is trying to lower his/her golf score

The person has identified certain “factors” that influences the golf score

- The type of driver used (oversized or regular sized)
- The type of ball used (balata or three piece)
- Walking and carrying the golf clubs or riding in a golf cart
- Drinking water or drinking “something else” while playing
- Playing in the morning or playing in the afternoon
- Playing when it is cool or playing when it is hot
- The type of golf shoe spike worn (metal or soft) 8. Playing on a windy day or playing on a calm day.

continued

Best guess

- Select an arbitrary combination of these factors, test them, and see what happens

One factor at a time(OFAT)

- The OFAT method consists of selecting a starting point, or baseline set of levels, for each factor, and then successively varying each factor over its range with the other factors held constant at the baseline level. After all tests are performed, a series of graphs are usually constructed showing how the response variable is affected by varying each factor with all other factors held constant

Factorial experiment

- factors are varied together, instead of one at a time

An example

In many integrated circuit manufacturing steps, wafers are completely coated with a layer of material such as silicon dioxide or a metal.

The unwanted material is then selectively removed by etching through a mask, thereby creating circuit patterns, electrical interconnects, and areas in which diffusions or metal depositions are to be made.

A plasma etching process is widely used for this operation, particularly in small geometry applications.

An engineer is interested in investigating the relationship between the RF power setting and the etch rate for this tool.

The objective of an experiment like this is to model the relationship between etch rate and RF power, and to specify the power setting that will give a desired target etch rate.

The engineer is interested in a particular gas (C_2F_6) and gap (0.80 cm) and wants to test four levels of RF power: 160, 180, 200, and 220 W.

It was decided to test five wafers at each level of RF power.

This is an example of a single-factor experiment with a = 4 levels of the factor and n = 5 replicates.

The 20 runs should be made in random order.

This randomized test sequence is necessary to prevent the effects of unknown nuisance variables, perhaps varying out of control during the experiment, from contaminating the results.

- Suppose that we were to run the 20 test wafers in the original nonrandomized order (that is, all five 160 W power runs are made first, all five 180 W power runs are made next, and so on).
- If the etching tool exhibits a warm-up effect such that the longer it is on, the lower the observed etch rate readings will be, the warm-up effect will potentially contaminate the data and destroy the validity of the experiment.

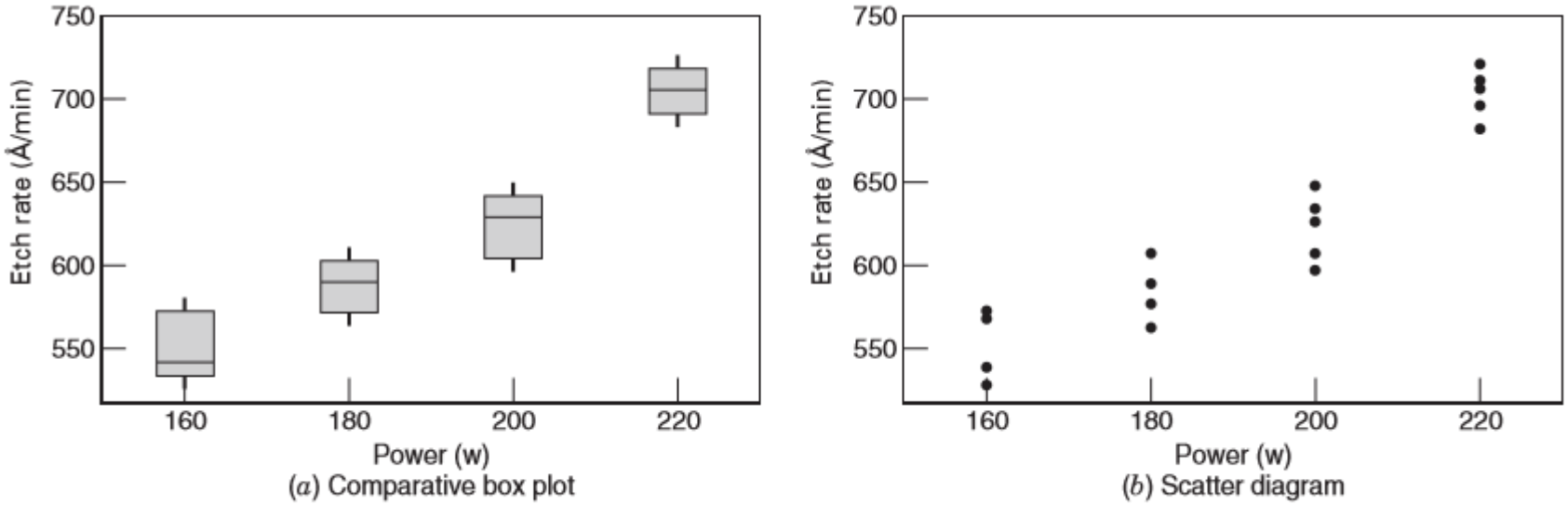
A sample data set

Test Sequence	Excel Random Number (Sorted)	Power
1	12417	200
2	18369	220
3	21238	220
4	24621	160
5	29337	160
6	32318	180
7	36481	200
8	40062	160
9	43289	180
10	49271	200
11	49813	220
12	52286	220
13	57102	160
14	63548	160
15	67710	220
16	71834	180
17	77216	180
18	84675	180
19	89323	200
20	94037	200

■ **TABLE 3.1**

Etch Rate Data (in Å/min) from the Plasma Etching Experiment

Power (W)	Observations 1	2	3	4	5	Totals	Averages
160	575	542	530	539	570	2756	551.2
180	565	593	590	579	610	2937	587.4
200	600	651	610	637	629	3127	625.4
220	725	700	715	685	710	3535	707.0



■ **FIGURE 3.2** **Box plots and scatter diagram of the etch rate data**

Both graphs indicate that etch rate increases as the power setting increases.

There is no strong evidence to suggest that the variability in etch rate around the average depends on the power setting. On the basis of this simple graphical analysis, we strongly suspect that (1) RF power setting affects the etch rate and (2) higher power settings result in increased etch rate.

Suppose that we wish to be more objective in our analysis of the data.

Specifically, suppose that we wish to test for differences between the mean etch rates at all a = 4 levels of RF power.

Thus, we are interested in testing the equality of all four means. It might seem that this problem could be solved by performing a t-test for all six possible pairs of means.

However, this is not the best solution to this problem.

First of all, performing all six pairwise t-tests is inefficient. It takes a lot of effort.

Second, conducting all these pairwise comparisons inflates the type I error. Suppose that all four means are equal, so if we select $\alpha=0.05$, the probability of reaching the correct decision on any single comparison is 0.95. However, the probability of reaching the correct conclusion on all six comparisons is considerably less than 0.95, so the type I error is inflated.

The appropriate procedure for testing the equality of several means is the analysis of variance.

The Analysis of Variance

Suppose we have a treatments or different levels of a single factor that we wish to compare. The observed response from each of the treatments is a random variable.

■ **TABLE 3.2**

Typical Data for a Single-Factor Experiment

Treatment (Level)	Observations				Totals	Averages
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\cdots$	y_{an}	$y_{a.}$	$\bar{y}_{a.}$
					$y_{..}$	$\bar{y}_{..}$

Each value represents the jth observation taken under factor level or treatment i. There will be, in general, n observations under the ith treatment.

Models of the data

$$y_{ij} = \mu_i + \epsilon_{ij} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases}$$

y_{ij} is the ijth observation, μ_i is the mean of the ith factor level or treatment, and ϵ_{ij} is a random error component that incorporates all other sources of variability in the experiment including measurement, variability arising from uncontrolled factors, differences between the experimental units (such as test material, etc.) to which the treatments are applied, and the general background noise in the process (such as variability over time, effects of environmental variables, and so forth).

The errors are assumed to have mean zero, so that $E(y_{ij}) = \mu_i$

The above equation is also known as means model.

μ_i can be written as $\mu_i = \mu + \tau_i, \quad i = 1, 2, \ldots, a$

The means equation can be written as :

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases}$$

This equation is termed as: effects model

In this form of the model, μ is a parameter common to all treatments called the overall mean, and Tau_i is a parameter unique to the ith treatment called the ith treatment effect

Both the means model and the effects model are linear statistical models; that is, the response variable y_{ij} is a linear function of the model parameters.

Although both forms of the model are useful, the effects model is more widely encountered in the experimental design literature.

The above equations are called the one-way or single-factor analysis of variance (ANOVA) model because only one factor is investigated.

It is required that the experiment be performed in random order so that the environment in which the treatments are applied (often called the experimental units) is as uniform as possible. Thus, the experimental design is a completely randomized design.

Our objectives will be to test appropriate hypotheses about the treatment means and to estimate them.

For hypothesis testing, the model errors are assumed to be normally and independently distributed random variables with mean zero and variance σ^2. The variance σ^2 is assumed to be constant for all levels of the factor.

Fixed or random effects

The above equations describe two different situations with respect to the treatment effects. First, the a treatments could have been specifically chosen by the experimenter. In this situation, we wish to test hypotheses about the treatment means, and our conclusions will apply only to the factor levels considered in the analysis. The conclusions cannot be extended to similar treatments that were not explicitly considered. We may also wish to estimate the model parameters (μ, τ_i, σ^2). This is called the **fixed effects model**.

