CHAPTER 1

Introduction

The term *tribology*, meaning the *science* and *technology of friction, lubrication*, and *wear*, is of recent origin (Lubrication Engineering Working Group, 1966), but its practical aspects reach back to prehistoric times. The importance of tribology has greatly increased during its long history, and modern civilization is surprisingly dependent on sound tribological practices.

The field of tribology affects the performance and life of all mechanical systems and provides for reliability, accuracy, and precision of many. Tribology is frequently the pacing item in the design of new mechanical systems. Energy loss through friction in tribo-elements is a major factor in limits on energy efficiency. Strategic materials are used in many tribo-elements to obtain the required performance.

Experts estimate that in 1978 over 4.22×10^6 Tjoule (or four quadrillion Btu) of energy were lost in the United States due to simple friction and wear – enough energy to supply New York City for an entire year (Dake, Russell, and Debrodt, 1986). This translates to a $20 billion loss, based on oil prices of about $30 per barrel. Most frictional loss occurs in the chemical and the primary metal industries. The metalworking industry's share of tribological losses amount to 2.95×10^4 Tjoule in friction and 8.13×10^3 Tjoule in wear; it has been estimated that more than a quarter of this loss could be prevented by using surface modification technologies to reduce friction and wear in metal working machines. The unsurpassed leader in loss due to wear is mining, followed by agriculture.

1.1 Historical Background

There is little evidence of tribological practices in the early Stone Age. Nevertheless, we may speculate that the first fires made by humans were created by using the heat of friction. In later times hand- or mouth-held bearings were developed for the spindles of drills, which were used to bore holes and start fires. These bearings were often made of wood, antlers, or bone; their recorded use covers some four millennia. Among the earliest-made bearings were door sockets, first constructed of wood or stone and later lined with copper, and potter's wheels, such as the one unearthed in Jericho, dated 2000 BC. The wheel contained traces of bitumen, which might have been used as a lubricant.

Lubricants were probably used on the bearings of chariots, which first appeared ca. 3500 BC (McNeill, 1963). One of the earliest recorded uses of a lubricant, probably water, was for transportation of the statue of Ti ca. 2400 BC. Considerable development in tribology occurred in Greece and Rome beginning in the fourth century BC, during and after the time of Aristotle. Evidence of advanced lubrication practices during Roman times is provided by two pleasure boats that sank in Lake Nemi, Italy, ca. AD 50; they contain what might be considered prototypes of three kinds of modern rolling-element bearings. The Middle Ages saw a further improvement in the application of tribological principles, as evidenced by the development of machinery such as the water mill. An excellent account

of the history of tribology up to the time of Columbus is given by Dowson (1973). See also Dowson's *History of Tribology* (Dowson, 1979).

The basic laws of friction were first deduced correctly by da Vinci (1519), who was interested in the music made by the friction of the heavenly spheres. They were rediscovered in 1699 by Amontons, whose observations were verified by Coulomb in 1785. Coulomb was able to distinguish between static friction and kinetic friction but thought incorrectly that friction was due only to the interlocking of surface asperities. It is now known that friction is caused by a variety of surface interactions. These surface interactions are so complex, however, that the friction coefficient in dry sliding still cannot be predicted.

The scientific study of lubrication began with Rayleigh, who, together with Stokes, discussed the feasibility of a theoretical treatment of film lubrication. Reynolds (1886) went even further; he detailed the theory of lubrication and discussed the importance of boundary conditions. Notable subsequent work was done by Sommerfeld and Michell, among others. However, for many years the difficulty of obtaining two-dimensional solutions to Reynolds' pressure equations impeded the application of lubrication theory to bearing design. This impediment was finally removed with the arrival of the digital computer (Raimondi and Boyd, 1958).

