—Review—

Applications of Statistics to Medical Science, II
Overview of Statistical Procedures for General Use

Hiroshi Watanabe

Department of Mathematics, Nippon Medical School

Abstract

Procedures of statistical analysis are reviewed to provide an overview of applications of statistics for general use. Topics that are dealt with are inference on a population, comparison of two populations with respect to means and probabilities, and multiple comparisons. This study is the second part of series in which we survey medical statistics. Arguments related to statistical associations and regressions will be made in subsequent papers.
(J Nippon Med Sch 2012; 79: 31–36)

Key words: statistical test, statistical inference

1. Introduction

In the previous work[1], we discussed a conceptual framework of statistical tests and statistical inferences. The present paper intends to give an overview of applications of statistics for general use.

The aim of statistics is to draw inferences on populations from samples that are randomly chosen from populations. Statistical inference is made as point estimation, interval estimation, and a statistical test for population parameters. In this paper, we describe typical analysis procedures from these 3 viewpoints.

Textbooks[2,3] are rich sources of information on medical statistics. In particular, we find medical examples of the statistical procedures presented in this paper.

2. Inference on a Population

In this section, we describe how to infer parameters of a population.

Point Estimation

Let M be a population with mean μ and variance σ^2. To infer the values of μ and σ^2 from a sample x_1, x_2, $\cdots$, x_n that are randomly chosen from M, we compute a sample mean $\bar{x}$ and an unbiased variance v^2 defined by

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{1}{n}(x_1+x_2+\cdots+x_n), \tag{1}$$

$$v^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2. \tag{2}$$

The expectation values of $\bar{x}$ and of v^2 are equal to μ and σ^2, respectively. Furthermore, if the sample size n is large, values of $\bar{x}$ and of v^2 concentrate in the neighborhoods of μ and of σ^2, respectively. In this sense, we write

$$\mu\approx\bar{x},\ \sigma^2\approx v^2. \tag{3}$$

This is the reason we use $\bar{x}$ and v^2 as estimators of μ and σ^2, respectively.

Values of the estimator $\bar{x}$ of μ are scattered

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

around the value of μ. The magnitude of an error E associated with this estimation is measured by (the square root of) the variance of $\bar{x}$. Because the variance of $\bar{x}$ is equal to σ^2/n and σ^2 has the estimator v^2, the error E is estimated by the statistic $v/\sqrt{n}$, which we call a standard error of $\bar{x}$ and write

$$\mathrm{SE}(\bar{x}) = \frac{v}{\sqrt{n}}. \tag{4}$$

Interval Estimation

Let us assume that the population M obeys a normal distribution. Then, the statistic

$$t = \frac{\bar{x}-\mu}{\mathrm{SE}(\bar{x})} = \frac{\bar{x}-\mu}{v/\sqrt{n}} \tag{5}$$

obeys a t distribution with $n - 1$ degrees of freedom. On the basis of this fact, we obtain an interval estimation for μ as follows. Denote the $(1 - \alpha/2) \times 100$th percentile of the t distribution with $n - 1$ degrees of freedom by $t_{n-1}(1 - \alpha/2)$. Then, the confidence interval of μ with confidence level $1 - \alpha$ is given by

$$\bar{x} - \delta < \mu < \bar{x} + \delta \tag{6}$$

with

$$\delta = t_{n-1}(1-\alpha/2)\,\mathrm{SE}(\bar{x}), \tag{7}$$

which follows from the bound $|t| < t_{n-1}(1 - \alpha/2)$.

Statistical Test

Let μ_0 be an arbitrary constant. Then, the null hypothesis H_0: $\mu = \mu_0$ is tested by means of the confidence interval (6). A test with significance level α is implemented by the following rule:

Accept H_0 if and only if $\mu = \mu_0$ belongs to the interval (6), i.e., if and only if $|\bar{x} - \mu_0| < \delta$ holds,

where δ is defined by (7). This test is referred to as a **one-sample *t* test.**

Remarks. 1) A one-sample t test is, in fact, available in the situation where the distribution of the population might slightly deviate from a normal distribution. We refer to this property of the one-sample t test as *robustness*. However, if the distribution is far from being normal and the sample size is not large, we should use nonparametric statistical methods, such as the *Wilcoxon signed-rank test*, for the null hypothesis that the (population) median is μ_0.

2) The one-sample t test for the null hypothesis $\mu = 0$ is especially important for comparison of *paired samples*. See section 3.

3) The definition (5) of t has the form "(estimator − parameter)/SE". We shall see later further examples of this form.

4) Interval estimation of σ^2 is possible based on the fact that $(n - 1)\,v^2/\sigma^2$ obeys a chi-square distribution with $n - 1$ degrees of freedom.