Alternatively, the a treatments could be a random sample from a larger population of treatments. In this situation, we should like to be able to extend the conclusions (which are based on the sample of treatments) to all treatments in the population, whether or not they were explicitly considered in the analysis. Here, the Tau_i are random variables, and knowledge about the particular ones investigated is relatively useless. Instead, we test hypotheses about the variability of the Tau_i and try to estimate this variability. This is called the **random effects model** or components of variance model.

Analysis of the fixed effects model

Let $\bar{y}_{i.}$ represent the average of the observations under the ith treatment. Similarly, let y.. represent the grand total of all the observations and $\bar{y}_{..}$ represent the grand average of all the observations.

$$y_{i.} = \sum_{j=1}^{n} y_{ij} \qquad \bar{y}_{i.} = y_{i.}/n \qquad i = 1, 2, \ldots, a$$

$$y_{..} = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij} \quad \bar{y}_{..} = y_{..}/N$$

where N = an is the total number of observations.

We are interested in testing the equality of the a treatment means.

The appropriate hypotheses are:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_a$$
$$H_1: \mu_i \neq \mu_j \quad \text{for at least one pair } (i, j)$$

In the effects model, we break the ith treatment mean μ_i into two components such that $\mu_i = \mu + \tau_i$.

We can define μ as:

$$\frac{\sum_{i=1}^{a} \mu_i}{a} = \mu$$

Based on the above, we can conclude: $\sum_{i=1}^{a} \tau_i = 0$

Thus, an equivalent way to write the above hypotheses is in terms of the treatment effects:

$$H_0: \tau_1 = \tau_2 = \cdots \tau_a = 0$$
$$H_1: \tau_i \neq 0 \qquad \text{for at least one } i$$

The name analysis of variance is derived from a partitioning of total variability into its component parts. The total corrected sum of squares

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2$$

is used as a measure of overall variabilitv in the data.

$$\sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2 = n \sum_{i=1}^{a} (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2$$

The above equation is the fundamental ANOVA identity.

It states that the total variability in the data, as measured by the total corrected sum of squares, can be partitioned into a sum of squares of the differences between the treatment averages and the grand average plus a sum of squares of the differences of observations within treatments from the treatment average.

The difference between the observed treatment averages and the grand average is a measure of the differences between treatment means, whereas the differences of observations within a treatment from the treatment average can be due to only random error.

So the previous equation can be rewritten as:

$$SS_T = SS_{\text{Treatments}} + SS_E$$

where $SS_{\text{Treatments}}$ is called the sum of squares due to treatments (i.e., between treatments), and SSE is called the sum of squares due to error (i.e., within treatments). There are an = N total observations; thus, SS_T has N - 1 degrees of freedom. There are a levels of the factor (and a treatment means), so $SS_{\text{Treatments}}$ has a - 1 degrees of freedom. Finally, there are n replicates within any treatment providing n - 1 degrees of freedom with which to estimate the experimental error. Because there are a treatments, we have a(n - 1) = an - a =N - a degrees of freedom for error.

$SS_E/(N - a)$is a pooled estimate of the common variance within each of the a treatments

If there were no differences between the a treatment means, we could use the variation of the treatment averages from the grand average to estimate σ^2

$$\frac{SS_{\text{Treatments}}}{a-1} = \frac{n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2}{a-1}$$

The quantities

$$MS_{\text{Treatments}} = \frac{SS_{\text{Treatments}}}{a-1}$$

and

$$MS_E = \frac{SS_E}{N-a}$$

are called **mean squares**. We now examine the **expected values** of these mean squares.

$$E(MS_E) = \sigma^2$$

$$E(MS_{\text{Treatments}}) = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1}$$

Thus, a $MS_E = SS_E/(N - a)$ estimates σ^2, and, if there are no differences in treatment means (which implies that $Tau_i = 0$), $MS_{Treatments} = SS_{Treatments}/(a - 1)$ also estimates σ^2. However, note that if treatment means do differ, the expected value of the treatment mean square is greater than σ^2

A test of the hypothesis of no difference in treatment means can be performed by comparing $MS_{Treatments}$ and MS_E

Let us see next how that can be done.

We have assumed that the errors ϵ_{ij} are normally and independently distributed with mean zero and variance σ^2.

The observations y_{ij} are normally and independently distributed with mean $\mu + \tau_i$ variance σ^2. Thus, SST is a sum of squares in normally distributed random variables; consequently, it can be shown that SST/σ^2 is distributed as chi-square with N - 1 degrees of freedom.

It can be shown that SSE/σ^2 is chi-square with N - a degrees of freedom and that $SS_{Treatments}/\sigma^2$ is chi-square with a - 1 degrees of freedom if the null hypothesis $H_0: \tau_i = 0$ is true. However, all three sums of squares are not necessarily independent because $SS_{Treatments}$ and SS_E add to SS_T.

If the null hypothesis of no difference in treatment means is true, the ratio

$$F_0 = \frac{SS_{\text{Treatments}}/(a-1)}{SS_E/(N-a)} = \frac{MS_{\text{Treatments}}}{MS_E}$$

is distributed as F with a - 1 and N - a degrees of freedom.

The previous equation is the test statistic for the hypothesis of no differences in treatment means.

We should reject H_0 and conclude that there are differences in the treatment means if

$$F_0 > F_{\alpha,a-1,N-a}$$

where F0 is computed from the previous equation.

Next we list the formulations to find the various sum of squares.

■ **TABLE 3.3**

The Analysis of Variance Table for the Single-Factor, Fixed Effects Model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Between treatments	$SS_{\text{Treatments}} = n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2$	$a - 1$	$MS_{\text{Treatments}}$	$F_0 = \dfrac{MS_{\text{Treatments}}}{MS_E}$
Error (within treatments)	$SS_E = SS_T - SS_{\text{Treatments}}$	$N - a$	MS_E	
Total	$SS_{\text{T}} = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2$	$N-1$		

To illustrate the analysis of variance, return to the first example discussed in Section 3.1. Recall that the engineer is interested in determining if the RF power setting affects the etch rate, and she has run a completely randomized experiment with four levels of RF power and five replicates. For convenience, we repeat here the data from Table 3.1:

We will use the analysis of variance to test $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ against the alternative H_1: some means are different. The sums of squares required are computed using Equations 3.8, 3.9, and 3.10 as follows:

RF Power (W)	Observed Etch Rate (Å/min)					Totals	Averages
	1	2	3	4	5	$y_{i.}$	$\bar{y}_{i.}$
160	575	542	530	539	570	2756	551.2
180	565	593	590	579	610	2937	587.4
200	600	651	610	637	629	3127	625.4
220	725	700	715	685	710	3535	707.0
						$y_{..} = 12{,}355$	$\bar{y}_{..} = 617.75$

Appendix 4 (to be used in the next slide)

$F_{0.05,\nu_1,\nu_2}$

ν_2 \ ν_1	Degrees of Freedom for the Numerator (ν_1)											
	1	2	3	4	5	6	7	8	9	10	12	15
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15

Degrees of Freedom for the Denominator (ν_2)

$$SS_T = \sum_{i=1}^{4}\sum_{j=1}^{5} y_{ij}^2 - \frac{y_{..}^2}{N}$$

$$= (575)^2 + (542)^2 + \cdots + (710)^2 - \frac{(12{,}355)^2}{20}$$

$$= 72{,}209.75$$

$$SS_{\text{Treatments}} = \frac{1}{n}\sum_{i=1}^{4} y_{i.}^2 - \frac{y_{..}^2}{N}$$

$$= \frac{1}{5}[(2756)^2 + \cdots + (3535)^2] - \frac{(12{,}355)^2}{20}$$

$$= 66{,}870.55$$

$$SS_E = SS_T - SS_{\text{Treatments}}$$

$$= 72{,}209.75 - 66{,}870.55 = 5339.20$$

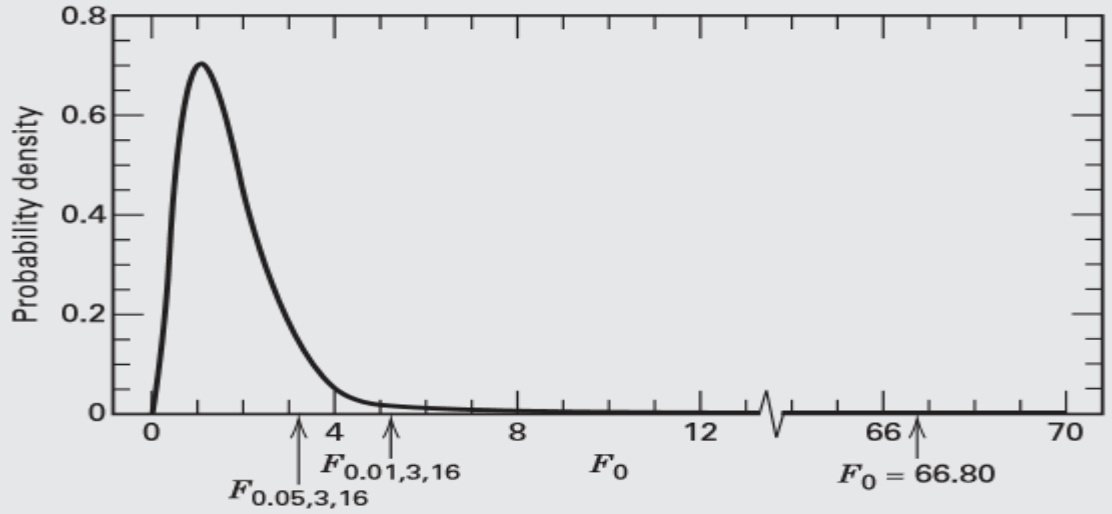
Usually, these calculations would be performed on a computer, using a software package with the capability to analyze data from designed experiments.

