CHAPTER 4 MEASUREMENT

4 MEASUREMENT

> *Until the phenomena of any branch of Knowledge have been submitted to measurement and number it cannot assume the status and dignity of a science. Sir Francis Galton*

Many breakthroughs in science come directly from advances in measurement. Claims that sensory acuity is a sign of intelligence (Wissler, 1901), or that a mother's rejecting personality causes her child's autism (Schreibman & Koegel, 1975), or that prefrontal lobotomy effectively cures mental illness (Valenstein, 1986) are examples of theories that have been rejected after being evaluated using good measures. Improved measures of personality, intelligence, and cognitive abilities often lead to progress in understanding human behavior

To do this, they combined three previously published paper-and-pencil measures into their own custom-made stress index. The measures were: (1) a life event scale, on which participants reported the number of stressful life events that they had experienced in the previous year (events such as divorce or the death of a close friend; (2) a perceived stress scale, on which they reported the extent to which their lives felt overwhelming; and (3) a negative mood scale, measuring the degree to which they felt upset, shaky, irritated, sad, etc. Participants' scores on the combined measure indicated their levels of stress, from high to low. Since each component was known to be reliable, the composite also was expected to be reliable. The validity of the composite was unknown. (Procedures for assessing the reliability and validity of measures are discussed in Chapters 5, Correlation, and 12, Planning the Study.)

Cohen et al. decided to directly expose their high- and low-stressed subjects to disease by giving them nose drops containing an infectious dose of a respiratory virus and then quarantining them in hotel rooms. The subjects were told the nature of the study before volunteering; they knew that the virus was only a cold virus and would do no lasting harm; and they were paid for their participation. Exposing them to a measured dose of a virus, followed by a quarantine, ensured that high- and low-stressed subjects were exposed equally to the disease. The

alternative to this procedure, measuring exposure, would have been very difficult.

Whether or not a subject was infected was determined by blood tests. Subjects were considered to be infected if the virus or antibodies to it could be isolated in their blood. A physician also rated the severity of the subjects' colds.

The researchers also measured a variety of other variables so that if stress and vulnerability to the cold virus were found to be related, alternative explanations of the results could be tested. For example, because smoking affects susceptibility to disease, the amount of the chemical cotinine in the blood was measured; this is an accurate index of how much a person smokes. The participants also took two standard personality scales to check whether the stress index was measuring personality differences rather than stress. How much the subjects exercised, their diets, and the quality of their sleep also were assessed by questionnaire.

All of the measures used in this study are subject to error, but to different degrees. The least error is expected for the measures that participants cannot influence, like the blood analysis. Such measures are called *nonreactive.*

Measures are *nonreactive* if subjects can have no control or influence over their outcomes; that is, if the act of measurement itself cannot result in a reaction from the subjects that could bias the results.

The questionnaires, interviews, and personality scales used in the study, by contrast, were *reactive.*

> Measures are *reactive* if they are made with the subjects' awareness, and if this awareness could lead to a bias in their results.

Even seemingly small modifications in measurement can result in major scientific advances. In his studies of children's intelligence, jean Piaget (1952) switched from the standard measure, counting the number of correct answers on an intelligence test, to studying how children explain their answers, both right and wrong. With this change, Piaget started a line of research that revealed the logic of

children's thinking. B. F. Skinner (1938) observed the rate of bar pressing of pigeons and rats instead of choosing other possible measures of learning. This choice was critical for his discovery of the effects *of* different schedules of reinforcement.

Given the importance of measurement, researchers must plan their measurement strategies with care. Independent and dependent variables must be operationally defined. Variables that the investigator is trying to control may have to be measured to ensure that they remain constant throughout an experiment. Even uncontrolled variables have to be measured if the experimenter plans on using statistical controls. Whether the results of a study are clear-cut or not will depend in large part on the type and quality of its measurements.

4.1 THE VARIETY OF MEASURES IN PSYCHOLOGICAL RESEARCH

Psychologists have at their disposal all the measurement techniques that have been developed in the physical and medical sciences. They can use computer-controlled displays to study perception, radio telemetry to track wild animals, deep-sea sonar to follow whales, and magnetic brain imaging to study brain dysfunction in children with attention deficit hyperactive disorder.

Psychologists also can choose from a vast assortment of psychological tests and observational schemes. *Tests in Print IV* (Murphy, Conoley, & Impara, 1994), a directory of the available commercial tests, lists over 3,000 tests for comparing peoples' aptitudes, abilities, interests, emotions, sensory acuities, personalities, disabilities, attitudes, and disorders. The questions used in surveys often are published along with results so that other researchers can use them. Ethograms profiling the behavioral repertoires of different animals also are available.

Psychological studies often use a combination of measures, some physical, some previously published psychological measures, some custom-made for the study. This blend of instruments is well illustrated in a study by Sheldon Cohen and his colleagues (Cohen, Tyrrell, & Smith, 1991) on the effects of psychological stress on

susceptibility to the common cold. The researchers faced two measurement challenges in their study: (1) how to measure or manipulate stress, and (2) how to measure or manipulate exposure to the cold virus.

Ethically sound procedures for manipulating stress are severely limited and are not likely to result in the high levels of stress that would be needed to affect resistance to disease. Although it might be possible to show subjects a stressful movie or recount a sad story, for example, such brief events most likely would not have a lasting impact. An alternative would be to select participants who already have different levels of stress in their lives. Cohen et al. decided on this procedure. Participants who want to create a favorable impression on investigators, for example, might present themselves on questionnaires and personality scales as less stressed than they actually are. On the other hand, some people may exaggerate their problems to gain an investigator's sympathy. Another problem with reactive measures is that the measurement itself may change the subject, thereby introducing other kinds of error into the study. Filling out the questionnaire on exercise and diet, for example, could suggest the benefits of exercise and a good diet to participants, leading them to change their normal routines.

Whenever possible, nonreactive measures should be used to supplement reactive measures. In the stress study, for example, the physiological measure of smoking, the amount of cotinine in the blood, served as a check on participants' self-reports of the number of cigarettes they smoked. The agreement between the results of these measures confirmed the validity of cotinine as an index of smoking and also helped to establish the self-report measure as valid for future research.

The only problem with using nonreactive measures is that they may be difficult to obtain. To collect blood samples, for instance, requires that a medical professional be on hand; such intrusive procedures may not be suitable for most psychological studies. Other nonreactive measures, such as archival records or covert observation, may violate subjects' rights to privacy. And nonreactive measures simply are not available for many variables. Cohen et al. most likely did not include a nonreactive measure of stress because such measures have not been

developed. Webb, Campbell, Schwartz, Sechrest, and Grove's 1981 book, *Unobtrusive Measures: Nonreactive Research in the Social Sciences,* is a good source of ideas on ways to measure nonreactively.

When nonreactive measures are unavailable, multiple reactive measures should be considered.

> Checking the agreement between multiple measures of the same variable gives investigators a way to evaluate error.

