—Review—

Applications of Statistics to Medical Science (1) Fundamental Concepts

Hiroshi Watanabe

Department of Mathematics, Nippon Medical School

Abstract

The conceptual framework of statistical tests and statistical inferences are discussed, and the epidemiological background of statistics is briefly reviewed. This study is one of a series in which we survey the basics of statistics and practical methods used in medical statistics. Arguments related to actual statistical analysis procedures will be made in subsequent papers.
(J Nippon Med Sch 2011; 78: 274–279)

Key words: statistical test, statistical inference, study design

1. Introduction

Statistics is used in medicine because of the need to adapt medical investigations to principles of natural science and because medical phenomena are nondeterministic by nature.

The principle of natural science is to formulate hypotheses for certain phenomena and to verify the hypotheses by objectively considering the available evidence. To make objective verification possible, we must quantitatively formulate the hypothesis using mathematical terms.

In the case of medical science, phenomena may be considered random to some degree. What is observed is the result of necessity and chance, and medical phenomena are governed by both deterministic and nondeterministic laws. The nondeterministic nature requires a probabilistic view in theoretical considerations, and the principle of objectivity requires statistical methods to analyze experimental results.

Statistics has therefore been indispensable in medical investigations. However, statistics has a long history and many techniques have been developed, making a comprehensive understanding of its scope difficult to obtain. Furthermore, for medical uses, it is not sufficient to manage pure statistics, and we need to be acquainted with the links between medicine and statistics.

This study is one of a series in which we survey statistical methods applied to medical science with focuses on the following aspects: 1) basic concepts in statistics, 2) typical statistical analysis procedures, and 3) characteristic methods used in medical statistics. In this study, we discuss basic concepts in statistics, i.e., statistical tests and statistical inferences, and briefly review epidemiological ideas for study designs. Arguments related to actual statistical analysis procedures and characteristic methods used in medicine will be made in subsequent papers.

Correspondence to Hiroshi Watanabe, Department of Mathematics, Nippon Medical School, 2–297–2 Kosugi-cho, Nakahara-ku, Kawasaki, Kanagawa 211–0063, Japan
E-mail: watmath@nms.ac.jp
Journal Website (http://www.nms.ac.jp/jnms/)

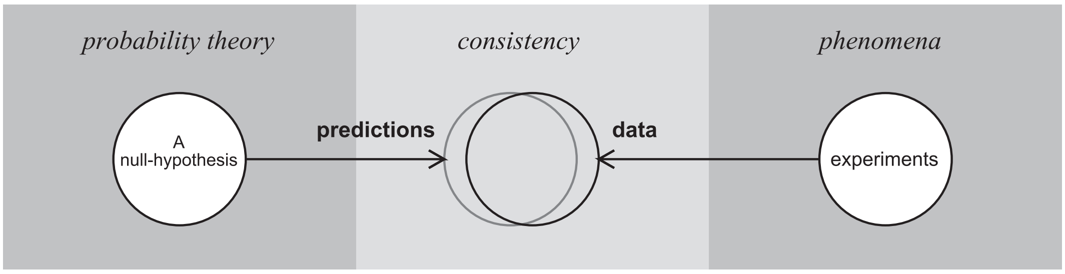

Fig. 1 Schematic view of a statistical test. We perform experiments to obtain experimental data, while we have a null hypothesis that gives predictions. The predictions are compared with experimental data to test the consistency between the null hypothesis and experiments.

2. Principle of Statistical Tests

In this section, we review the fundamental principle of statistical tests from a general viewpoint.

Nondeterministic Phenomena and Probability Theory

When we analyze nondeterministic phenomena such as medical problems in natural science context, we need to be aware of the concept of probability. Based on the concept of probability, we can apply the principle of natural science to nondeterministic phenomena by 1) formulating a probabilistic model as a hypothesis for the nondeterministic phenomena and 2) verifying the hypothesis by means of experimental data (In this section, we use the term “experiment” in a broad sense, i.e., observations without any manipulative actions are included.).