The ANOVA is summarized in Table 3.4. Note that the RF power or between-treatment mean square (22,290.18) is many times larger than the within-treatment or error mean square (333.70). This indicates that it is unlikely that the treatment means are equal. More formally, we can compute the F ratio $F_0 = 22{,}290.18/333.70 = 66.80$ and compare this to an appropriate upper-tail percentage point of the $F_{3,16}$ distribution. To use a fixed significance level approach, suppose that the experimenter has selected $\alpha = 0.05$. From Appendix Table IV we find that $F_{0.05,3,16} = 3.24$. Because $F_0 = 66.80 > 3.24$, we reject H_0 and conclude that the treatment means differ; that is, the RF power setting significantly affects the mean etch rate. We could also compute a P-value for this test statistic. Figure 3.3 shows the reference distribution ($F_{3,16}$) for the test statistic F_0. Clearly, the P-value is very small in this case. From Appendix Table A-4, we find that $F_{0.01,3,16} = 5.29$ and because $F_0 > 5.29$, we can conclude that an upper bound for the P-value is 0.01; that is, $P < 0.01$ (the exact P-value is $P = 2.88 \times 10^{-9}$).

■ **TABLE 3.4**

ANOVA for the Plasma Etching Experiment

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	*P*-Value
RF Power	66,870.55	3	22,290.18	$F_0 = 66.80$	<0.01
Error	5339.20	16	333.70		
Total	72,209.75	19			



■ **FIGURE 3.3** The reference distribution ($F_{3,16}$) for the test statistic F_0 in Example 3.1

We now present estimators for the parameters in the single-factor model $y_{ij} = \mu + \tau_i + \epsilon_{ij}$

$$\hat{\mu} = \bar{y}_{..}$$

$$\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..}, \qquad i = 1, 2, \ldots, a$$

A confidence interval estimate of the ith treatment mean may be easily determined. The mean of the ith treatment is $\mu_i = \mu + \tau_i$

A point estimator of i would be $\hat{\mu}_i = \hat{\mu} + \hat{\tau}_i = \bar{y}_{i.}$

a 100(1-α) percent confidence interval on the ith treatment mean μ_i is

$$\bar{y}_{i.} - t_{\alpha/2,N-a}\sqrt{\frac{MS_E}{n}} \le \mu_i \le \bar{y}_{i.} + t_{\alpha/2,N-a}\sqrt{\frac{MS_E}{n}}$$

A 100(1 - α) percent confidence interval on the difference in any two treatments means, say μ_i - μ_j

$$\bar{y}_{i.} - \bar{y}_{j.} - t_{\alpha/2,N-a}\sqrt{\frac{2MS_E}{n}} \le \mu_i - \mu_j \le \bar{y}_{i.} - \bar{y}_{j.} + t_{\alpha/2,N-a}\sqrt{\frac{2MS_E}{n}}$$

Using the data in Example 3.1, we may find the estimates of the overall mean and the treatment effects as $\hat{\mu} = 12{,}355/20 = 617.75$ and

$$\begin{aligned}\hat{\tau}_1 &= \bar{y}_{1.} - \bar{y}_{..} = 551.20 - 617.75 = -66.55\\ \hat{\tau}_2 &= \bar{y}_{2.} - \bar{y}_{..} = 587.40 - 617.75 = -30.35\\ \hat{\tau}_3 &= \bar{y}_{3.} - \bar{y}_{..} = 625.40 - 617.75 = 7.65\\ \hat{\tau}_4 &= \bar{y}_{4.} - \bar{y}_{..} = 707.00 - 617.75 = 89.25\end{aligned}$$

A 95 percent confidence interval on the mean of treatment 4 (220W of RF power) is computed from Equation 3.12 as

$$707.00 - 2.120\sqrt{\frac{333.70}{5}} \le \mu_4 \le 707.00 + 2.120\sqrt{\frac{333.70}{5}}$$

or

$$707.00 - 17.32 \le \mu_4 \le 707.00 + 17.32$$

Thus, the desired 95 percent confidence interval is $689.68 \le \mu_4 \le 724.32$.

In some single-factor experiments, the number of observations taken within each treatment may be different. We then say that the design is unbalanced.

The manual computational formulas for SS_T and $SS_{\text{Treatments}}$ become

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{N}$$

$$SS_{\text{Treatments}} = \sum_{i=1}^{a} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N}$$

No other changes are required in the analysis of variance

There are two advantages in choosing a balanced design. First, the test statistic is relatively insensitive to small departures from the assumption of equal variances for the a treatments if the sample sizes are equal. This is not the case for unequal sample sizes. Second, the power of the test is maximized if the samples are of equal size.

Model adequacy checking

The single factor model works on the assumption:

The errors are normally and independently distributed with mean zero and constant but unknown variance σ^2. If these assumptions are valid, the analysis of variance procedure is an exact test of the hypothesis of no difference in treatment means.

In practice, however, these assumptions will usually not hold exactly. Consequently, it is usually unwise to rely on the analysis of variance until the validity of these assumptions has been checked.

Violations of the basic assumptions and model adequacy can be easily investigated by the examination of residuals.

$$e_{ij} = y_{ij} - \hat{y}_{ij}$$

Where $\hat{y}_{ij}$ is an estimate of the corresponding observation y_{ij} obtained as follows:

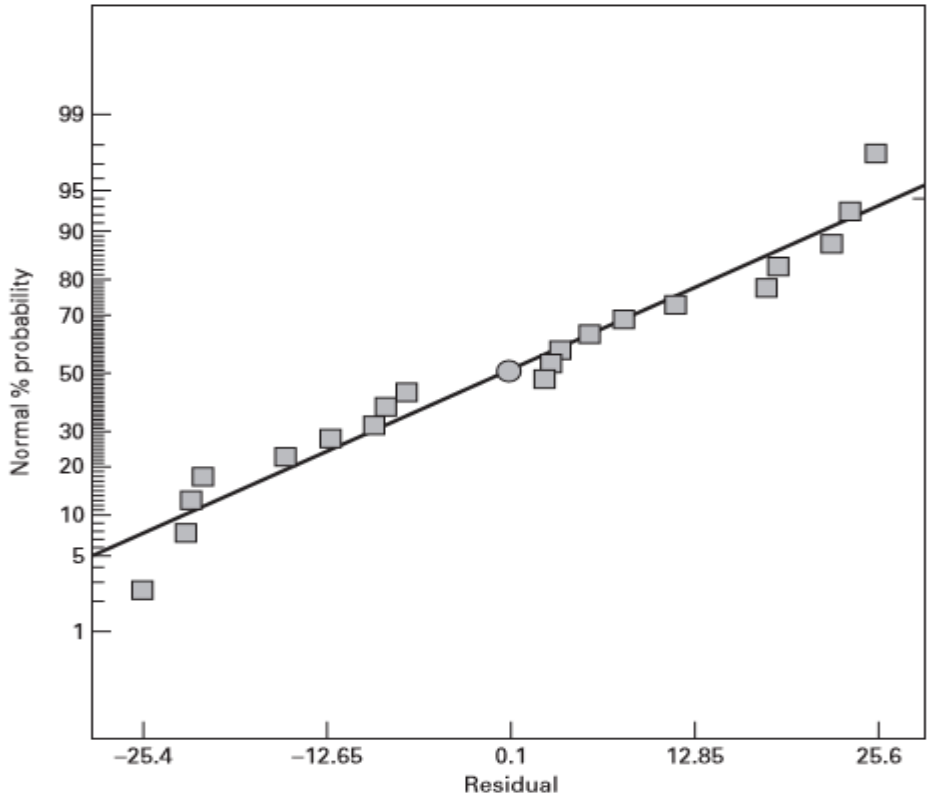
$$\begin{aligned}\hat{y}_{ij} &= \hat{\mu} + \hat{\tau}_i \\ &= \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) \\ &= \bar{y}_{i.}\end{aligned}$$

If the model is adequate, the residuals should be structureless; that is, they should contain no obvious patterns.

Normality assumption

A check of the normality assumption could be made by plotting a histogram of the residuals.

One way is to construct a normal probability plot of the residuals. If the underlying error distribution is normal, this plot will resemble a straight line.



Because the F test is only slightly affected, we say that the analysis of variance (and related procedures such as multiple comparisons) is robust to the normality assumption.

A very common defect that often shows up on normal probability plots is one residual that is very much larger than any of the others. Such a residual is often called an outlier. The presence of one or more outliers can seriously distort the analysis of variance, so when a potential outlier is located, careful investigation is called for.

A rough check for outliers may be made by examining the standardized residuals:

$$d_{ij} = \frac{e_{ij}}{\sqrt{MS_E}}$$

If the errors ϵ_{ij} are $N(0, \sigma^2)$, the standardized residuals should be approximately normal with mean zero and unit variance.