Agreement between the results of different measures establishes that the observations are not uniquely tied to a particular method. In psychotherapy evaluation research, for example, psychologists' ratings of their patients' improvement often are checked against the patients' own ratings. In the stress study, the severity of the participants' colds was rated both by a physician and by the subjects themselves. The close agreement between their ratings provided 2vidence for the validity of both measures.

Cohen et al. found that a higher percentage of high- than low-stressed subjects caught cold after being exposed to the virus. This result was replicated with five types of viruses and shown not to be due to differences in the personalities, diet, exercise, or smoking habits of the high- and low-stressed subjects. The study provides perhaps the best evidence available to date that psychological stress can lower a person's resistance to disease.

The results of this stress study were reported using numerical scales. The stress scale varied from 3, low stress, to 12, high stress; the severity of a subject's cold ranged from 0, no cold, to 4, severe cold; cotinine levels were recorded in parts per unit volume; scores on the personality scales were numbers in the range from 20 to 80; weight was measured in kilograms. Although these scales all involve numerical values, they are not interpreted or analyzed in the same way to interpret scale scores, researchers must know both the *scale type* of the measure and whether the scale construction is based on *standards* or *norms.* We turn first to the distinction between scale types.

4.2 Scales of Measurement

The idea of different types of measurement scales was developed by S. S. Stevens (1946), a psychologist who studied sensation and developed many of the basic sensory scales used today. Stevens argued that there are four basic measurement scales: *ratio, interval, ordinal,* and *nominal,* distinguished from each other by four properties that determine how scores on the scale can be interpreted. The first property is *equality.*

> *A* scale has the property of *equality* if two subjects who are assigned the same score are equal on what is being measured.

If Bob and Jane both are measured as 68 inches tall, then they actually are the same height. The scale of height has equality, the most basic property of a scale; without equality, you do not have a scale at all.

The second property is *rank order.*

> *A* scale has *rank order* if higher scale scores always indicate more of what is being measured.

The scale of height, for example, has rank order since higher numbers, 69 inches, 70 inches, etc., indicate taller people. The numbers assigned to players on a basketball team do not. Player number 23 does not have more of a trait than player number 8. Although qualitative measures, like sex, sometimes are coded numerically, for example, Female 2, Male 1, these "scale scores" also do not have the property of rank order.

The third property is *equal intervals.*

> *A* scale has *equal intervals* if equal-sized differences in scale scores always indicate equal-sized differences in the amount of what is being measured.

For height, the difference between 69 inches and 71 inches, a difference of 2 inches, is the same as the difference in height between

52 and 54 inches. A scale with equal intervals has a constant unit of measurement. The Richter earthquake scale is an example of a common scale that does not have equal intervals. A difference of 1 unit on this scale indicates an increase in the energy of the quake by a multiplication of 10, not the addition of a constant amount of energy. The difference in energy between a 6.0 and a 7.0 earthquake is much greater than the difference between earthquakes of 1.0 and 2.0 on the scale. The fourth and last property is *equal ratios.*

A scale has *equal ratios* if ratios of scores are meaningful.

Height has this property, so two heights can be meaningfully compared by computing their ratio. If Paul is 7 feet tall and Tim is 3.5 feet tall, it is permissible to say that Paul is twice as tall as Tim. By contrast, ratios are not directly interpretable on the Richter scale. An earthquake of 6.0 is not twice as severe as a quake of 3.0; it is 1,000 times more severe.

We have numbered these properties from 1 to 4 to indicate their interrelationship. If a scale has property 4, it also must have all the properties with lower numbers, that is, properties 3,2, and 1. Similarly, if a scale has property 3, it must have properties 2 and 1; and if a scale has property 2, it must have property 1. Because of this structure, Stevens pointed out that the four properties describe only four types of scales. These scales are shown in Table 1 along with their properties.

TABLE 1 SCALES OF MEASUREMENT AND THEIR PROPERTIES

	Property			
	1	2	3	4
Scale	*Equality*	*Rank Order*	*Equal Intervals*	*Equal Ratios*
Ratio	Yes	Yes	Yes	Yes
Interval	Yes	Yes	Yes	No
Ordinal	Yes	Yes	No	No
Nominal	Yes	No	No	No

Ratio scales have all four properties. Many scales in the physical sciences, such as distance, weight, voltage, current, force, are ratio scales. These scales all lave a natural zero point; that is, a score of zero means the absence of the property; zero weight, for example, means literally no weight. Few of today's psychological scales are ratio scales; and the ratio scales we are familiar with, scales of sensations, like the sone scale of noise, were all constructed by Stevens himself. A noise of 10 sones will sound twice as loud to the average person as a noise scaled 5 on this scale.

Interval scales have properties 3, 2, and 1.

The best known interval scales are Fahrenheit and Celsius temperature scales. Because the zero points on these scales do not correspond to the absence of heat, ratios cannot be computed to compare two temperatures; a 100-degree day is not twice as hot as a 50-degree ay, for example. But the scale has a constant unit, so a 5-degree difference, from 0 to 15 degrees, is the same increase in temperature as a 5-degree difference from 95 to 100 degrees. The Kelvin temperature scale, by contrast, does have an absolute zero, — 459.7 degrees Fahrenheit, the temperature at which all molecular motion stops. Ratios on this scale are meaningful; a 100-degree Fahrenheit ay (310.9 °K) is 1.1 times as hot as a 50 degree day (283.1 °K).

Whether a scale has interval properties is verified by experimentation. If an object with a scale score of, say, 5 is "added to" an object with a scale score of 10, for the scale to have the interval property the combination must yield a score of 15. For example, if a 5-pound object is placed on top of a 10-pound object, the combination will weigh 15 pounds. For every interval scale, a process of combining, or adding, two objects to get a new third object must be found so that the new object's scale score is in agreement with arithmetical rules (see Cohen & Nagel, 1934).

Research to verify scale properties is straightforward with measures of weight, length, voltage, etc., but has not been possible for psychological scales, like self-confidence, degree of depression, or intelligence. Consider intelligence; say Bob has an IQ of 80, Jill's 1Q is

60, and Mary's is 140. Is Mary's intelligence equal to the combination of Bob's and Jill's intelligence (80 + 60 = 140)? For this question to be meaningful, there has to be some concrete means to "add" IQ scores. For example, if Bob and Jill took the IQ test together could they get Mary's score? Probably not. Because no one has thought of a way of verifying the interval or ratio properties of such measures, such scales can only provide information about the rank ordering of people.

TABLE 2 BEAUFORT WIND SCALE

—— '—	*Wind Speed*		
Beaufort	*(Km/hr)*	*(mph)*	*Description*
0	below 1	below	Calm
1	1-5	1-3	Light air
2	6-11	4-7	Light breeze
3	12-19	8-12	Gentle breeze
4	20-28	13-18	Moderate
5	29-38	19-24	Fresh breeze
6	39-49	25-31	Strong breeze
7	50-61	32-38	Moderate gale
8	62-74	39-46	Fresh gale
9	75-88	47-54	Strong gale
10	89-102	55-63	Whole gale
11	103-117	64-75	Storm
12	above 117	above 75	Hurricane

From Microsoft Encarta, 1994.