In contrast to friction, the scientific study of wear is more recent. As sliding wear, a term often used to define progressive removal of material due to relative motion at the surface, is caused by the same type of interaction as friction, the quantitative prediction of wear rate is fraught with the same difficulties. The situation is even more gloomy, as under normal conditions the value of the coefficient of friction between different metal pairs changes by one order of magnitude at most, while corresponding wear rates can change by several orders. Although there have been attempts to predict wear rate, Archard's formula (Archard, 1953) being perhaps the most noteworthy in this direction, for the foreseeable future at least, the designer will have to rely on experimentation and handbook data (see Peterson and Winer, 1980).

1.2 Tribological Surfaces

Even early attempts to develop a theory of friction recognized the fact that all practically prepared surfaces are rough on the microscopic scale. The aspect ratio and the absolute height of the hills, or *asperities*, and valleys one observes under the microscope vary greatly, depending on material properties and on the method of surface preparation. Roughness height may range from 0.05 μm or less on polished surfaces to 10 μm on medium-machined surfaces, to even greater values on castings. Figure 1.1 shows a size comparison of the various surface phenomena of interest in tribology.

When two solid surfaces are brought into close proximity, actual contact will be made only by the asperities of the two surfaces, specifically along areas over which the atoms of one asperity surface are within the repulsive fields of the other.[1] The *real area of contact* A_r, which is the totality of the individual asperity contact areas, is only a fraction of the *apparent*

[1] The equilibrium spacing of atoms is on the order of 0.2–0.5 nm (2–5 Angstrom); at distances less than the equilibrium spacing, the repulsive forces dominate, while at greater distances the forces of attraction are influential. The equilibrium spacing changes with temperature; macroscopically we recognize this change as thermal expansion.

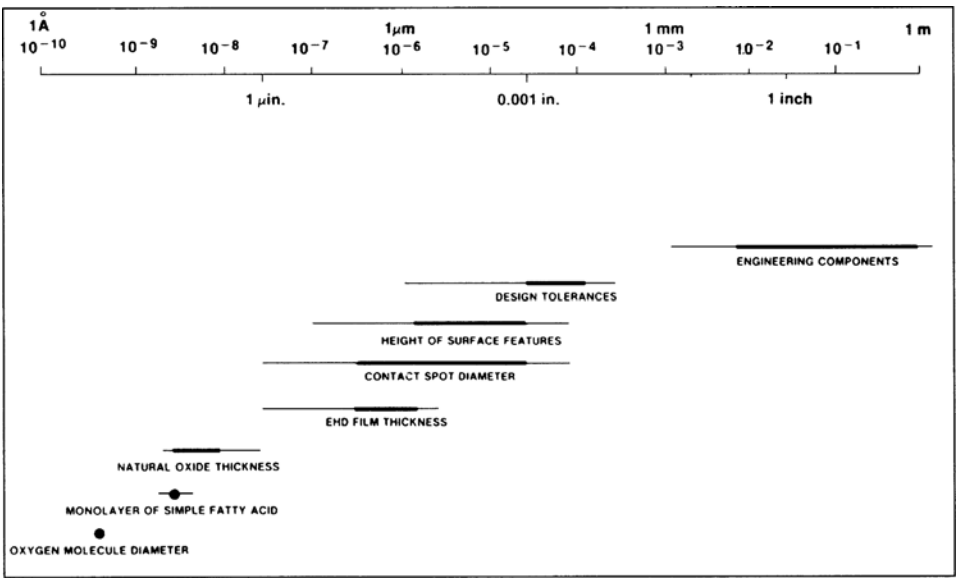

Figure 1.1. Comparative size of surface-related phenomena. (Reprinted with permission from Williamson, J. B. P. The shape of surfaces. In Booser, E. R. *CRC Handbook of Lubrication*. Copyright CRC Press, Boca Raton, Florida, © 1984.)

area of contact, perhaps as small as 1/100,000 at light loads. The areas of individual asperity contacts are typically 1 to 5 μm across and 10 to 50 μm apart.

The topography of engineering surfaces indicates features of four different length scales: (1) *error of form* is a gross deviation from shape of the machine element, (2) *waviness* is of a smaller scale and may result from heat treatment or from vibration of the workpiece or the tool during machining, (3) *roughness* represents closely spaced irregularities and includes features that are intrinsic to the process that created the surface, and (4) surface features on the *atomic scale* are important for the recording industry and in precision machining.