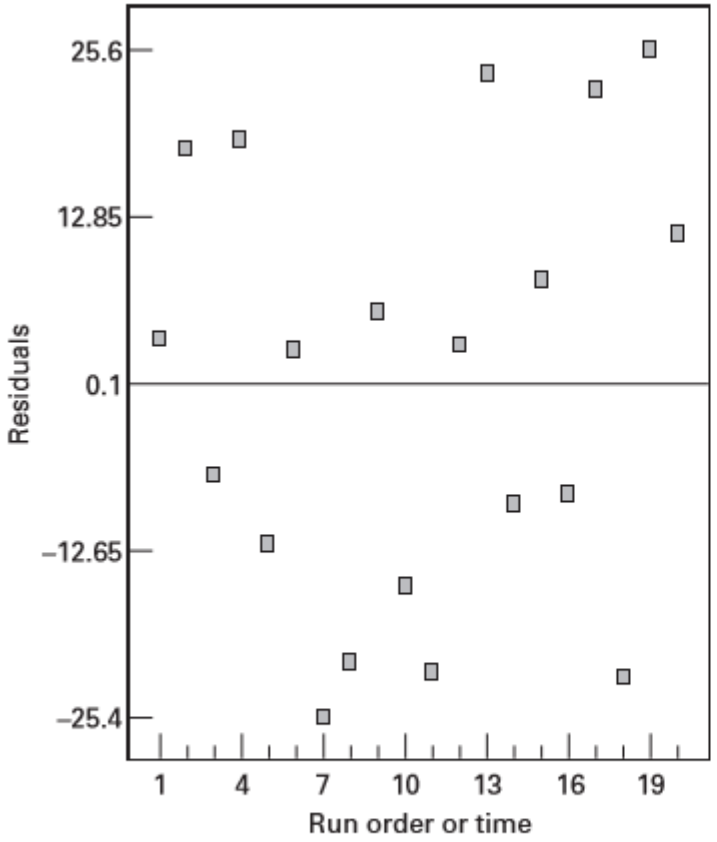
Plotting the residuals in time order of data collection is helpful in detecting strong correlation between the residuals.

A tendency to have runs of positive and negative residuals indicates positive correlation. This would imply that the independence assumption on the errors has been violated. This is a potentially serious problem and one that is difficult to correct, so it is important to prevent the problem if possible when the data are collected.

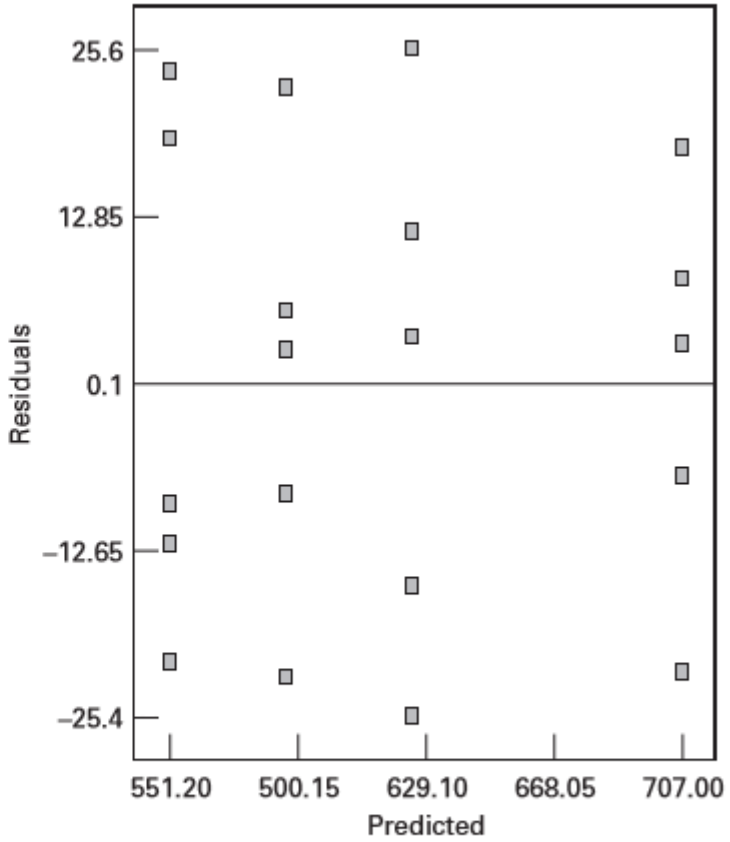
Proper randomization of the experiment is an important step in obtaining independence.

Sometimes the skill of the experimenter (or the subjects) may change as the experiment progresses, or the process being studied may “drift” or become more erratic.

This will often result in a change in the error variance over time. This condition often leads to a plot of residuals versus time that exhibits more spread at one end than at the other.



■ **FIGURE 3.5** Plot of residuals versus run order or time



■ **FIGURE 3.6** Plot of residuals versus fitted values

Plot of Residuals Versus Fitted Values

If the model is correct and the assumptions are satisfied, the residuals should be structureless; in particular, they should be unrelated to any other variable including the predicted response. A simple check is to plot the residuals versus the fitted values .

This plot should not reveal any obvious pattern.

A defect that occasionally shows up on this plot is nonconstant variance. Sometimes the variance of the observations increases as the magnitude of the observation increases. This would be the case if the error or background noise in the experiment was a constant percentage of the size of the observation.

If the assumption of homogeneity of variances is violated,the Ftest is only slightly affected in the balanced (equal sample sizes in all treatments) fixed effects model.

The usual approach to dealing with nonconstant variance when it occurs for the above reasons is to apply a variance-stabilizing transformation and then to run the analysis of variance on the transformed data. In this approach, one should note that the conclusions of the analysis of variance apply to the transformed populations.

Example on transformation

For example, if the observations follow the Poisson distribution, the **square root transformation** $y_{ij}^* = \sqrt{y_{ij}}$ or $y_{ij}^* = \sqrt{1 + y_{ij}}$ would be used. If the data follow the lognormal distribution, the **logarithmic transformation** $y_{ij}^* = \log y_{ij}$ is appropriate. For binomial data expressed as fractions, the **arcsin transformation** $y_{ij}^* = \arcsin \sqrt{y_{ij}}$ is useful. When there is no obvious transformation, the experimenter usually *empirically* seeks a transformation that equalizes the variance regardless of the value of the mean.

Statistical Tests for Equality of Variance.

$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_a^2$

H_1: above not true for at least one σ_i^2

A widely used procedure is Bartlett's test. The procedure involves computing a statistic whose sampling distribution is closely approximated by the chi-square distribution with a 1 degrees of freedom when the a random samples are from independent normal populations. The test statistic is

$$\chi_0^2 = 2.3026 \frac{q}{c}$$

where

$$q = (N - a)\log_{10} S_p^2 - \sum_{i=1}^{a} (n_i - 1)\log_{10} S_i^2$$

$$c = 1 + \frac{1}{3(a-1)}\left(\sum_{i=1}^{a} (n_i - 1)^{-1} - (N - a)^{-1}\right)$$

$$S_p^2 = \frac{\sum_{i=1}^{a} (n_i - 1)S_i^2}{N - a}$$

and S_i^2 is the sample variance of the ith population.

we reject H_0 only when

$$\chi_0^2 > \chi_{\alpha,a-1}^2$$

where $\chi_{\alpha,a-1}^2$ is the upper α percentage point of the chi-square distribution with $a - 1$ degrees of freedom. The P-value approach to decision making could also be used.

This test is highly sensitive to normality assumption

An example

In the plasma etch experiment, the normality assumption is not in question, so we can apply Bartlett's test to the etch rate data. We first compute the sample variances in each treatment and find that $S_1^2 = 400.7$, $S_2^2 = 280.3$, $S_3^2 = 421.3$, and $S_4^2 = 232.5$. Then

$$S_p^2 = \frac{4(400.7) + 4(280.3) + 4(421.3) + 4(232.5)}{16} = 333.7$$

$$q = 16 \log_{10}(333.7) - 4[\log_{10}400.7 + \log_{10}280.3 + \log_{10}421.3 + \log_{10}232.5] = 0.21$$

$$c = 1 + \frac{1}{3(3)}\left(\frac{4}{4} - \frac{1}{16}\right) = 1.10$$

and the test statistic is

$$\chi_0^2 = 2.3026 \frac{(0.21)}{(1.10)} = 0.43$$

From Appendix Table III, we find that $\chi_{0.05,3}^2 = 7.81$ (the P-value is $P = 0.934$), so we cannot reject the null hypothesis. There is no evidence to counter the claim that all five variances are the same. This is the same conclusion reached by analyzing the plot of residuals versus fitted values.

Some other transformation types

■ **TABLE 3.9**

Variance-Stabilizing Transformations

Relationship Between σ_y and μ	α	$\lambda = 1 - \alpha$	Transformation	Comment
$\sigma_y \propto$ constant	0	1	No transformation	
$\sigma_y \propto \mu^{1/2}$	1/2	1/2	Square root	Poisson (count) data
$\sigma_y \propto \mu$	1	0	Log	
$\sigma_y \propto \mu^{3/2}$	3/2	−1/2	Reciprocal square root	
$\sigma_y \propto \mu^2$	2	−1	Reciprocal	

Interpreting results

The factors involved in an experiment can be either quantitative or qualitative.

A quantitative factor is one whose levels can be associated with points on a numerical scale, such as temperature, pressure, or time.