Ordinal or *rank order scales* have properties 2 and 1.

A newspaper may rank order best-selling books from 1 to 10. This scale does not have a constant unit and the ratios of scores are not meaningful. The 5th best-selling book does not sell twice as much as the 10th best-seller, nor are the differences in sales the same between 1st and 2nd place as between 2nd and 3rd place. The Beaufort wind speed scale and the Mohs hardness scale are commonly used ordinal scales. The Beaufort scale, which is presented in Table 2, classifies wind speed into 13 categories from 0, calm, to 12, hurricane. You can see by looking at the, wind speeds that this is not a ratio or interval

scale. The difference between scale scores of 0 and 2,3 mph, is not equal to the difference between scale scores of 10 and 12, 12 mph.

The Mohs hardness scale, presented in Table 3, is based on the operation of scratching one material against another. The harder of the two materials will scratch the other material, but not the other way around. Topaz will scratch quartz, but quartz is not hard enough to scratch topaz. A new material, like your fingernail, is measured by comparing it to these standard 10 materials. A fingernail will scratch gypsum but not calcite, so the scale score for a fingernail is 2.5. This scale does not have ratio or interval properties. It is an ordinal scale.

TABLE 3 MOHS HARDNESS SCALE

Mineral	*Hardness*	*Common Tests*
Talc	1	Scratched
Gypsum	2	by fingernail
Calcite	3	Scratched by copper coin
Fluorite	4	Scratched by a knife blade
Apatite	5	or window glass
Feldspar	6	Scratches a knife
Quartz	7	blade
Topaz	8	or
Corundum	9	window glass
Diamond	10	Scratches all common materials

From Microsoft Encarta, 1994.

The *nominal scale* only has property 1, equality.

If numbers are assigned with these scales (e.g., Democrats = 1; Republicans = 2), they are only for convenience in naming. Nominal scale scores cannot be compared numerically.

The type of scale used in measurement is critical because it limits the type of analysis that is possible on the scale scores. The mean and standard deviation, for example, are not meaningful when computed on scores from nominal and some ordinal scales. If you were to rank order 10 people on intelligence and compute the mean of the

rankings, you would get a mean of 5.5. This number would reflect only the number of subjects studied and would say nothing about the average intelligence of the 10 people. Because nominal scales and many ordinal scales do not have normal distributions, statistics that assume this type of distribution are not recommended for use with them. Specific statistical procedures have been developed for analyzing data for these types of scales (see Liao, 1994).

4.3 Standard and Norm Based Scales

Physical measures are based on standards, objects with known properties that serve as the official definition of the unit for measuring the property. Prescientific standards were the foot, literally a person's foot, the hand width, and the distance from the elbow to the end of the middle finger, the cubit. Today's standards are considerably more precise. The meter is defined as the distance traveled by light in a vacuum in 1/299,792,458 of a second (Wikipedia). The kilogram is defined by a cylinder of platinum-iridium alloy kept in France.

Psychological measures are not based on such standards. There is no person housed in Washington, D.C., who is the standard for "average neuroticism," although most people could nominate an acquaintance for this standard. Psychological measures are *norm-based,* meaning that the score for an individual is interpreted by comparing his/her score with the scores of a group of people who define the norms for the test. A person scores average on an intelligence test whose score is equal to the average of this group of people.

Norm-based measurement is unique to psychology and other social sciences. Although we know of no physical measure that is interpreted with norms, this type of measurement is common in psychology. In developing the Wechsler intelligence scales, for instance, the tests (there are three tests to span the age range from young children to adults) were given to a representative national sample of people of different ages. Their performances set the norms, that is, what is considered a high, average, or low score on the test.

The logic of norm-based tests was developed in the mid 1800s by Sir Francis Galton. Galton also developed the basic statistics for reporting norm based scores—percentiles—and the statistic used in determining

the reliability and validity of measures—the correlation coefficient. In the next section of this chapter, we discuss how Galton made these discoveries and explain the logic of norm-based tests. This logic is essential for understanding modern psychological measurement and analysis.

4.4 The Beginnings of Norm-Based Measurement

Galton got involved in measuring individual differences because such measurement was essential to the success of his scientific program to improve the human race. Galton defined the nature extreme in the nature versus nurture debate—the question of whether differences between people in abilities, attitudes, and other characteristics are due to experience (nurture) or biological inheritance (nature, Galton's position). It was Galton who introduced the word heredity into English (from the French) to refer to the process of biological transmission of traits from parents to offspring.

Given his extreme biological position, it is not surprising that Galton saw educational programs as a waste of time and money. For Galton, social reform required intervention in the process of breeding itself. He coined the word *eugenics* (the science of improving the human race by judicious mating and other means that give more suitable people the advantage in having children) to name such reform. His eugenics program would encourage desirable people to have many children and discourage undesirable people from having any children at all. Then, by virtue of the laws of heredity, the human race would improve generation by generation. To carry out his program, Galton needed to discover the laws of heredity. The then "state of the art" experimental methods of John Stuart Mill were not useful to Galton. Galton's problems in heredity could not be stated in terms of cause and effect. The characteristics of parents do not cause the characteristics of their offspring in any one-to-one manner. Galton needed methods for measuring traits and other methods for dealing with the co-relation or correlation between the traits of parents and offspring. He wanted to know, for example, if parents were intelligent, what that implied about the intelligence of their children. Galton had to invent these methods.

Galton's grandfather had amassed a fortune selling muskets to the British army during the war against Napoleon; the grandfather's factory turned them out at the rate of one per minute. When his father died, Galton inherited his share of the family fortune, which was enough so that he did not have to work again, he quit medical school and devoted his life to science (Galton, 1909). Galton's wealth was the capital that financed the beginning of statistics in the social sciences. (Students today might wish England hadn't needed so many muskets!)

Galton's wealth financed his scientific work on a wide range of topics, including geography and meteorology. After he read *Origin of the Species* in 1859, book his cousin, Charles Darwin, has just published, his interests focused on eugenics. In Darwin's theory, the future of a species is determined by natural election—the survival of the fittest. Galton thought it would be an error to trust le future of the human race to the capriciousness of natural selection. It would be a far better world, he argued, if the future were engineered through eugenics. Galton's work on methods stemmed directly from problems he faced in his eugenics program:

- Galton wanted a way of describing the scores of a large group (population) of people on a trait like intelligence. This description would be the standard of comparison for any future changes on the trait brought about by eugenics.
- He needed a way of describing the degree of change on the trait, so he could find out if future generations were improving compared to the present one. His programs would be evaluated by these changes.
- He needed a quantitative measure of intelligence. This measure was needed to select the people to encourage to breed.
- He needed to identify which traits were biologically determined, because these were the traits that eugenics programs could influence. To do this, he had to find a measure of the degree to which offspring are similar to their parents on a trait.