However, the results of experiments in nondeterministic phenomena are themselves non-deterministic and random. This poses several problems such as determining what deduction is to be made from data obtained by chance. For example, when we observe that the head of a coin turns up 51 times in 100 tosses, there is no rationale for insisting that 50 must be the average and the remaining 1 must be due to statistical fluctuation.

Role of Statistics

Statistical analysis helps us to assess a hypothesis related to specific nondeterministic phenomena based on experimental data.

To identify the mechanism responsible for some effect, we must construct an appropriate theoretical model to describe the effect. However, in most cases, such a task belongs to a more advanced stage of the investigation. The first step that is required is to determine whether or not the effect may have occurred by chance. Using statistical analysis, we can deduce that if there was no real effect, such data would not have been obtained by chance, so the data can be regarded as evidence of the existence of some real effect. In other words, by applying statistics, we can test the consistency between the resulting data and the hypothesis stating that there is no real effect. This hypothesis is called a null hypothesis (**Fig. 1**).

To determine whether or not there is a real effect, we need not address the mechanisms behind the effect, but it suffices to test the null hypothesis. However, even if we succeed in ascertaining that there is an effect, a further specific study should be devoted to identifying its mechanisms.

Principle of Statistical Tests

To test a hypothesis, it is necessary to analyze experimental data, i.e., a sample.

In the context of epidemiology, a **population** refers to a large ensemble which is the totality of the subjects in our scope, and a **sample** is a small part of the population consisting of subjects who are actually observed. The aim of statistics is to deduce the properties of a population by analyzing a sample. The uncertainty of statistical assertion is due to the fact that we observe only a sample and not an entire population.

On the other hand, in the context of experiments that take the form of repeated trials, populations and samples are defined in a slightly different manner. In the case of repeated trials, we obtain data in the form of a sequence of events. The totality of what can be observed in a single trial is called a **population**, and the result of repeated trials, which is a sequence of events, is called a **sample**. For example, in a coin-tossing problem, the result of a single toss is a head or tail. Then, the population is the set {head, tail} and a sample resulting from repeated tosses is a sequence such as (head, tail, tail, ... , head). The uncertainty of statistical assertion in this case is due to the fact that the sample size is finite.

In general, statistical analysis proceeds as follows: we first formulate a null hypothesis for a population and then test it by means of a sample. If the sample is considered to be rare under the null hypothesis, it is reasonable to deduce that the null hypothesis is false, because the sample is statistically inconsistent with the null hypothesis. In other words, we **reject** the null hypothesis. In contrast, if the sample is not considered to be rare under the null hypothesis, we cannot eliminate the possibility that the sample was obtained from a population that satisfies the null hypothesis. In this case, we cannot reject the null hypothesis, that is, we **accept** it. Acceptance of the null hypothesis does not necessarily imply that it has been confirmed. Instead, it implies that we lack sufficient evidence to reject it.

Reliability of Statistical Tests

In the process of performing the statistical test described above, we need to assess the *rareness* for each sample. The rareness is expressed in terms of a probability, and we refer to it as a ***P*-value**. Given a *P*-value for a sample, we can state the rule for statistical decision as follows. We reject the hypothesis if the *P*-value is smaller than a prescribed value, e.g., 5%. The cut-off point of 5% is called the **significance level** of this statistical test.

The significance level indicates the degree of uncertainty of the statistical decision. Because we have only considered a sample of finite size, we cannot entirely eliminate errors from the statistical decision process. In a statistical test with a significance level of 5%, we may incorrectly reject (with probability of 5%) a null hypothesis, which is in fact true. For a significance level of 0%, the null hypothesis is never rejected. While such a decision is error-free, the test is trivial and has no practical effect. A non-zero significance level is a requirement for a non-trivial and sensible decision to be made.