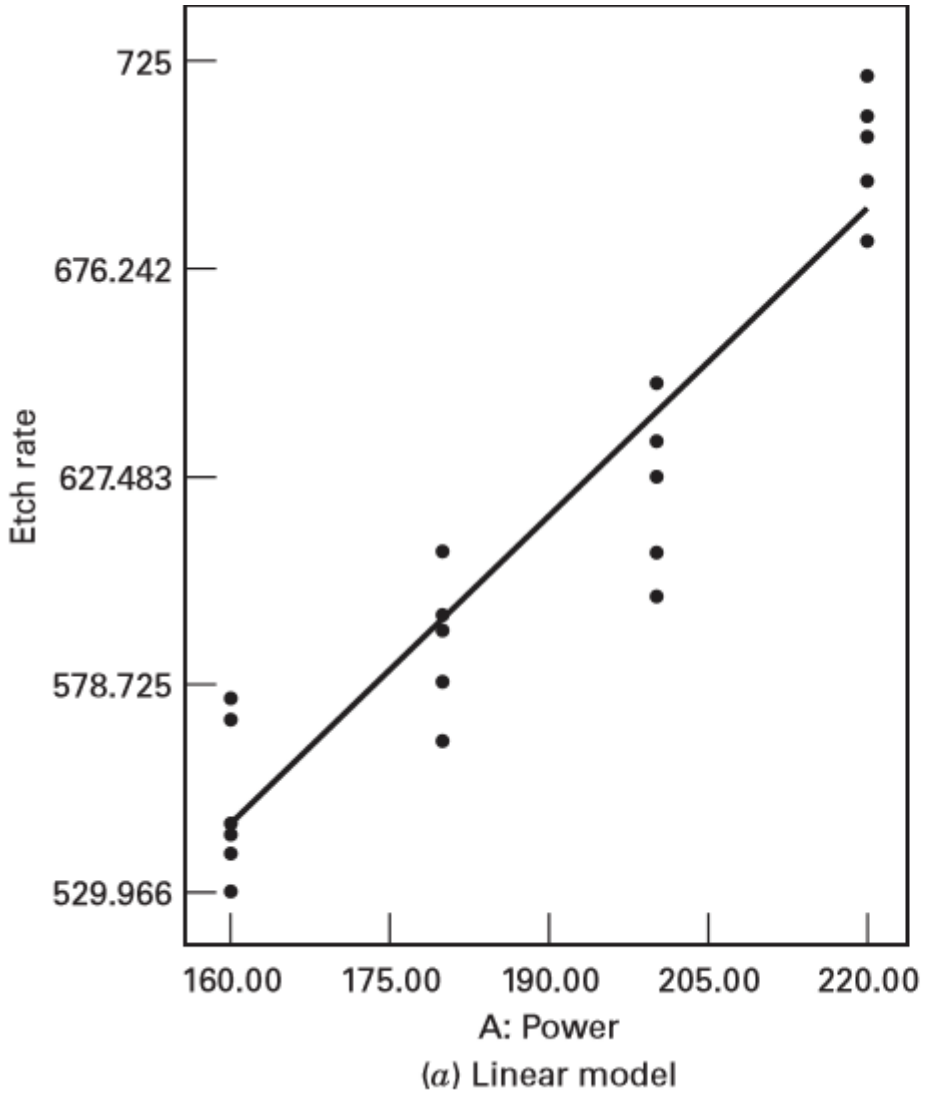
Qualitative factors, on the other hand, are factors for which the levels cannot be arranged in order of magnitude. Operators, batches of raw material, and shifts are typical qualitative factors because there is no reason to rank them in any particular numerical order.

the experimenter is frequently interested in developing an interpolation equation for the response variable in the experiment. This equation is an empirical model of the process that has been studied.

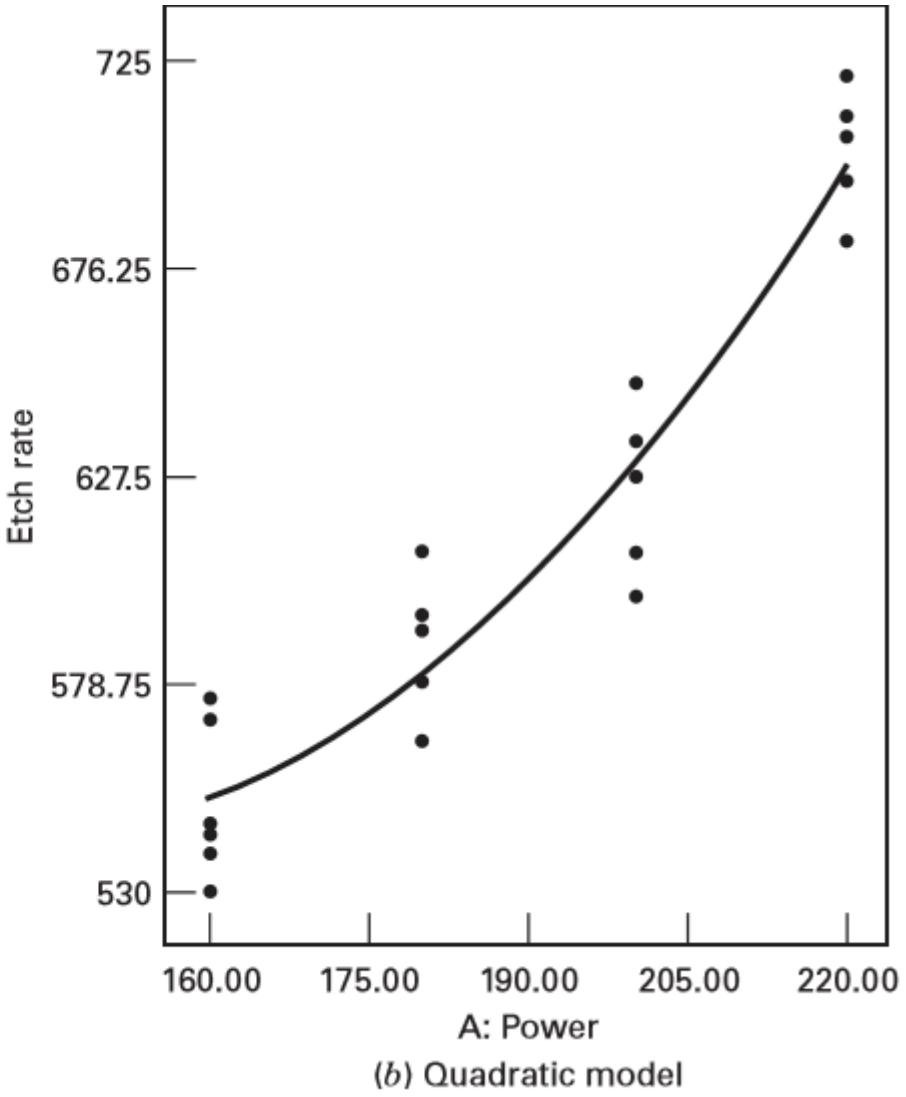
We initially try fitting a linear model to the data, say

$$y = \beta_0 + \beta_1 x + \epsilon$$

where β_0 and β_1 are unknown parameters to be estimated and ϵ is a random error term.



(*a*) Linear model



(*b*) Quadratic model

The quadratic model is made since the liner model does not produce a good fit.

The model is:

$$\hat{y} = 1147.77 - 8.2555\,x + 0.028375\,x^2$$

The empirical model is used to predict etch rate at power settings within the region of experimentation. In other cases, the empirical model could be used for process optimization

Graphical comparison of means

It is very easy to develop a graphical procedure for the comparison of means following an analysis of variance. Suppose that the factor of interest has a levels and that $\bar{y}_{1.}, \bar{y}_{2.}, \ldots, \bar{y}_{a.}$ are the treatment averages. If we know σ, any treatment average would have a standard deviation $\sigma/\sqrt{n}$. Consequently, if all factor level means are identical, the observed sample means $\bar{y}_{i.}$ would behave as if they were a set of observations drawn at random from a normal distribution with mean $\bar{y}_{..}$ and standard deviation $\sigma/\sqrt{n}$. Visualize a normal distribution capable of being slid along an axis below which the $\bar{y}_{1.}, \bar{y}_{2.}, \ldots, \bar{y}_{a.}$ are plotted. If the treatment means are all equal, there should be some position for this distribution that makes it obvious that the $\bar{y}_{i.}$ values were drawn from the same distribution. If this is not the case, the $\bar{y}_{i.}$ values that appear *not* to have been drawn from this distribution are associated with factor levels that produce different mean responses.

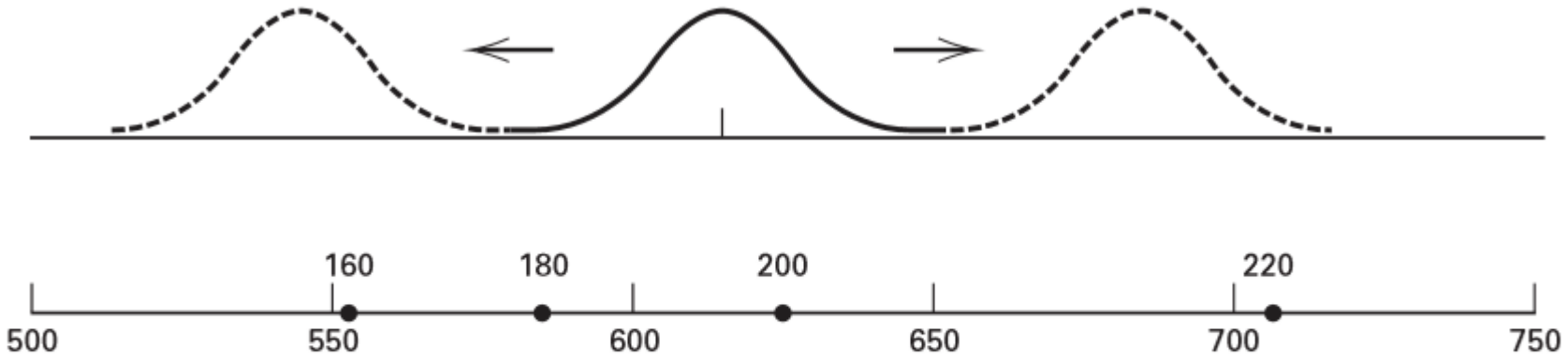
The only flaw in this logic is that σ is unknown. Box, Hunter, and Hunter (2005) point out that we can replace σ with $\sqrt{MS_E}$ from the analysis of variance and use a t distribution with a scale factor $\sqrt{MS_E/n}$ instead of the normal. Such an arrangement for the etch rate data of Example 3.1 is shown in Figure 3.11. Focus on the t distribution shown as a solid line curve in the middle of the display.