Galton's eugenics program is history, but his methods are now basic in modern research.

In the early 1860s, Galton began studying the inheritance of intelligence, there were no established measures of intelligence at that time. (The Binet-Simon scale was not available until 1911.) Galton had to find one. He settled for eminence, "high reputation," as his measure. He decided to examine the family trees of eminent men—judges, statesmen, scientists, etc.—to see if eminence "ran in families." The answer was yes. Galton found for judges that:

> More than *one in every nine* of them have been either father, son or brother to another judge. . . . There cannot, then, remain a doubt but that the peculiar type of ability that is necessary to a judge is often transmitted by descent. (Galton, 1864/1892, p. 62)

Galton found similar results for other categories of eminent men. We now know that this result, that eminence runs in families, does not necessarily mean that it is biologically inherited. Families share a common culture as well as common genes. Galton ignored the latter possibility and concluded that intelligence is highly heritable.

4.5 Describing Individual Differences

Galton was happy with the results of his study of eminence, but he was not satisfied with using ratings of "eminent" or "not eminent" to measure intelligence. He wanted to study heredity with *quantitative* measures. His dream was to obtain "exact measurements relating to every measurable faculty of body or mind, for two generations at least" (Galton, 1909, p. 244).

To this end, Galton set up a unique laboratory in London at the International Health Exhibition. For threepence, visitors to his laboratory could take a series of tests and measures, and compare how they did with the results from other visitors. Galton's measurements included hearing and visual acuity; color sense; reaction time; pulling, squeezing, and hitting strength; as well as height, weight, and arm span.

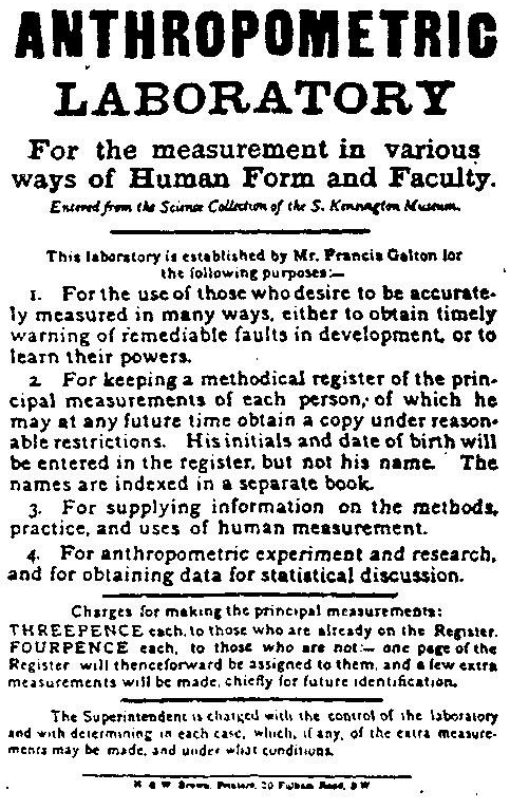
On first reading, it seems that this list of measures did not include the one trait of most interest to Galton—intelligence. Not so. For Galton

had included measures of sensory acuity and reaction time, the measures thought by scientists of his day to be the best indicators of intelligence. Scientists believed that intelligent people were quick to react and, like the fairy-tale princess who could feel a pea through a stack of mattresses, highly sensitive to stimuli. Retarded people were expected to be slow and insensitive.

4.5.1 Percentiles

Galton wanted to let visitors to his exhibit compare themselves to other visitors. To do this, he developed a novel statistical method called *centiles* or *percentiles.*

Percentiles are calculated from a group of scores. First, all the scores are rank ordered from high to low; then they are divided into 100 groups, with an equal number of scores in each group. If 500 people were tested, each of the 100 groups would contain 5 scores. The values dividing these groups are called percentiles and are numbered from 0 to 100.



ANTHROPOMETRIC LABORATORY

For the measurement in various ways of Human Form and Faculty.

Entered from the Science Collection of the S. Kensington Museum.

This laboratory is established by Mr. Francis Galton for the following purposes:—

1. For the use of those who desire to be accurately measured in many ways, either to obtain timely warning of remediable faults in development, or to learn their powers.
2. For keeping a methodical register of the principal measurements of each person, of which he may at any future time obtain a copy under reasonable restrictions. His initials and date of birth will be entered in the register, but not his name. The names are indexed in a separate book.
3. For supplying information on the methods, practice, and uses of human measurement.
4. For anthropometric experiment and research, and for obtaining data for statistical discussion.

Charges for making the principal measurements: THREEPENCE each, to those who are already on the Register. FOURPENCE each, to those who are not:— one page of the Register will thenceforward be assigned to them, and a few extra measurements will be made, chiefly for future identification.

The Superintendent is charged with the control of the laboratory and with determining in each case, which, if any, of the extra measurements may be made, and under what conditions.

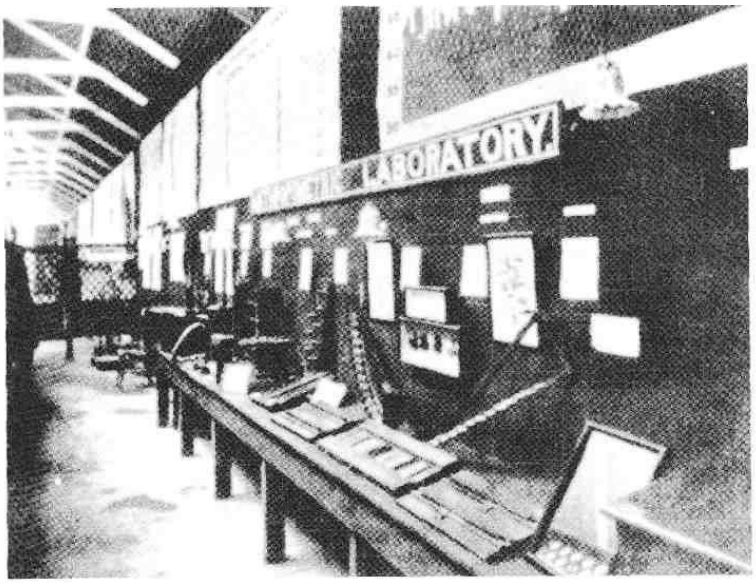
H. & W. Brown, Printers, 20 Fulham Road, S.W.

Poster advertising Galton's Anthropometric Laboratory.

A percentile is one of the values that divide a set of scores into 100 groups of equal frequency. One percent of the scores fall below the value of the 1st percentile, 2% fall below the 2nd percentile, 80% fall below the 80th percentile, and so on.

The 50th percentile, called the *median,* divides the scores into two equal sized groups; 50% are below the median, and 50% above.

For the heights of males, Galton found that the 50th percentile was at 67.9 riches, the 70th percentile at 69.2, and the 90th percentile at 71.3. A visitor who was 69 inches tall would know that about 70% *of* the visitors were shorter and 30% were taller than he was.