One of the methods used for describing surface roughness consists of drawing a fine stylus across it. The stylus is usually a conical diamond with a radius of curvature at its tip of the order of 2 μm. The movement of the stylus is amplified, and both vertical and horizontal movements are recorded electronically for subsequent statistical analysis. The instrument designed to accomplish this is the *profilometer*. Clearly, such an instrument is limited in resolution by the diameter and the radius of curvature of the tip of the stylus. A profilometer trace[2] of an engineering surface is shown in Figure 1.2.

Two modern instruments, the scanning *electron microscope* and the *transmission electron microscope* (Sherrington and Smith, 1988), have resolution higher than profilometers and are employed extensively in surface studies. *Optical interferometers*, which can record surface profiles without distortion or damage, have recently come into use thanks to advances

[2]That the vertical amplification is typically 10–1000 times greater than the horizontal one has led to the popular misconception that engineering surfaces support steep gradients. Machined surfaces have aspect ratios normally found in the topography of the Earth, the slopes rarely exceeding 5–10°; Figure 1.2 is a distortion of this.

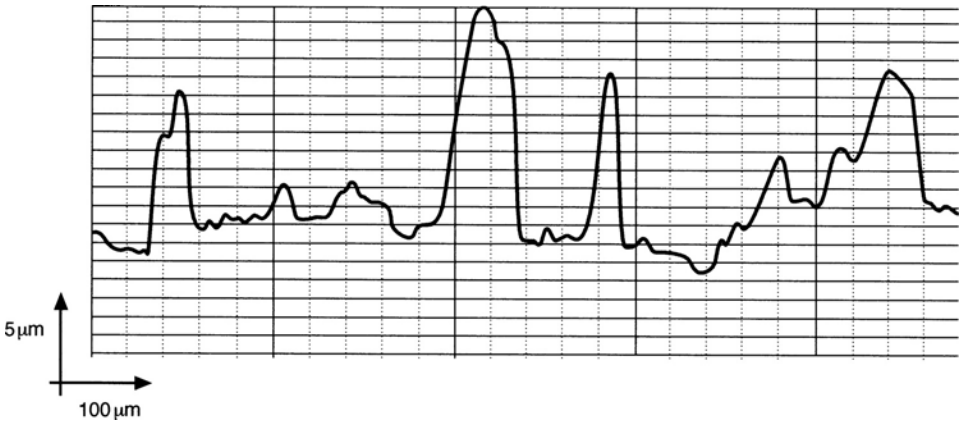

Figure 1.2. Profilometer trace of a rolled metal specimen. The vertical magnification is 20 times the horizontal magnification.

in microprocessors. Vertical resolution of the order of 1 nm has been achieved by optical interferometers, although the maximum measurable height is somewhat limited by the depth of focus of these instruments (Bhushan, Wyant, and Meiling, 1988). The *atomic force microscope* measures the forces between a probe tip and the surface and has been used for topographical measurement of surfaces on the nanometers scale. Its modification, known as the *friction force microscope* (Ruan and Bhushan, 1994) is used for friction studies on the atomic scale. Details of these recent additions to the arsenal of the surface scientist can be found in the excellent review article by Bhushan, Israelachvili, and Landman (1995).

To discuss surface roughness quantitatively, let $\xi(x)$ represent the height of the surface above an arbitrary datum at the position x, and let $\bar{\xi}$ be its mean value as depicted in Figure 1.3. Furthermore, denote by $|\eta(x)|$ the vertical distance between the actual surface at x and the mean. Surface roughness is often characterized in terms of the *arithmetic average*, R_a, of the absolute value of surface deviations from the mean