To sketch the t distribution in Figure 3.11, simply multiply the abscissa t value by the scale factor

$$\sqrt{MS_E/n} = \sqrt{330.70/5} = 8.13$$

and plot this against the ordinate of t at that point. Because the t distribution looks much like the normal, except that it is a little flatter near the center and has longer tails, this sketch is usually easily constructed by eye. If you wish to be more precise, there is a table of abscissa t values and the corresponding ordinates in Box, Hunter, and Hunter (2005). The distribution can have an arbitrary origin, although it is usually best to choose one in the region of the $\bar{y}_i$.

Visualize sliding the t distribution in Figure 3.11 along the horizontal axis as indicated by the dashed lines and examine the four means plotted in the figure. Notice that there is no location for the distribution such that all four averages could be thought of as typical, randomly selected observations from the distribution. This implies that all four means are not equal; thus, the figure is a graphical display of the ANOVA results. Furthermore, the figure indicates that all four levels of power (160, 180, 200, 220 W) produce mean etch rates that differ from each other.



■ **FIGURE 3.11** Etch rate averages from Example 3.1 in relation to a t distribution with scale factor $\sqrt{MS_E/n} = \sqrt{330.70/5} = 8.13$

Contrast

Many multiple comparison methods use the idea of a contrast. Consider the plasma etching experiment. Because the null hypothesis was rejected, we know that some power settings produce different etch rates than others, but which ones actually cause this difference? We might suspect at the outset of the experiment that 200 W and 220 W produce the same etch rate, implying that we would like to test the hypothesis:

$$H_0: \mu_3 = \mu_4$$
$$H_1: \mu_3 \neq \mu_4$$

or equivalently

$$H_0: \mu_3 - \mu_4 = 0$$
$$H_1: \mu_3 - \mu_4 \neq 0$$

If we had suspected at the start of the experiment that the average of the lowest levels of power did not differ from the average of the highest levels of power, then the hypothesis would have been

$$H_0: \mu_1 + \mu_2 = \mu_3 + \mu_4$$
$$H_1: \mu_1 + \mu_2 \neq \mu_3 + \mu_4$$

or

$$H_0: \mu_1 + \mu_2 - \mu_3 - \mu_4 = 0$$
$$H_1: \mu_1 + \mu_2 - \mu_3 - \mu_4 \neq 0$$

In general, a **contrast** is a linear combination of parameters of the form

$$\Gamma = \sum_{i=1}^{a} c_i \mu_i$$

where the **contrast constants** $c_1, c_2, \ldots, c_a$ sum to zero; that is, $\sum_{i=1}^{a} c_i = 0$. Both of the above hypotheses can be expressed in terms of contrasts:

$$H_0: \sum_{i=1}^{a} c_i \mu_i = 0$$
$$H_1: \sum_{i=1}^{a} c_i \mu_i \neq 0 \qquad \textbf{(3.25)}$$

The contrast constants for the hypotheses in Equation 3.23 are $c_1 = c_2 = 0$, $c_3 = +1$, and $c_4 = -1$, whereas for the hypotheses in Equation 3.24, they are $c_1 = c_2 = +1$ and $c_3 = c_4 = -1$.

Equation 3.23 and 3.24 are the hypothesises mentioned in the previous slide

Testing hypotheses involving contrasts can be done in two basic ways. The first method uses a t-test.

The following statistics is calculated to test the hypothesis mentioned in the previous slide (3.25)

$$t_0 = \frac{\sum_{i=1}^{a} c_i \bar{y}_{i.}}{\sqrt{\frac{MS_E}{n} \sum_{i=1}^{a} c_i^2}}$$

The null hypothesis would be rejected if $|t_0|$ exceeds $t_{\alpha/2,N-a}$.

The second approach uses an F test. Now the square of a t random variable with degrees of freedom is an F random variable with 1 numerator and v denominator degrees of freedom. Therefore, we can obtain

$$F_0 = t_0^2 = \frac{\left(\sum_{i=1}^{a} c_i \bar{y}_{i.}\right)^2}{\frac{MS_E}{n} \sum_{i=1}^{a} c_i^2}$$

The null hypothesis would be rejected if $F_0 > F_{\alpha,1,N-a}$.

Orthogonal contrasts

Two contrasts with coefficients {ci} and {di} are orthogonal if

$$\sum_{i=1}^{a} c_i d_i = 0$$

or, for an unbalanced design, if

$$\sum_{i=1}^{a} n_i c_i d_i = 0$$

For a treatments, the set of a - 1 orthogonal contrasts partition the sum of squares due to treatments into a - 1 independent single-degree-of-freedom components. Thus, tests performed on orthogonal contrasts are independent.

There are many ways to choose the orthogonal contrast coefficients for a set of treatments. Usually, something in the nature of the experiment should suggest which comparisons will be of interest.

Treatment	Coefficients for Orthogonal Contrasts	
1 (control)	−2	0
2 (level 1)	1	−1
3 (level 2)	1	1

Generally, the method of contrasts (or orthogonal contrasts) is useful for what are called preplanned comparisons. That is, the contrasts are specified prior to running the experiment and examining the data.

Examining the data to select comparisons of potential interest is often called data snooping

Comparing pair of treatment means

In many practical situations, we will wish to compare only pairs of means. Frequently, we can determine which means differ by testing the differences between all pairs of treatment means. Thus, we are interested in contrasts of the form $\Gamma = \mu_i - \mu_j$ for all $i \neq j$.

Suppose that we are interested in comparing all pairs of treatment means and that the null hypotheses that we wish to test are $H_0: \mu_i = \mu_j$ for all $i \neq j$.

Tukey's Test:

Suppose that, following an ANOVA in which we have rejected the null hypothesis of equal treatment means, we wish to test all pairwise mean comparisons:

$$H_0: \mu_i = \mu_j$$
$$H_1: \mu_i \neq \mu_j$$

Tukey's procedure makes use of the distribution of the studentized range statistic

$$q = \frac{\bar{y}_{\max} - \bar{y}_{\min}}{\sqrt{MS_E/n}}$$

where $\bar{y}_{\max}$ and $\bar{y}_{\min}$ are the largest and smallest sample means, respectively, out of a group of p sample means. Appendix Table VII contains values of $q_\alpha(p, f)$, the upper α percentage points of q, where f is the number of degrees of freedom associated with the MS_E.

For equal sample sizes, Tukey's test declares two means significantly different if the absolute value of their sample differences exceeds

$$T_\alpha = q_\alpha(a, f)\sqrt{\frac{MS_E}{n}}$$

Appendix (to be used in next slide)

$q_{0.05}(p, f)$

	p								
f	2	3	4	5	6	7	8	9	10	11
1	18.1	26.7	32.8	37.2	40.5	43.1	45.4	47.3	49.1	50.6
2	6.09	8.28	9.80	10.89	11.73	12.43	13.03	13.54	13.99	14.39
3	4.50	5.88	6.83	7.51	8.04	8.47	8.85	9.18	9.46	9.72
4	3.93	5.00	5.76	6.31	6.73	7.06	7.35	7.60	7.83	8.03
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17
6	3.46	4.34	4.90	5.31	5.63	5.89	6.12	6.32	6.49	6.65
7	3.34	4.16	4.68	5.06	5.35	5.59	5.80	5.99	6.15	6.29
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05
9	3.20	3.95	4.42	4.76	5.02	5.24	5.43	5.60	5.74	5.87
10	3.15	3.88	4.33	4.66	4.91	5.12	5.30	5.46	5.60	5.72
11	3.11	3.82	4.26	4.58	4.82	5.03	5.20	5.35	5.49	5.61
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.40	5.51
13	3.06	3.73	4.15	4.46	4.69	4.88	5.05	5.19	5.32	5.43
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	5.31
16	3.00	3.65	4.05	4.34	4.56	4.74	4.90	5.03	5.15	5.26

Tukey's test example

To illustrate Tukey's test, we use the data from the plasma etching experiment in Example 3.1. With $\alpha = 0.05$ and $f = 16$ degrees of freedom for error, Appendix Table VII gives $q_{0.05}(4, 16) = 4.05$. Therefore, from Equation 3.35,

$$T_{0.05} = q_{0.05}(4, 16)\sqrt{\frac{MS_E}{n}} = 4.05\sqrt{\frac{333.70}{5}} = 33.09$$

Thus, any pairs of treatment averages that differ in absolute value by more than 33.09 would imply that the corresponding pair of population means are significantly different. The four treatment averages are

$$\bar{y}_{1.} = 551.2 \quad \bar{y}_{2.} = 587.4$$
$$\bar{y}_{3.} = 625.4 \quad \bar{y}_{4.} = 707.0$$

and the differences in averages are

$$\bar{y}_{1.} - \bar{y}_{2.} = 551.2 - 587.4 = -36.20^*$$
$$\bar{y}_{1.} - \bar{y}_{3.} = 551.2 - 625.4 = -74.20^*$$
$$\bar{y}_{1.} - \bar{y}_{4.} = 551.2 - 707.0 = -155.8^*$$
$$\bar{y}_{2.} - \bar{y}_{3.} = 587.4 - 625.4 = -38.0^*$$
$$\bar{y}_{2.} - \bar{y}_{4.} = 587.4 - 707.0 = -119.6^*$$
$$\bar{y}_{3.} - \bar{y}_{4.} = 625.4 - 707.0 = -81.60^*$$

The starred values indicate the pairs of means that are significantly different. Note that the Tukey procedure indicates that all pairs of means differ. Therefore, each power setting results in a mean etch rate that differs from the mean etch rate at any other power setting.