Galton's laboratory at the International Health Exhibition.

Table 4 shows the results for some of Galton's measures, collected in 1884. Besides the simplicity of the percentile system, what is striking is the marked difference in the size of people just over 130 years ago. Back then the 50th percentile for men's weights was 143 pounds; today the 50th percentile man weighs 191 pounds!

TABLE 4 PERCENTILES

					Values surpassed by percents, below										
					95%	*90%*	*80%*	*70%*	*60%*	*50%*	*40%*	*30%*	*20%*	*10%*	*5%*
Character measured	*Age*	*Unit of measurement*	*Sex*	*No. of Persons*	*Values unreached by percents, below*										
					5%	*10%*	*20%*	*30%*	*40%*	*50%*	*60%*	*70%*	*80%*	*90%*	*95%*
Height standing without shoes	23–51	Inches	M.	811	63·2	64·5	65·8	66·5	67·3	67·9	68·5	69·2	70·0	71·3	72·4
		"	F.	770	58·9	59·9	61·3	62·1	62·7	63·3	63·9	64·6	65·3	66·4	67·3
Height sitting from seat of chair	23–51	Inches	M.	1013	33·6	34·2	34·9	35·3	35·4	36·0	36·3	36·7	37·1	37·7	38·2
		"	F.	775	31·8	32·3	32·9	33·3	33·6	33·9	34·2	34·6	34·9	35·6	36·0
Span of Arms	23–51	Inches	M.	811	65·0	66·1	67·2	68·2	69·0	69·9	70·6	71·4	72·3	73·6	74·8
		"	F.	770	58·6	59·5	60·7	61·7	62·4	63·0	63·7	64·5	65·4	66·7	68·0
Weight in ordinary indoor clothes	23–26	Pounds	M.	520	121	125	131	135	139	143	147	150	156	165	172
		"	F.	276	102	105	110	114	118	122	129	132	136	142	149
Vital or Breathing Capacity	23–26	Cubic Inches	M.	212	161	177	187	199	211	219	226	236	248	277	290
			F.	277	92	102	113	124	131	138	144	151	164	177	186
Strength of Pull as archer with bow	23–26	Pounds	M.	519	56	60	64	68	71	74	77	80	82	89	96
		"	F.	276	30	32	34	36	38	40	42	44	47	51	54
Strength of Squeeze with strongest hand	23–26	Pounds	M.	519	67	71	76	79	82	85	88	91	95	100	104
		"	F.	276	36	39	43	47	49	52	55	58	62	67	72
Swiftness of Blow	23–26	Feet per second	M.	516	13.2	14.1	15.2	16.2	17.3	18.1	19.1	20.0	20.9	22.3	23.6
			F.	271	9.2	10.1	11.3	12.1	12.8	13.4	14.0	14.5	15.1	16.3	16.9
Keenness of Sight by distance of reading diamond type	23–26	Inches	M.	398	13	17	20	22	23	25	26	28	30	32	34
		"	F.	433	10	12	16	19	22	24	26	27	29	31	32

Values surpassed and values unreached, by various percentages of the persons measured at the Anthropometric Laboratory at the late International Health Exhibition.

No one could have devised a simpler method for conveying a person's relative standing on a measure, and no one has since. Educators today still use Galton's percentiles to report the results of standard tests, such as the SATs; physicians use percentiles to evaluate the height and weight of children; and psychologists use these statistics to describe a person's standing on all kinds of measures of personality and ability.

4.5.2 The Normal Distribution

When the exhibition closed, Galton moved his laboratory to the South Kensington Museum, where he collected data for six more years. He used the data for a variety of projects, including replicating a remarkable discovery that had been made some 30 years earlier by M. A. Quetelet, a Belgian astronomer.

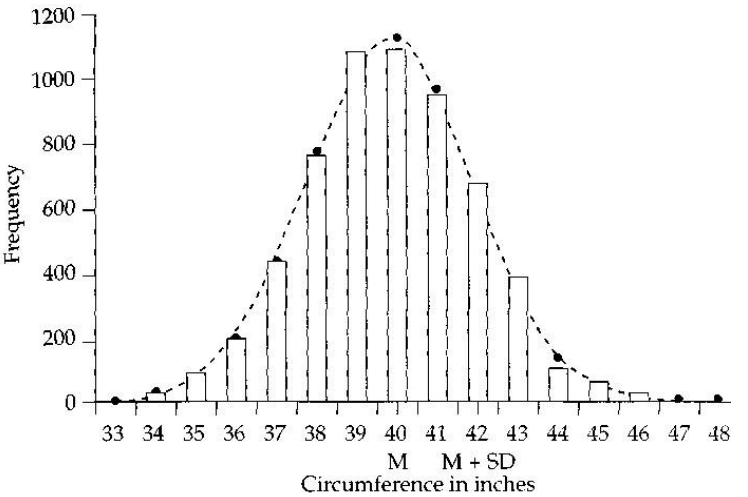


FIGURE 1 Histogram of the circumferences of the chests of Scottish soldiers (based onQuetelet, 1849).

Quetelet discovered that differences between people in height and other physical measures have a simple mathematical form—a form that allows accurate description of the physical characteristics of thousands of people using only two numbers (Quetelet, 1849). In one demonstration Quetelet arranged the chest circumferences of 5,738 army recruits in a special pattern which we now call a *histogram.*

> A *histogram* is a graph showing the frequency of occurrence of different values of a measure.

The graph is made up of a series of rectangles; the width of each rectangle indicates an interval of values on the measure and the height of the rectangles is in proportion to the frequency of cases that have values in that interval of scores.

Quetelet took the continuum of chest measures from 33 to 48 inches and divided it into a series of consecutive categories each with a width of 1 inch. He then placed each of his 5,738 measurements into its appropriate category.

The resulting histogram (shown in Figure 1) had a "bell-curve" shape that was familiar to Quetelet. Astronomers had been using this curve to describe the distribution of errors they made, for example, in locating stars in the heavens. It was called the *normal distribution* or *Gaussian distribution* (after the mathematician Carl Gauss).

> A *normal distribution* is a theoretical frequency distribution that is specified by a mathematical equation. The distribution has the shape of the cross-section of a bell; the high point of the curve is at the median, and the curve is symmetric around the median.

The form of the ideal normal distribution, shown by the dotted line in Figure 1, is described mathematically by a formula that depends upon only two values, the *mean* and *standard deviation* of the measure.

> The *mean* is the average value of a measure. It is computed by adding all the scores and dividing by the number of scores.

For a normal distribution, the mean is equal to the median; this is not true for every distribution.

> The *standard deviation* is an index of the width of a frequency distribution. The smaller the standard deviation, the closer the scores cluster around the mean score. The

greater the standard deviation, the greater the differences between the scores and the mean. The standard deviation is computed by calculating the average squared distance of the scores from the mean and taking the square root of this value.