$$R_a = \frac{1}{L}\int_{-L/2}^{L/2} |\eta(x)|dx, \tag{1.1}$$

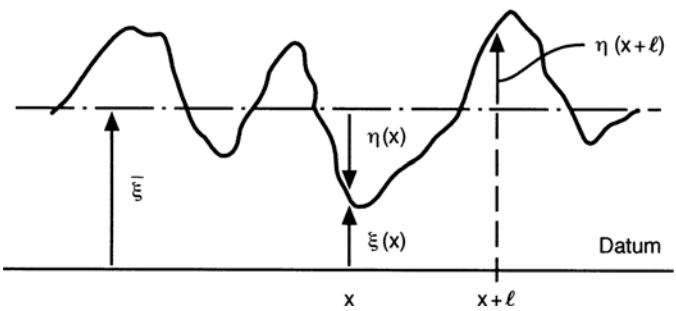

Figure 1.3. Schematics of a surface showing mean surface height, $\bar{\xi}$, and surface deviation from mean height, $\eta(x)$.

or in terms of its *standard deviation* [i.e., root mean square (rms)], R_q, defined by

$$R_q^2 = \frac{1}{L}\int_{-L/2}^{L/2} \eta^2(x)\,dx. \tag{1.2}$$

where L is the sample length.

The rms value, Eq. (1.2), is some 10–20% greater than the R_a value for many common surfaces; for surfaces with Gaussian distribution $R_q = 1.25R_a$. Typical values of R_a for metals prepared by various machining methods are: turned, 1–6 μm; course ground, 0.5–3 μm; fine ground, 0.1–0.5 μm; polished, 0.06–0.1 μm; and super finished, 0.01–0.06 μm.

Another quantity used in characterizing surfaces is the *autocorrelation function,* $R(\ell)$, it has the definition (see Figure 1.3)

$$R(\ell) = \frac{1}{L}\int_{-L/2}^{L/2} \eta(x)\eta(x+\ell)\,dx. \tag{1.3}$$

$R(\ell)$ attains its maximum value at $\ell = 0$, equal to ${R_q}^2$, then vanishes rapidly as ℓ is increased. Its normalized value, $r(\ell) = R(\ell)/R_q^2$, is called the *autocorrelation coefficient*. Peklenik (1968) analyzed surfaces that were produced by different machining techniques and proposed a surface classification based on the shape of the correlation function and the magnitude of the correlation length $\lambda_{0.5}$, defined by $R(\lambda_{0.5}) = 0.5$.

The Fourier cosine transform, $P(\omega)$, of the autocorrelation function

$$P(\omega) = \frac{2}{\pi}\int_0^\infty R(\ell)\cos(\omega\ell)\,d\ell, \tag{1.4}$$

is a quantity particularly suitable to the study of machined surfaces (see Figure 1.5), since it clearly depicts and separates strong surface periodicities that may result from the machining process (i.e., waviness).

There are other numerical characteristics of surfaces in use; to define these we make recourse to probability theory. To this end consider the random *variable* ξ, representing the height of the surface at some position x relative to an arbitrary datum, and examine the event $\xi < y$, signifying that the random variable ξ has a value less than the number y. The probability of this event occurring, designated by $P(\xi < y)$, is a function of y. Define the *integral distribution function* by $F(y) = P(\xi < y)$, then $F(-\infty) = 0$, $F(+\infty) = 1$ and $0 \le F(y) \le 1$. The random variable ξ is considered known if its integral distribution, $F(y)$, is given.

For any two numbers y_2 and y_1, where $y_2 > y_1$, the probability of the event $\xi < y_2$ is given by the sum of the probabilities that $\xi < y_1$ and $y_1 \le \xi < y_2$ or

$$\begin{aligned} P(\xi < y_2) &= P(\xi < y_1 \quad \text{or} \quad y_1 \le \xi < y_2) \\ &= P(\xi < y_1) + P(y_1 \le \xi < y_2). \end{aligned} \tag{1.5}$$

From Eq. (1.5) we find that

$$\begin{aligned} P(y_1 \le \xi < y_2) &= P(\xi < y_2) - P(\xi < y_1) \\ &= F(y_2) - F(y_1), \end{aligned} \tag{1.6}$$