Comparing Treatment Means with a Control (Dunnet's test)

In many experiments, one of the treatments is a control, and the analyst is interested in comparing each of the other a 1 treatment means with the control. Thus, only a 1 comparisons are to be made.

Suppose that treatment a is the control and we wish to test the hypotheses

$$H_0: \mu_i = \mu_a$$

$$H_1: \mu_i \neq \mu_a$$

For each hypothesis, we compute the observed differences in the sample means

$$|\bar{y}_{i.} - \bar{y}_{a.}| \qquad i = 1, 2, \ldots, a-1$$

The null hypothesis $H_0: \mu_i = \mu_a$ is rejected using a type I error rate α if

$$|\bar{y}_{i.} - \bar{y}_{a.}| > d_{\alpha}(a-1, f)\sqrt{MS_E\left(\frac{1}{n_i} + \frac{1}{n_a}\right)}$$

where the constant $d_{\alpha}(a-1, f)$ is given in Appendix Table VIII. (Both two- and one-sided tests are possible.) Note that α is the **joint significance level** associated with all $a-1$ tests.

An example

To illustrate Dunnett's test, consider the experiment from Example 3.1 with treatment 4 considered as the control. In this example, $a = 4$, $a - 1 = 3$, $f = 16$, and $n_i = n = 5$. At the 5 percent level, we find from Appendix Table VIII that $d_{0.05}(3, 16) = 2.59$. Thus, the critical difference becomes

$$d_{0.05}(3, 16)\sqrt{\frac{2MS_E}{n}} = 2.59\sqrt{\frac{2(333.70)}{5}} = 29.92$$

(Note that this is a simplification of Equation 3.42 resulting from a balanced design.) Thus, any treatment mean that differs in absolute value from the control by more than 29.92 would be declared significantly different. The observed differences are

$$1 \text{ vs. } 4: \bar{y}_{1.} - \bar{y}_{4.} = 551.2 - 707.0 = -155.8$$
$$2 \text{ vs. } 4: \bar{y}_{2.} - \bar{y}_{4.} = 587.4 - 707.0 = -119.6$$
$$3 \text{ vs. } 4: \bar{y}_{3.} - \bar{y}_{4.} = 625.4 - 707.0 = -81.6$$

Note that all differences are significant. Thus, we would conclude that all power settings are different from the control.

$d_{0.05}(a - 1, f)$
Two-Sided Comparisons

	$a - 1$ = Number of Treatment Means (Excluding Control)		
f	1	2	3
5	2.57	3.03	3.29
6	2.45	2.86	3.10
7	2.36	2.75	2.97
8	2.31	2.67	2.88
9	2.26	2.61	2.81
10	2.23	2.57	2.76
11	2.20	2.53	2.72
12	2.18	2.50	2.68
13	2.16	2.48	2.65
14	2.14	2.46	2.63
15	2.13	2.44	2.61
16	2.12	2.42	2.59

Determining sample size

In any experimental design problem, a critical decision is the choice of sample size—that is, determining the number of replicates to run. Generally, if the experimenter is interested in detecting small effects, more replicates are required than if the experimenter is interested in detecting large effects.

Here we discuss the approach using OC curve

An operating characteristic (OC) curve is a plot of the type II error probability of a statistical test for a particular sample size versus a parameter that reflects the extent to which the null hypothesis is false. These curves can be used to guide the experimenter in selecting the number of replicates so that the design will be sensitive to important potential differences in the treatments.

We consider the probability of type II error of the fixed effects model for the case of equal sample sizes per treatment, say

$$\begin{aligned}\beta &= 1 - P\{\text{Reject } H_0 | H_0 \text{ is false}\} \\ &= 1 - P\{F_0 > F_{\alpha,a-1,N-a} | H_0 \text{ is false}\}\end{aligned}$$

To evaluate the probability statement in Equation 3.43 (mentioned in the last slide), we need to know the distribution of the test statistic F_0 if the null hypothesis is false. It can be shown that, if H0 is false, the statistic $F_0 = MS_{Treatments}/MS_E$ is distributed as a noncentral F random variable with a - 1 and N - a degrees of freedom and the noncentrality parameter δ. If δ = 0, the noncentral F distribution becomes the usual (central) F distribution.

Operating characteristic curves given in Chart V of the Appendix are used to evaluate the probability statement in Equation 3.43. These curves plot the probability of type II error (β) against a parameter Φ, where

$$\Phi^2 = \frac{n\sum_{i=1}^{a}\tau_i^2}{a\sigma^2} \qquad (3.44)$$

The quantity Φ^2 is related to the noncentrality parameter δ. Curves are available for $\alpha = 0.05$ and $\alpha = 0.01$ and a range of degrees of freedom for numerator and denominator.

In using the OC curves, the experimenter must specify the parameter Φ and the value of σ^2. This is often difficult to do in practice. One way to determine Φ is to choose the actual values of the treatment means for which we would like to reject the null hypothesis with high probability. Thus, if $\mu_1, \mu_2, \ldots, \mu_a$ are the specified treatment means, we find the τ_i in Equation 3.48 as $\tau_i = \mu_i - \overline{\mu}$, where $\overline{\mu} = (1/a)\Sigma_{i=1}^{a}\mu_i$ is the average of the individual treatment means. The estimate of σ^2 may be available from prior experience, a previous experiment or a preliminary test (as suggested in Chapter 1), or a judgment estimate. When we are uncertain about the value of σ^2, sample sizes could be determined for a range of likely values of σ^2 to study the effect of this parameter on the required sample size before a final choice is made.

A significant problem with this approach to using OC curves is that it is usually difficult to select a set of treatment means on which the sample size decision should be based. An alternate approach is to select a sample size such that if the difference between any two treatment means exceeds a specified value, the null hypothesis should be rejected. If the difference between any two treatment means is as large as D, it can be shown that the minimum value of Φ^2 is

$$\Phi^2 = \frac{nD^2}{2a\sigma^2} \tag{3.45}$$

Because this is a minimum value of Φ^2, the corresponding sample size obtained from the operating characteristic curve is a conservative value; that is, it provides a power at least as great as that specified by the experimenter.

Discovering dispersion effects

We have focused on using the analysis of variance and related methods to determine which factor levels result in differences among treatment or factor level means. It is customary to refer to these effects as location effects.

If there was inequality of variance at the different factor levels, we used transformations to stabilize the variance to improve our inference on the location effects.

In some problems, however, we are interested in discovering whether the different factor levels affect variability; that is, we are interested in discovering potential dispersion effects. This will occur whenever the standard deviation, variance, or some other measure of variability is used as a response variable.

To illustrate these ideas, consider the data in Table 3.15, which resulted from a designed experiment in an aluminum smelter.

■ TABLE 3.15

Data for the Smelting Experiment

Ratio Control Algorithm	Observations					
	1	2	3	4	5	6
1	4.93(0.05)	4.86(0.04)	4.75(0.05)	4.95(0.06)	4.79(0.03)	4.88(0.05)
2	4.85(0.04)	4.91(0.02)	4.79(0.03)	4.85(0.05)	4.75(0.03)	4.85(0.02)
3	4.83(0.09)	4.88(0.13)	4.90(0.11)	4.75(0.15)	4.82(0.08)	4.90(0.12)
4	4.89(0.03)	4.77(0.04)	4.94(0.05)	4.86(0.05)	4.79(0.03)	4.76(0.02)

■ TABLE 3.16

Analysis of Variance for the Natural Logarithm of Pot Noise

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-Value
Ratio control algorithm	6.166	3	2.055	21.96	<0.001
Error	1.872	20	0.094		
Total	8.038	23			

An analysis of variance was performed to determine whether the different ratio control algorithms affect average cell voltage. This revealed that the ratio control algorithm had no location effect; that is, changing the ratio control algorithms does not change the average cell voltage.

An analysis of variance was performed to determine whether the different ratio control algorithms affect average cell voltage. This revealed that the ratio control algorithm had no **location effect**; that is, changing the ratio control algorithms does not change the average cell voltage. (Refer to Problem 3.38.)

To investigate dispersion effects, it is usually best to use

$$\log(s) \quad \text{or} \quad \log(s^2)$$

as a response variable since the log transformation is effective in stabilizing variability in the distribution of the sample standard deviation. Because all sample standard deviations of pot voltage are less than unity, we will use

$$y = -\ln(s)$$

as the response variable. Table 3.16 presents the analysis of variance for this response, the natural logarithm of "pot noise." Notice that the choice of a ratio control algorithm affects pot noise; that is, the ratio control algorithm has a **dispersion effect**. Standard tests of model adequacy, including normal probability plots of the residuals, indicate that there are no problems with experimental validity.