The mean's location on the histogram in Figure 1 is shown with the letter M; here it is the average chest size of recruits. The standard deviation is the distance, in units of the trait being measured (for chest size the unit is inches), from the mean to the place on the curve marked by M + SD. The height of the normal curve at M + SD is about 60% of its height at M, its maximum height.

4.5.3 Percentiles and the Normal Curve

Galton could have used Quetelet's normal distribution method to describe the distribution of peoples' scores on his measures; there is a mathematical relationship between the normal curve and percentiles, the measure that Galton preferred. Percentiles can be calculated from the mean and standard deviation of normally distributed traits. Figure 2 shows that the mean is at the 50th percentile; M + SD is at the 84th percentile; M - SD is at the 16th percentile.

Once the values of M and SD are calculated for a measure, the values at various percentiles can be read from Figure 2. Using Galton's data for the height of males, M = 67.9 inches and SD = 2.5 inches; so the 16th percentile is M - SD = 67.9 - 2.5 = 65.4 inches; the 50th percentile is 67.9 inches; and the 84th percentile is M + SD = 67.9 + 2.5 = 70.4 inches. Figure 2 also shows the percent of cases falling within intervals expressed using M and SD.

Galton published the results of his measurements at the exhibition using percentiles rather than means and standard deviations. He did this because he thought percentiles would be easier for people to understand and use to determine their standings on the measures. Yet in scientific journals today, it is the mean and standard deviation, not percentiles, that are used routinely to describe distributions. The reason is that these statistics afford a more economical description than percentiles; only two numbers, M and SD, are needed to generate all the percentiles (for normal distributions).

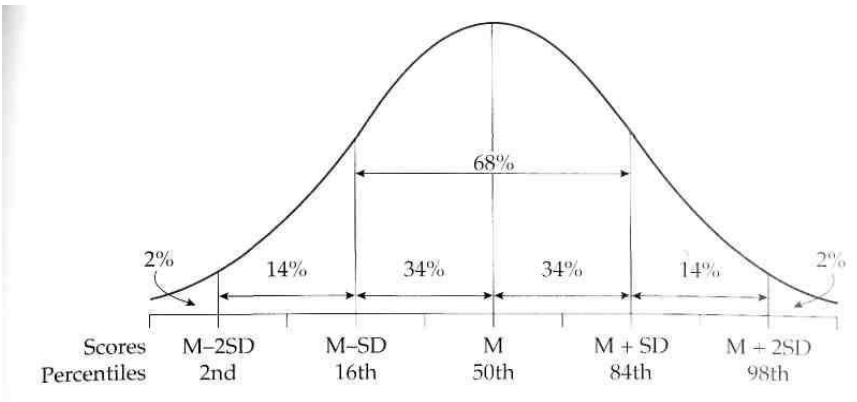


FIGURE 2
Percentiles and the normal curve, with percents under curve rounded to whole numbers.

4.5.4 Why the Normal Distribution Is So Common

Galton's explanation for why human characteristics are normally distributed was different from Quetelet's. Quetelet thought that "errors of creation" had a normal distribution because this distribution was characteristic of all errors. Galton based his explanation on a mathematical theorem discovered by Carl Gauss. Gauss's *central limit theorem* was concerned with the distribution of a variable that is formed by adding together a series of other variables. Gauss showed that such aggregate variables would tend more toward a normal distribution as the number of variables added together became larger. This would occur *regardless* of the distribution of the variables being added together. The theorem predicted that, in practice, any aggregate of variables would have approximately a normal distribution.

Height and weight, the measures studied by Quetelet and Galton, were such aggregates. Height is the sum of the heights of a series of body parts, the height of the foot bone plus the ankle bone, etc., up to the head bone. Weight is the sum of the weights of all separate body parts. Today many psychological measures are purposely constructed so that the score on the scale is an aggregate of other scores. Scale scores then will tend to have a normal distribution. This method of scale construction is discussed in the next section.

The discovery that many human characteristics are normally distributed was exciting to Galton because it showed that individual

differences follow exact mathematical laws. Quetelet's beautiful results were possible because of quantitative measures. Galton was impressed with Quetelet's work. He thought that progress in studying heredity also would come only with refinements in measurement that would allow quantification of complex human traits, like intelligence.

This view is not unique to Galton. It is shared by many scientists. Peter Medawar, for example, who has written extensively on the scientific method, concludes:

> *The art of research is that of making a problem soluble by finding out ways of getting at it. . . . Very often a solution turns on devising some means of quantifying phenomena or states that have hitherto been assessed in terms of "rather more," "rather less," or "a lot of," or workhorse of scientific literature—"marked." (Medawar, 1979, p. 18)*

4.5.5 Galton's Scaling Method

Height and weight are easily quantified, but Galton was not particularly interested in these variables; they were simply convenient to try out his methods, Galton wanted to study intelligence. Like height and weight, intelligence could be considered an aggregate—an aggregate of different skills, such as sensory acuity, quick thinking, ability in arithmetic, problem solving, etc. He expected that a quantitative measure of intelligence would show the normal distribution; jut no measures were available to test this hypothesis.

To construct a quantitative interval scale of intelligence, Galton devised a clever method of scaling that would produce an interval scale from an ordinal rank order scale. The ordinal measure would come from teachers' judgments of their pupils. One hundred pupils, say would be assigned a score from 1 to 100 to indicate their rank order; the student with the highest intelligence would get a score of 100.

These ordinal scores then would be transformed to an interval scale of intelligence using the properties of the normal curve. First, the student with average intelligence, the one who scored 50 on the ordinal scale (the 50th percentile in the group of 100 students), would be given an

arbitrary intelligence score of, say, 100. Next, the student who scored 84 on the ordinal scale (the 84th percentile) would be given an intelligence score one standard deviation above the mean, since the 84th percentile is exactly one standard deviation above the mean on the normal curve. Setting the standard deviation to 15, this student would be assigned a score of 100 + 15 = 115. The value of 15 is arbitrary. Any positive number could be chosen for the standard deviation. The rest of the scores are not arbitrary; the remaining scores on the ordinal scale would be converted to intelligence scores by translating their percentile scores to the corresponding scores a normal distribution with a mean of 100 and a standard deviation of 15. For example, an ordinal score of 98 (the 98th percentile) is two standard deviations above the mean, and would be assigned a score of 100 + 15 + 15 = 130. An ordinal score of 16 would get an intelligence scale score of 100 - 15 = 85. This new scale, Galton argued, would be an interval scale of intelligence. In support of this conclusion, he showed that this method would work for physical characteristics, like height and weight. If you rank order 100 people on height and then go through the scaling procedure, you will end up with an interval scale of height. So Galton expected the method to work for intelligence as well. In discussing Galton's scaling method, Stigler (1986) points out that his argument is based only on an analogy to height. It is not necessarily true that if the method works for height it also will work for intelligence. Direct evidence is needed that the intelligence scale has equal intervals. As we mentioned earlier, such evidence has not been found.