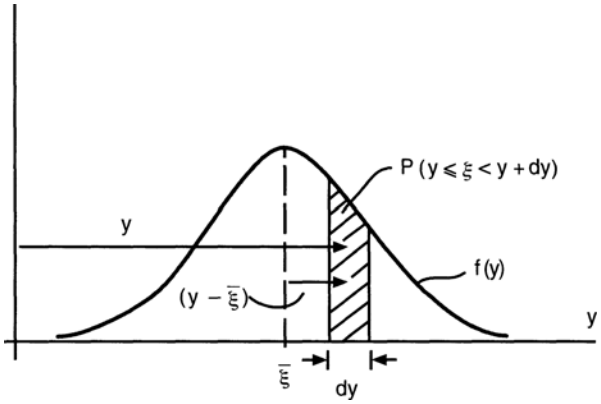

Figure 1.4. Illustration of the probabilistic terminology used.

In the case of a continuous random quantity the distribution function is differentiable. Define the *probability density function* or *probability distribution* by

$$f(y) = \lim_{\Delta y \to 0} \frac{F(y + \Delta y) - F(y)}{\Delta y}. \tag{1.7}$$

From here we can show that the probability that the random variable ξ has a value between y and $y + dy$ is

$$P(y \leq \xi < y + dy) = F(y + dy) - F(y) = f(y)\,dy,$$

and that the probability that ξ is located between the numbers a and b is

$$P(a \leq \xi < b) = \int_a^b f(y)\,dy.$$

Instead of the probability density function $f(y)$ itself, its various moments are often employed. The first *initial moment*, given by

$$\bar{\xi} = \int_{-\infty}^{\infty} y f(y)\,dy, \tag{1.8}$$

is the *mean value* of the random variable ξ (Figure 1.4). It is equivalent to R_a of Eq. (1.1).

The fluctuation about the mean can now be defined by $\eta = \xi - \bar{\xi}$; this is the (random) quantity appearing in Eqs. (1.1) and (1.2).

The first *central moment*, i.e., the moment about the mean, of the probability density function is zero. Its second central moment

$$\sigma^2 = \int_{-\infty}^{\infty} (y - \bar{\xi})^2 f(y)\,dy \tag{1.9}$$

is nonnegative, and it is called the *variance* of the random variable ξ. The square root of the variance is termed the *standard deviation* and is equivalent to the rms. of the deviation from the mean, $\sigma = R_q$.

Many variables that express the results of physical, biological, or medical experiments are, at least to first approximation, distributed according to

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-(y - \bar{\xi})^2/2\sigma^2], \tag{1.10}$$

the so-called *normal* or *Gaussian* distribution. For this reason, the normal distribution has played an important role in the development of statistical theory, and one frequently

encounters Eq. (1.10) in applications. We note from Eq. (1.10) that if the random variable ξ is normally distributed, it is characterized completely by its mean value $\bar{\xi}$ and its standard deviation σ. The simplicity in representation this affords is the reason why there is often great compulsion to declare a distribution Gaussian even though it may deviate from Eq. (1.10).

Other statistical quantities in use for surface characterization are the third and fourth (nondimensional) central moments, the *skewness, Sk*, and the *kurtosis* or "hump," *K*, respectively

$$Sk = \frac{1}{\sigma^3}\int_{-\infty}^{\infty}(y-\bar{\xi})^3 f(y)\,dy, \quad K = \frac{1}{\sigma^4}\int_{-\infty}^{\infty}(y-\bar{\xi})^4 f(y)\,dy. \tag{1.11}$$

Both Sk and K are dimensionless numbers; $Sk = 0$ indicates perfect symmetry, while K is small for a flat, broad distribution. For normal distribution $Sk = 0$ and $K = 3$.

There are many ways to statistically characterize surface roughness. Which of the characterizations is best is application dictated.

It has been shown recently (Sayles and Thomas, 1978) that the value of the various averages defined here changes with the sampling length L, i.e., surface roughness is a nonstationary random function of position. It is then more amenable to treatment by fractal methods (Majumdar and Bhushan, 1990; Wang and Komvopoulos, 1994).