Definitions of *P*-value

The question that arises is how it is possible to assess the rareness of a sample and obtain its *P*-value. For example, let us consider the coin-tossing problem and assume the null hypothesis that the coin is fair, i.e., the head will turn up with a probability of 1/2 in a single toss. As a result of 5 tosses, the sample (head, head, head, head, head) would be rare, while the sample (head, tail, tail, head, head) would not be extremely rare, because we expect that the result will be "head" on average 2 or 3 times in 5 tosses. However, both samples appear with the same probability of $1/2^5 = 1/32$. This implies that the rareness of a sample is not measured by the probability that the sample may appear. Instead, the key point is the number of times that the head faces up, because there is only one sample in which the head faces up all 5 times, while there are 10 samples in which the head faces up 3 times in 5 trials. In other words, we can define the *P*-value based on the probability that the number of heads assumes a given value. According to this definition, the *P*-values of the above two samples are 1/32 and 10/32, respectively.

In general, a *statistic* is a quantity whose value is determined by a sample. The number of heads in coin tossing is an example of a statistic that yields the *P*-value of a sample in the statistical test. However, there would be another statistic that yields another *P*-value. As a matter of fact, the actual definition of the *P*-value is not unique, but "to define the *P*-value" is synonymous with "to choose a statistical test." Suppose that we perform two separate statistical tests using the same data. The probabilities that we may reject a true null hypothesis are the same, if the significance levels are the same. In this situation, we can compare the *powers* of these statistical tests. Namely, there may be a difference between the probabilities that we may correctly reject a false null hypothesis.

3. Principle of Statistical Inferences

In this section, we review the fundamental principle of statistical inferences.

Point Estimation

In principle, fundamental laws of physics, such as classical mechanics and quantum mechanics, can be applied to any phenomenon observed in nature. In practice, however, even a familiar phenomenon should be dealt with by using a suitable phenomenological theory instead of the fundamental laws of physics. For example, results of coin tossing are described by a phenomenological probabilistic model instead of classical mechanics. One of the advantages of phenomenological approaches is that only a few parameters are included in the theory, and it suffices to determine their values so that the theory may be consistent with experimental data. In the case of the coin-tossing problem, the probabilistic model is characterized by the probability p with which the coin shows the head in a single toss. The value of p is estimated by experiments: if the head faces up k times in n trials, we may deduce that p will be k/n. This is an example of statistical inference called **point estimation**.

The human body is so complicated that the fundamental laws of physics cannot be applied to it in a straightforward way. Nevertheless, biometric quantities sometimes (approximately) obey a simple statistical law, such as a normal distribution. This is because many factors that weakly interact with each other take part in such a phenomenon. We are then able to describe the statistical aspects of such a biometric quantity by determining a few parameters, because typical statistical distributions are characterized by one or two parameters. For example, a binomial distribution, which is the probability law in the coin-tossing problem, has one parameter p that denotes the single-tossing probability, and a normal distribution has two parameters: mean and variance.

Here, we should make a distinction between the value of a parameter that is determined by a sample and its "true" value. We generically denote them by μ and μ_*, respectively. The true value μ_* is an unknown constant that characterizes the statistical distribution obeyed by the biometric quantity, while μ is its estimator, whose value is determined by a sample. In other words, μ is a statistic that approximates μ_*. For example, in the coin-tossing problem, the true value p_* of the single-tossing probability has the estimator $p = k/n$.

Using the notations μ_* and μ introduced above, we can express the point estimate as follows:

$$\mu_* \approx \mu. \tag{1}$$

This means that the true value μ_* is likely to lie in the neighborhood of the value of μ determined by a sample.

Interval Estimation

The point estimate (1) lacks information regarding the accuracy of the estimation, and hence an **interval estimate** expressed as

$$\mu_1 < \mu_* < \mu_2 \tag{2}$$

is desirable, where the limits μ_1 and μ_2 are determined by a sample similarly as μ in (1).