Since Galton's method requires the judges to be familiar with the intelligence of all the students (to rank order them), such a method would be impractical for measuring intelligence in clinical work. Clinicians need measures that can be administered to one person, or to larger samples for research. But a practical variant of Galton's procedure is used in modern measurement: A questionnaire is devised that is made up of a series of, say, 50 to 100 items, and presented to subjects one at a time. The items for an ability test could be problems to solve; for a personality test, questions about feelings, beliefs, and behaviors. Each item is scored either 1 or 0, depending on the subject's answer. A "correct" answer gets the higher score. The subject's score on the scale is the sum of the item scores. Since this is

an aggregate score, it would be expected, by the central limit theorem, to tend to have a normal distribution and, following Galton's logic, to be an interval scale.

The Wechsler Adult Intelligence Scale (WAIS), today's most popular clinical scale, consists of a series of items from different content areas related to intelligence: vocabulary, general information, arithmetic problems, similarities, etc. The subject's scores on the individual items are added together to give a total score for the test. Since these scores are aggregates, they are normally distributed. Similarly, the Minnesota Multiphasic Personality Scale, MMPI, the most popular measure of personality, consists of over 500 true/false items. The scores on particular sets of items are added to form aggregates measuring personality traits that have normal distributions.

4.6 Normal or Non-Normal? – The Logic of Statistical Tests

Quetelet's and Galton's demonstrations that human characteristics have normal distributions, together with the work of astronomers and other scientists showing that this distribution describes errors of measurement, resulted in the normal distribution acquiring almost a cult status. In the 1890s, scientists revered the normal distribution as a "universal law of nature," a distribution with limitless applications.

Usually when extravagant claims are made in science, there are skeptics ready to challenge them. In this case, the skeptic was Karl Pearson, Galton's friend, colleague, and fellow eugenicist. Pearson had a clear interest in showing that the normal distribution did not have limitless applications. In fact, he had developed an elaborate system of equations for describing different *nonnormal distributions.*

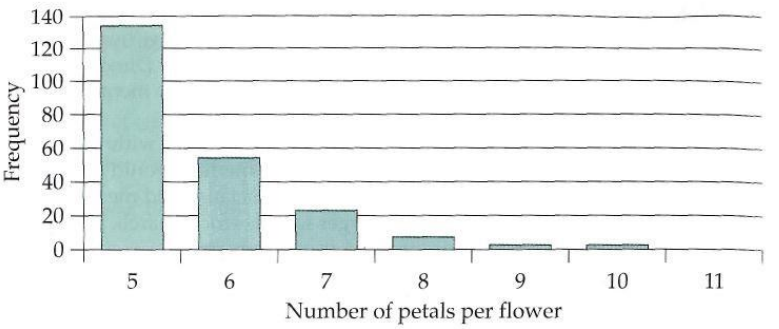


FIGURE 3

Skewed distribution of buttercup petals. (From Pearson, 1900.)

Pearson knew that many distributions of natural events are not normally distributed. One whimsical example he cited (Pearson, 1900) was the distribution of buttercup petals, which is a *skewed* distribution, not a normal distribution.

> In a *skewed distribution,* the mean and median are not equal, as they are in the normal distribution. A distribution is *positively skewed* if the mean is greater than the median and *negatively skewed* if the mean is less than the median.

Figure 3 shows the positively skewed distribution for 222 buttercups. Pearson suspected that the data cited by other scientists to demonstrate normal distributions really were not a good fit to the normal curve either! Professor Merriman, for example, illustrated the normal distribution of errors using data from one thousand rifle shots at a target by soldiers of the U.S. Army. For this demonstration, different areas of the target were marked with the numbers from 1 to 11. Table 5 shows the number of shots that hit each of these different areas, called the *observed frequencies,* and the number of shots that were expected to hit these areas if the distribution of errors was normal, the *expected frequencies.*

If you compare the observed and expected frequencies, they look fairly close. As expected, the highest number of shots, 212, hit area 6, and areas 1 and .1 were hit the least number of times, also as

expected. The observed and expected frequencies for the other areas also appear to be in close agreement. 5ut visually comparing the frequencies is not an objective procedure. Different people certainly would disagree, as Pearson and Merriman did, on whether the observed frequencies were a good fit with the expected frequencies.

What was needed, then, was an objective test to resolve such differences of opinion. Pearson developed such a test, the first widely used *statistical test.*

> *A statistical test* is a mathematical procedure to compare observed results with theoretically expected results in order to reach a conclusion as to whether or not the observations fit the theory.

TABLE 5 DISTRIBUTION OF SHOOTING ERRORS (FROM PEARSON, 1900)

Area	*Observed Frequency*	*Expected Frequency for Normal Distribution*
1	**1**	**1**
2	**4**	**6**
3	**10**	**27**
4	**89**	**67**
5	**190**	**162**
6	**212**	**242**
7	**204**	**240**
8	**193**	**157**
9	**79**	**70**
10	**16**	**26**
11	**2**	**2**

The test Pearson developed for this case is now called the Pearson chi-square, X^2, test. (The Greek letter chi is pronounced ki.)

The first step in the test is to state exactly the theory or hypothesis to be tested. For Pearson, the hypothesis was that the shooting errors had a normal distribution. The second step was to calculate a test statistic, a numerical value, to measure the similarity between the observed frequencies and the expected frequencies. This statistic is

called X^2. It is calculated by subtracting each corresponding observed and expected frequency squaring these differences, dividing by the expected frequency, and then adding up the values.

The formula for X^2 is:

$$X^2 = \sum (O - E)^2/E$$

where O is an observed frequency, E is the corresponding expected frequency, and the summation is over the set of observed frequencies. The smaller the value of X^2, the less the discrepancy between the two sets of frequencies, and the better the fit between the observations and the normal curve. When $X^2 = 0$, the corresponding frequencies are equal. Using Professor Merriman's data, Pearson calculated $X^2 = 45.8$.

Next, Pearson determined what values of X^2 to expect if the distribution of shooting errors was normally distributed. Note that even if the errors are really normally distributed, we would not expect X^2 to be exactly zero. Pearson showed that in this case X^2 is expected to be a small positive number and to vary, sometimes being larger, sometimes smaller, over replications of the study. Pearson figured out the exact probability distribution of X^2, assuming the distribution of errors was really normal (see Figure 4).

This distribution shows that values of X^2 from, say, 3 to 15 are most likely to occur, while large values of $\%^2$, indicating a poor fit, and values of X^2 close to zero, indicating an extremely good fit, are unlikely to occur.