Figure 1.5 shows statistical characteristics of some machined surfaces:

Manufacturing processes	R_a (μm)	σ (μm)	Sk	K	Peak to valley height (μm)	Figure
Shaping, fine	8.0	11.0	0	2.8	47.0	1.5 (a)
Milling	2.3	2.7	+0.22	2.4	13.0	1.5 (b)
Surface grinding	1.0	1.3	+0.17	3.1	15.0	1.5 (c)
Superfinish	0.18	0.25	+0.32	5.9	1.6	1.5 (d)

The asperity-height distribution of many engineering surfaces is approximately Gaussian. Several surface-finishing processes, such as bead-blasting, which are the cumulative result of a large number of random happenings, will encourage a Gaussian distribution.[3] Other processes, including wear, will destroy it. Figure 1.6 follows such a process. A mild steel pad lubricated with SAE-20 oil was worn against a finely ground hard steel flat (N.B., when plotted on probability paper, the Gaussian distribution appears as a straight line).

1.3 Friction

If two solid bodies, in direct or indirect surface contact, are made to slide relative to one another there is always a resistance to the motion called friction. Friction is beneficial in many instances, and we may even try to increase it. However, in other cases friction is energy consuming, and we endeavor to decrease it, although it may never be eliminated entirely.

[3]Let the n random variables $\xi_1, \ldots, \xi_n$ be independent. Then the *central limit theorem* asserts, under very general conditions, that in the limit as $n \to \infty$ the standardized sum $(\xi - \bar{\xi})/\sigma$ approaches Gaussian distribution (Cramer, 1955). Here $\bar{\xi} = \bar{\xi}_1 + \cdots + \bar{\xi}_n$, $\sigma^2 = \sigma_1^2 + \cdots + \sigma_n^2$ and $\xi = \xi_1 + \cdots + \xi_n$.