For example, we consider a biometric quantity χ that obeys the normal distribution with mean μ_* and variance 1, where μ_* is unknown. An interval estimate for μ_* based on a single measurement of χ is obtained as follows. Since $\chi - \mu_*$ obeys the standard normal distribution, i.e., the normal distribution with mean 0 and variance 1, the following inequality holds with a probability of 95%:

$$-1.96 < \chi - \mu_* < 1.96$$

or equivalently

$$\chi - 1.96 < \mu_* < \chi + 1.96. \tag{3}$$

Therefore, if we obtain an experimental value $\chi = 10.00$, we expect that μ_* is likely to satisfy the following bound:

$$8.04 < \mu_* < 11.96. \tag{4}$$

This is the interval estimate for μ_*.

In general, if the inequality (2) holds with probability p for a randomly chosen sample, the interval that is obtained from (2) for a particular sample is called a **confidence interval** with **confidence level** p. The interval (4) is a confidence interval with confidence level 95%. Whatever definitions we give to the statistics μ_1 and μ_2 in (2) as functions of a sample (as in (3)), we may encounter

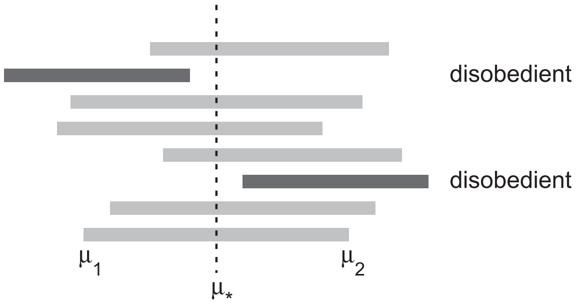

Fig. 2 A typical example of confidence intervals (shaded ones) for an unknown parameter μ_* produced by using eight samples. To yield confidence intervals, we employ the inequality $\mu_1 < \mu_* < \mu_2$ that holds with a certain probability. The endpoints μ_1 and μ_2 are statistics whose values are determined by a sample. In the above illustration, there are six "obedient" samples that give intervals containing the true value of μ_*, while the remaining two are "disobedient" in the sense that the produced intervals do not contain the true value of μ_*.

"disobedient" samples for which (2) does not hold **(Fig. 2)**. In fact, it is for this reason that procedures have been put in place to appropriately define statistics μ_1 and μ_2. This is in case the probability distribution for the biometric quantity under consideration is known to be a typical one, such as a normal, binomial, or Poisson distribution.

We conclude this section by mentioning the relationship that exists between statistical inferences in this section and statistical tests in section 2. Assume that a biometric quantity obeys a probability distribution with a parameter μ_* and let (2) be an inequality that gives a confidence interval with a confidence level of 95%. Furthermore, consider a null hypothesis that is expressed as $\mu_* = 0$, for example. Then, we can realize a statistical test with a significance level of 5% by setting the rule that we accept the null hypothesis if and only if a sample satisfies the following inequality:

$$\mu_1 < 0 < \mu_2.$$

In this sense, interval estimates include statistical tests.

4. Notions Connecting Medicine and Statistics

In this section, we discuss a few notions connecting medicine and statistics from an epidemiological viewpoint. For details, we refer to textbooks such as[1,2].

In general, medical phenomena are considered to be *outcomes* experienced by patients. People receive various *exposures*, such as making contact with patients or taking preventive injections, which can influence (cause or prevent) the outcomes. To establish existence of a cause-effect relationship between the outcome and exposure based on statistical analyses, various study designs are used for medical investigations. Some study design may be used simply to confirm statistical associations instead of a cause-effect relationship.

Study Designs

Designs of nonexperimental studies are described in terms of epidemiology. We present profiles of a few typical study designs below.