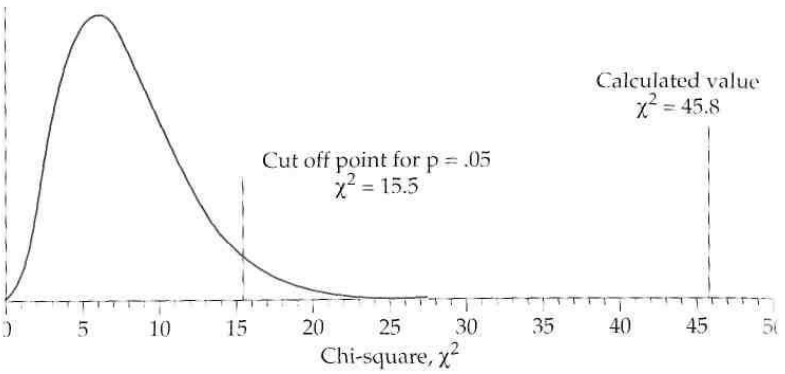


FIGURE 4

The distribution of chi-square if the shooting errors have a normal distribution.

The third step was to compare the calculated value of X^2 with the probability distribution of X^2- The calculated value was 45.8. The probability distribution shows that this is a very unlikely value. Pearson determined that the probability. *p,* of getting this value or an even larger value ($X^2 > 45.8$) was only .00000155.

Pearson based his decision about whether the distribution of shooting errors was normal or not normal on the probability, *p.* We do the same thing today in all statistical tests. We call *p* the *significance probability.*

> The *significance probability, p,* is the probability, if the hypothesis being tested is true, of getting the observed value of the test statistic or an even larger value. (The larger the value of the test statistic the poorer the fit of the data to the hypothesis being tested.)

In this case, the hypothesis being tested is that the distribution is normal. If *p* is too small, say, .05 or less, the fit is not good; that is, the observed data would be quite unlikely to occur if the distribution were normal. If, however, *p* is relatively large, greater than .05, then the fit is satisfactory, and the conclusion would be that the observed frequencies have an underlying normal distribution. The cutoff point of .05 is called the *alpha level,* α, of the statistical test.

> The *alpha level* is the critical value of *p* used in the statistical test. If *p* is equal to or smaller than the value of α *($p < \alpha$),* the hypothesis being tested is rejected; if *p* is greater than α *($p > \alpha$),* the hypothesis is not rejected. The alpha level is set by the experimenter before conducting the test. The value $\alpha = .05$ is commonly used.

Pearson furnished a table that allows researchers to determine the *p* value for any calculated value of X^2 for several different values of α. This chi-square table is published in modern statistics texts.

For Professor Merriman's data, p was very small, well below .05. In Pearson's words:

> *If shots are distributed on a target according to the normal law, then such a distribution as that cited by Mr. Merriman could only be expected to occur, on an average, some15 or 16 times in 10,000,000 trials. (Pearson, 1900, p. 355)*

Although the data appeared to confirm the hypothesis that they were normally distributed, Pearson's test indicated that, in fact, the distribution was not normal. Pearson had made his point, and the hypothesis that the errors had a normal distribution was rejected. The normal distribution was not a universal law of error.

You might expect that Pearson's demonstration would have resulted in a decline in the popularity of the normal curve, but this did not happen. In fact, the normal curve still is the major probability distribution taught today. The routine use of the mean and standard deviation to describe scores is based on the normal distribution being a good approximation to the actual distribution of the scores.

So Pearson's conclusion that the normal curve should have a narrow application in science was not convincing. Ironically, his X^2 test, the method he used to reach this conclusion, became immensely popular. It has become a standard method for analyzing non-normal data, that is, data from ordinal or nominal scales. The X^2 test is used, for example, in modern medical studies to see if a particular high incidence of a disease at a particular place is more than the frequency expected by chance. In studies of birth order, the test is used to study whether firstborns are more likely to be eminent than people with other birth orders. These are just two of the many possible applications of the X^2 test.

The use of statistical tests in data analysis is now virtually universal, and these tests have become standard for comparing observations with theoretical expectations. The most popular tests assume that the observed scores being analyzed are normally distributed.

In 1984, the Association for the Advancement of Science published an issue of their journal *Science 84* celebrating the top 20 discoveries of

the 20th century that have made a "significant impact on the way we think about ourselves and our world" (Hammond, 1984). The award winners included antibiotics, nuclear fission, Einstein's theory of relativity, the computer, television, birth control pills, and Pearson's chi-square test! In presenting the award for Pearson's discovery, Ian Hacking (1984) wrote:

> *The chi-square test was a tiny event in itself, but it was the signal for a sweeping transformation in the ways we interpret our numerical world. . . . For better or worse, statistical inference has provided an entirely new style of reasoning. The quiet statisticians have changed our world—not by discovering new facts or technical developments but by changing the ways we reason, experiment, and form our opinions about it. (Hacking, 1984, p. 70)*

4.7 KEY TERMS

Reactive and non-reactive measures

Multiple measures

Scale properties: equality, rank order, equal intervals, equal ratios

Scale types: ratio, interval, ordinal, and nominal

Standard vs. norm-based measures

Eugenics

Percentiles

Median

Histogram

Normal distribution

Mean, Standard deviation

Aggregate measures

Central limit theorem

Skewed distributions, positive and negative

Statistical test

Pearson's chi-square, X^{2},test

Significance probability, p

Alpha level, α

4.8 KEY PEOPLE

Sheldon Cohen et. Al.

Francis Galton

Charles Darwin

S.S. Stevens

M. A. Quetelet

Carl Gauss

Karl Pearson

4.9 REVIEW QUESTIONS

1. Cohen, Tyrrell,and Smith (1991) used multiple measures of their subjects' smoking and the severity of their colds. Explain why they used multiple measures.

2. Classify the following measures as reactive or nonreactive. Explain your answers.

Cotinine in the blood as a measure of smoking
Self-report of stress
Presence of antibodies in the blood
Wechsler IQ test
Subjects' weight

3. Describe the relationship between scale properties and scale types in Stevens' classification of scales

4. Identify the scale type and scale properties of the following:

weight in grams
temperature in degrees Fahrenheit
temperature in degrees Kelvin
Richter scale
Mohs hardness scale
Beaufort wind scale

5. Classify the following scales according to whether they involve standard or norm-based measurement:

weight in grams
Mohs hardness scale
Wechsler Adult Intelligence Scale
MMPI Depression scale

6. Explain how Galton's goal of developing a eugenics program led him to become interested in statistical methods.

7. Describe Galton's system of percentiles and explain why this system is important in norm-based measurement.

8. What is the advantage of describing sets of scores with the mean and standard deviation rather than percentiles? When would percentiles be preferred?

9. What percentiles correspond to the following scores on a normal curve?

two standard deviations below the mean
one standard deviation below the mean
at the mean
one standard deviation above the mean
two standard deviations above the mean

10. Compare Quetelet's and Galton's explanations for why human characteristics, like height, are normally distributed.

11. Describe how you would construct an interval scale of creativity using Galton's scaling method. Could you be sure that it was an interval scale? Why or why not?

12. Describe the logic of Pearson's chi-square test in your own words.

13. Explain why Pearson's statistical test was selected by the Association for the Advancement of Science as one of the top 20 discoveries of the 20th century.