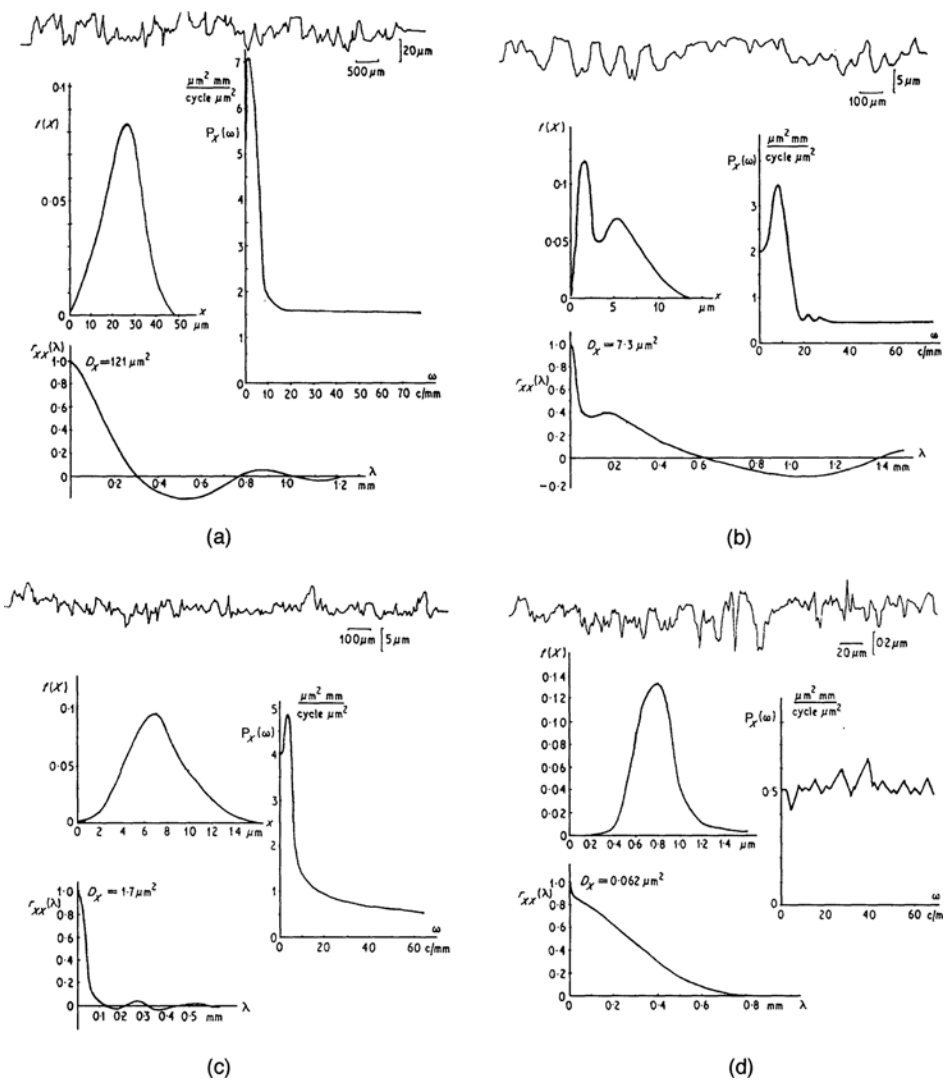

Figure 1.5. Examples of engineering surfaces (a) fine shaped; (b) milled; (c) surface ground (d) superfinished: their distributions, autocorrelation functions, power spectra. (Reprinted by permission of the Council of the Institution of Mechanical Engineers from Peklenik, J. New developments in surface characterization and measurements by means of random process analysis, *Proc. Inst. Mech. Engrs.* **182**, Pt. 3K, 108–126, 1968.)

Friction is present in all machinery, and it converts part of the useful kinetic energy to heat, thus decreasing the overall efficiency of the machine. About 30% of the power in an automobile (Hershey, 1966) and about 1.5% in a modern turbojet engine is wasted through friction. The two journal bearings of a large generator dissipate perhaps 0.75 MW or more. In 1951, G. Vogelpohl estimated that one-third to one-half of the world's energy production is consumed by friction (Fuller, 1956). Not all friction is undesirable, however, and in numerous instances we promote it, e.g., in brakes.

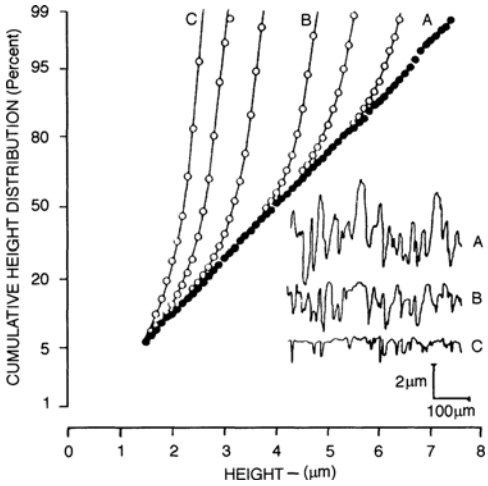

Figure 1.6. The effect of wear. The initial height distribution (A) and six non-Gaussian distributions (open circles) of a bead blasted surface represent, from right to left, progressive states. Height distributions of this form are typical of those created by stratified secondary preparation processes. (Reprinted with permission from Williamson, J. B. P. The shape of surfaces. In Booser, E. R. *CRC Handbook of Lubrication*. Copyright CRC Press, Boca Raton, Florida, © 1984.)

Laws of Friction

The two basic laws of friction:

1. Friction force F is proportional to the normal force W between surfaces,
2. Friction force is independent of the (apparent) area of contact,

were first deduced by da Vinci (1519) and discussed by Amontons (1699). Coulomb (1785) verified these laws experimentally.[4] Coulomb's observation that "kinetic friction is nearly independent of the sliding speed" is at times referred to as the third law of friction. The laws of friction have remained intact for more than 400 years, and even modern experimental research supports them in numerous cases.