A **cohort study** is the most basic study design in medicine. In a canonical cohort study, selections are first made from a population to obtain *cohorts*, each of which is representative of persons sharing the same state of exposure. Next, we observe what happens in each cohort for some period of time. On the basis of these observations, the cohorts are mutually compared with respect to aspects of *incidence* of outcomes, and we then identify the exposure that is responsible for the outcomes. This study design works successfully for frequent outcomes, but is not efficient for rare outcomes.

A **case-control study** is particularly well suited for examining rare outcomes. In this study design, cases with outcomes of interest are pooled together to form a *case group*. However, to correctly determine risk factors for outcomes, we need a *control group*, in which persons without the outcomes of interest are pooled. These groups are then compared with respect to aspects of exposure, and the relevant exposure is identified. This study design has the advantage of allowing us to efficiently use limited case resources. On the other hand, we should consider its disadvantages such as the likely introduction of various biases, including the confounding factors explained below.

A **cross-sectional study** can be used when we are interested in statistical properties of a population rather than a cause-effect relationship between items. In this study design, a selection is made from a population at a point of time (or during a short period of time) to obtain a sample of persons irrespective of exposure and outcomes. We then search for *statistical associations* between observed items in the sample. In principle, in this study design, we cannot determine a cause-effect relationship between items, because exposure and outcomes have equal significance. Furthermore, as the observation is not made over a long time period, what is observed is the *prevalence* of various health states of people instead of the *incidence* of health events. For the same reason, there is a tendency to detect more likely health states that continue for a long time rather than those that do not. This tendency may sometimes give rise to a paradoxical result.

Confounding

A statistical association should be distinguished from a cause-effect relationship especially in nonexperimental studies in which we cannot control exposure manually.

Suppose that we have found a statistically significant association between two events A and B. This does not imply that A causes B. For example, the rustles of leaves are not the cause of the wind, although they are statistically associated with each other. Furthermore, we cannot declare that either A causes B or B causes A, because another event C may cause both A and B. In general, if C causes B and if C is associated with A, then A and B have a statistical association, although A is not directly responsible for B. An event such as C is called a **confounding factor**.

For example, let us consider the following three matters: recovering from a disease (D), taking medicine (M), and drinking water (W). There are four possibilities: 1) W has no direct effect on D but M confounds their relation; 2) M has no direct effect on D but W confounds their relation; 3) neither has a direct effect on D; and 4) both have direct effects on D. For the last option, we may want to assess the net effect of M (or W) on D by controlling the confounding factor W (or M, respectively). We describe this viewpoint by means of another example below.

Consider a case-control study in which we want to conclude that a particular health habit (H) is a risk factor for a disease (D), when age (A) is a known risk factor for D. In this situation, A could be a confounding factor, if A and H have a statistical association. Therefore, it is necessary to eliminate the possible influence of A on D so that we can assess the net effect of H on D. To this end, when we sample the cases and controls, we can perform either specification or matching. In **specification**, we select only persons within a specific age region for both cases and controls. In **matching**, we select the control group so that i) a case and control possess pairs having the same age (**pairwise matching**) or ii) the control group has the same age distribution as the case group (**frequency matching**). Alternatively, when we analyze the data, we can stratify the data (**stratification**). To do this, the case group and control group are respectively decomposed into subgroups (strata) with respect to age, and the effect of H on D is analyzed in each stratum. Furthermore, to resolve confounding, we can perform a *multivariate analysis* that is based on some mathematical model on the effects of A and H on D.

Finally, we emphasize that study protocols must be explicitly written. In particular, definitions of procedures employed in investigations should be *operational* rather than conceptual.

References

1. Rothman KJ, Greenland A, Lash TL: Modern Epidemiology, 3rd ed., 2008; Lippincott Williams & Wilkins, Philadelphia.
2. Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB: Designing Clinical Research, 3rd ed., 2007; Lippincott Williams & Wilkins, Philadelphia.

(Received, July 12, 2011)
(Accepted, August 9, 2011)