The Random Effects Model

An experimenter is frequently interested in a factor that has a large number of possible levels. If the experimenter randomly selects a of these levels from the population of factor levels, then we say that the factor is random.

Because the levels of the factor actually used in the experiment were chosen randomly, inferences are made about the entire population of factor levels. We assume that the population of factor levels is either of infinite size or is large enough to be considered infinite.

Situations in which the population of factor levels is small enough to employ a finite population approach are not encountered frequently.

The linear statistical model is

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases} \qquad (3.47)$$

where both the treatment effects τ_i and ϵ_{ij} are random variables.

The variances σ_τ^2 and σ^2 are called **variance components**, and the model (Equation 3.47) is called the **components of variance** or **random effects model**. The observations in the random effects model are normally distributed because they are linear combinations of the two normally and independently distributed random variables τ_i and ε_{ij}. However, unlike the fixed effects case in which all of the observations y_{ij} are independent, in the random model the observations y_{ij} are only independent if they come from different factor levels. Specifically, we can show that the covariance of any two observations is

$$Cov\left(y_{ij}, y_{ij'}\right) = \sigma_\tau^2 \qquad j \neq j'$$

$$Cov\left(y_{ij}, y_{i'j'}\right) = 0 \qquad i \neq i'$$

Analysis of Variance for the Random Model

The basic ANOVA sum of squares identity

$$SS_T = SS_{\text{Treatments}} + SS_E \tag{3.48}$$

is still valid. That is, we partition the total variability in the observations into a component that measures the variation between treatments ($SS_{\text{Treatments}}$) and a component that measures the variation within treatments (SS_E). Testing hypotheses about individual treatment effects is not very meaningful because they were selected randomly, we are more interested in the **population** of treatments, so we test hypotheses about the variance component σ_τ^2.

$$\begin{aligned} H_0&: \sigma_\tau^2 = 0 \\ H_1&: \sigma_\tau^2 > 0 \end{aligned} \tag{3.49}$$

If $\sigma_\tau^2 = 0$, all treatments are identical; but if $\sigma_\tau^2 = 0$, variability exists between treatments. As before, SS_E/σ^2 is distributed as chi-square with $N - a$ degrees of freedom and, under the null hypothesis, $SS_{\text{Treatments}}/\sigma^2$ is distributed as chi-square with $a - 1$ degrees of freedom. Both random variables are independent. Thus, under the null hypothesis $\sigma_\tau^2 = 0$, the ratio

$$F_0 = \frac{\dfrac{SS_{\text{Treatments}}}{a-1}}{\dfrac{SS_E}{N-a}} = \frac{MS_{\text{Treatments}}}{MS_E} \tag{3.50}$$

is distributed as F with $a - 1$ and $N - a$ degrees of freedom.

The expected values of mean squares are:

$$E(MS_{\text{Treatments}}) = \sigma^2 + n\sigma_\tau^2$$

$$E(MS_E) = \sigma^2$$

From the expected mean squares, we see that under H_0 both the numerator and denominator of the test statistic (Equation 3.50) are unbiased estimators of σ^2, whereas under H_1 the expected value of the numerator is greater than the expected value of the denominator. Therefore, we should reject H_0 for values of F_0 that are too large. This implies an upper-tail, one-tail critical region, so we reject H_0 if $F_0 > F_{\alpha,a-1,N-a}$.

Estimating the Model Parameters

The parameters can be estimated as:

$$\hat{\sigma}^2 = MS_E$$

$$\hat{\sigma}_\tau^2 = \frac{MS_{\text{Treatments}} - MS_E}{n}$$

An example is discussed next

Example 3.11

A textile company weaves a fabric on a large number of looms. It would like the looms to be homogeneous so that it obtains a fabric of uniform strength. The process engineer suspects that, in addition to the usual variation in strength within samples of fabric from the same loom, there may also be significant variations in strength between looms. To investigate this, she selects four looms at random and makes four strength determinations on the fabric manufactured on each loom. This experiment is run in random order, and the data obtained are shown in Table 3.17. The ANOVA is conducted and is shown in Table 3.18. From the ANOVA, we conclude that the looms in the plant differ significantly.

■ **TABLE 3.17**
Strength Data for Example 3.11

	Observations				
Looms	**1**	**2**	**3**	**4**	$y_{i\cdot}$
1	98	97	99	96	390
2	91	90	93	92	366
3	96	95	97	95	383
4	95	96	99	98	388
					$1527 = y_{\cdot\cdot}$

Solution

The variance components are estimated by $\hat{\sigma}^2 = 1.90$ and

$$\hat{\sigma}_\tau^2 = \frac{29.73 - 1.90}{4} = 6.96$$

Therefore, the variance of any observation on strength is estimated by

$$\hat{\sigma}_y = \hat{\sigma}^2 + \hat{\sigma}_\tau^2 = 1.90 + 6.96 = 8.86.$$

Most of this variability is attributable to differences *between* looms.

■ **TABLE 3.18**
Analysis of Variance for the Strength Data

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	*P*-Value
Looms	89.19	3	29.73	15.68	<0.001
Error	22.75	12	1.90		
Total	111.94	15			

Estimation of the Overall Mean μ.

In many random effects experiments the experimenter is interested in estimating the overall mean . From the basic model assumptions it is easy to see that the expected value of any observation is just the overall mean. Consequently, an unbiased estimator of the overall mean is

$$\hat{\mu} = \bar{y}_{..}$$

So for Example 3.11 the estimate of the overall mean strength is

$$\hat{\mu} = \bar{y}_{..} = \frac{y_{..}}{N} = \frac{1527}{16} = 95.44$$

the $100(1-\alpha)\%$ CI on the overall mean is

$$\bar{y}_{..} - t_{\alpha/2,\, a(n-1)}\sqrt{\frac{MS_{\text{Treatments}}}{an}} \le \mu \le \bar{y}_{..} + t_{\alpha/2,\, a(n-1)}\sqrt{\frac{MS_{\text{Treatments}}}{an}} \qquad \textbf{(3.61)}$$

Nonparametric Methods in the Analysis of Variance

In situations where the normality assumption is unjustified, the experimenter may wish to use an alternative procedure to the F test analysis of variance that does not depend on this assumption.

Such a procedure has been developed by Kruskal and Wallis (1952). The Kruskal–Wallis test is used to test the null hypothesis that the a treatments are identical against the alternative hypothesis that some of the treatments generate observations that are larger than others.

Because the procedure is designed to be sensitive for testing differences in means, it is sometimes convenient to think of the Kruskal–Wallis test as a test for equality of treatment means. The Kruskal–Wallis test is a nonparametric alternative to the usual analysis of variance.

To perform a Kruskal–Wallis test, first rank the observations y_{ij} in ascending order and replace each observation by its rank, say R_{ij}, with the smallest observation having rank 1. In the case of ties (observations having the same value), assign the average rank to each of the tied observations. Let $R_{i.}$ be the sum of the ranks in the ith treatment. The test statistic is

$$H = \frac{1}{S^2}\left[\sum_{i=1}^{a}\frac{R_{i.}^2}{n_i} - \frac{N(N+1)^2}{4}\right] \tag{3.67}$$

where n_i is the number of observations in the ith treatment, N is the total number of observations, and

$$S^2 = \frac{1}{N-1}\left[\sum_{i=1}^{a}\sum_{j=1}^{n_i} R_{ij}^2 - \frac{N(N+1)^2}{4}\right] \tag{3.68}$$

Note that S^2 is just the variance of the ranks. If there are no ties, $S^2 = N(N+1)/12$ and the test statistic simplifies to

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{a}\frac{R_{i.}^2}{n_i} - 3(N+1)$$

When the number of ties is moderate, there will be little difference between Equations 3.68 and 3.69, and the simpler form (Equation 3.69) may be used. If the n_i are reasonably large, say $n_i \geq 5$, H is distributed approximately as χ^2_{a-1} under the null hypothesis. Therefore, if

$$H > \chi^2_{\alpha,a-1}$$

the null hypothesis is rejected. The P-value approach could also be used.

An example

The data from Example 3.1 and their corresponding ranks are shown in Table 3.20. There are ties, so we use Equation 3.67 as the test statistic. From Equation 3.67

$$S^2 = \frac{1}{19}\left[2869.50 - \frac{20(21)^2}{4}\right] = 34.97$$

and the test statistic is

$$H = \frac{1}{S^2}\left[\sum_{i=1}^{a}\frac{R_{i.}^2}{n_i} - \frac{N(N+1)^2}{4}\right]$$

$$= \frac{1}{34.97}[2796.30 - 2205]$$

$$= 16.91$$

Data and Ranks for the Plasma Etching Experiment in Example 3.1

Power							
160		180		200		220	
y_{1j}	R_{1j}	y_{2j}	R_{2j}	y_{3j}	R_{3j}	y_{4j}	R_{4j}
575	6	565	4	600	10	725	20
542	3	593	9	651	15	700	17
530	1	590	8	610	11.5	715	19
539	2	579	7	637	14	685	16
570	5	610	11.5	629	13	710	18
$R_{i.}$	17		39.5		63.5		90

Because $H > \chi^2_{0.01,3} = 11.34$, we would reject the null hypothesis and conclude that the treatments differ. (The P-value for $H = 16.91$ is $P = 7.38 \times 10^{-4}$.) This is the same conclusion as given by the usual analysis of variance F test.