This is not true, however, for the origin of friction as discussed by Coulomb. At first Coulomb inclined toward the view that friction is produced by molecular adhesion between the interacting surfaces, which is somewhat in line with present-day theories. Later Coulomb rejected this in favor of the view that friction is produced by interlocking surface asperities. According to this theory, the frictional force is the force required to lift the load over the asperities. Considering that sliding down the asperities releases as much energy as was spent on climbing up, Coulomb's friction is nondissipative, as was first pointed out by Leslie in 1804.

[4]To derive Amontons' laws, we need the assumption that the real area of contact is proportional to the normal load $A_r = qW$, where q is a constant. If now we denote the friction force per unit area by τ, we have for the friction force $F = \tau A_r$, and Amontons' laws follow at once. In the adhesion theory of friction of Bowden and Tabor (1986), the constant q is made equal to the yield pressure p_0.

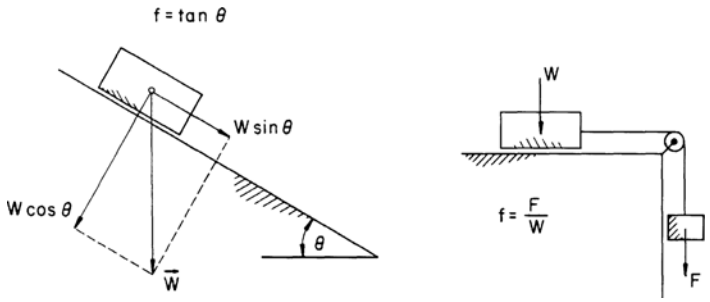

Figure 1.7. Elementary methods of measuring friction.

Most current theories recognize that frictional force in metals arises from three sources: (1) the force necessary to shear *adhesive junctions*, formed at the real area of contact between the asperities; (2) the deformation force, due to the *ploughing* of the asperities of the harder metal through the asperities of the softer one; and (3) *asperity deformation*, which is responsible for the static coefficient of friction – Suh (1986) lists the force required for this as the third source of frictional force. Though these three forces, and the three effects causing them, are not independent, it is customary to treat friction as a result of adhesion interactions, plowing interactions, and asperity deformations. In elastomers, elastic and viscoelastic effects dominate, while in ceramics the type of bonding (ionic in MgO and Al_2O_3 and covalent in TiC, diamond, and SiC) limits plastic flow and the high plastic strains associated with junction growth, at room temperature.

The idea of formation of adhesive junctions (cold welding) over the area of real contact seems frivolous at first, until one considers ultraclean metallic surfaces. When such surfaces are brought together in high vacuum ($P < 10^{-8}$ Pa), the atoms of the real area of contact approach one another across the interface. When they are within 2 nm (20 Angstrom), long distance, weak van der Waals forces are first experienced. As the interfacial distance is decreased to 0.2–0.1 nm, a full metallic bond will form and the pieces weld together. The experiments of Buckley (1977) have been concerned with the force required to overcome this so-called *cold welding*. The adhesive forces are sometimes greater than the forces necessary to press the metals together. However, the metallic bond is completely broken if extended to 0.5 nm, thus a surface film of this thickness signifies that only weak van der Waals forces are acting. As a result, one should expect considerable reduction in adhesive strength. These ideas recently have been confirmed by molecular dynamics simulations (Landman, Luedtke, and Ringer, 1992).

Two elementary methods of measuring static friction, both considered by Leonardo da Vinci, are illustrated in Figure 1.7. Though these methods are quick and convenient, they have had limited success due to the response of the systems being too slow for variations in the coefficient of friction to be detected. Once the body has started moving it will accelerate under constant force, for in general $f_{\text{static}} > f_{\text{kinetic}}$. Even such a variation in friction can hardly be detected by these simple methods. More sophisticated devices for measuring friction are described by Bowden and Tabor (1986), who identify cleanliness of the surface as the single most important factor in achieving repeatable friction results. Surface contaminants, even when present in a layer only one molecule thick, are capable of drastically modifying the friction coefficient because of the reduction in adhesive interactions. Table 1.1 lists f_{static} and f_{dynamic} for various surface pairs under both dry and greasy (lubricated) conditions.