3. Comparison of Two Populations: Means

In this section, we describe how to compare means of two populations.

Point Estimation

Let M_x and M_y be populations with means μ_x and μ_y, respectively, and with the same variance

$$\sigma_x^2 = \sigma_y^2. \tag{8}$$

To compare means, we estimate the difference $\mu_x - \mu_y$. Let $x_1, x_2, \cdots, x_m$ and $y_1, y_2, \cdots, y_n$ be samples randomly chosen from M_x and M_y, respectively. Denote their sample means by $\bar{x}$ and $\bar{y}$, respectively, and their unbiased variances by v_x^2 and v_y^2, respectively. The difference $\mu_x - \mu_y$ has the estimator $\bar{x} - \bar{y}$ and the squared standard error $\mathrm{SE}(\bar{x} - \bar{y})^2$ of $\bar{x} - \bar{y}$ is given by

$$\mathrm{SE}(\bar{x}-\bar{y})^2 = \frac{(m-1)v_x^2 + (n-1)v_y^2}{m+n-2}\left(\frac{1}{m}+\frac{1}{n}\right) \tag{9}$$

under the assumption (8).

Interval Estimation

Let M_x and M_y be normally distributed populations with (8). Then, the statistic

$$t = \frac{(\bar{x}-\bar{y}) - (\mu_x - \mu_y)}{\mathrm{SE}(\bar{x}-\bar{y})} \tag{10}$$

obeys a t distribution with $m + n - 2$ degrees of freedom. On the basis of this fact, we obtain the following confidence interval for $\mu_x - \mu_y$:

$$\bar{x} - \bar{y} - \delta' < \mu_x - \mu_y < \bar{x} - \bar{y} + \delta' \tag{11}$$

with

$$\delta' = t_{m+n-2}\left(1 - \frac{1}{2}\alpha\right)\mathrm{SE}(\bar{x}-\bar{y}), \tag{12}$$

where $t_{m+n-2}(1-\alpha/2)$ denotes the $(1-\alpha/2)\times 100$th percentile of the t distribution with $m+n-2$ degrees of freedom.

Statistical Test

The null hypothesis H_0: $\mu_x = \mu_y$ is tested by means of the confidence interval (11). A test with significance level α is implemented by the following rule:

Accept H_0 if and only if $\mu_x - \mu_y = 0$ belongs to the interval (11), i.e., if and only if $|\bar{x} - \bar{y}| < \delta'$ holds,

where δ' is defined by (12). We refer to this test as a **two-sample *t* test**.

Remarks. 1) A two-sample t test is applied to unpaired samples. This means that the two samples must be independently chosen from two populations.

2) Suppose that we compare a set of data for patients obtained a year earlier with another set of data for the same patients obtained a month earlier. In this case, we cannot regard the sets of data as unpaired samples, because two results of a patient may have some dependence. These samples are called paired samples; the sizes of paired samples are the same ($m = n$), and x_i and y_i make a pair. Paired samples should be analyzed in the form of differences $x_i - y_i$ by using the method stated in section 2. We refer to this test as a *paired t test*.

3) When we cannot assume that the populations are normally distributed, we should use nonparametric statistical methods as *Wilcoxon rank sum test* to compare (population) medians.

4) We can test the assumption (8) by an F test using the statistic $F = v_x^2/v_y^2$. If we fail in accepting the hypothesis (8) as a result of the F test, *Welch's test* is available as an alternative of the two-sample t test.

4. Comparison of Two Populations: Probabilities

In this section, we describe procedures to compare probabilities.

Probabilities, Relative Risk, and Odds Ratio

Suppose that two properties X and Y are defined on a population M. For example, X means that a subject is *exposed* to a certain situation, and Y means that we observe a certain *outcome* for a subject. We decompose M into two subpopulations M_1 and M_2 according to whether X holds or not, i.e., an element of M belongs to M_1 (or M_2) if the element has (or does not have, resp.) the property X. Let p_1 and p_2 be the probabilities with which Y may hold in M_1 and in M_2, respectively.

To compare p_1 and p_2, we introduce the following ratios:

$$\text{relative risk} = \text{RR} = \frac{p_1}{p_2}, \tag{13}$$

$$\text{odds ratio} = \text{OR} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}. \tag{14}$$

If X and Y are statistically independent, we have $p_1 = p_2$ and RR = OR = 1.

Point Estimation

The population parameters p_1, p_2, RR, and OR are estimated by using a randomly chosen sample. Sample data for this purpose are categorical and presented in the form of a contingency table (**Table 1**). The estimators of p_1, p_2, RR, and OR are as follows:

$$p_1 \approx \frac{a}{a+b}, \quad p_2 \approx \frac{c}{c+d}, \tag{15}$$

$$\text{RR} \approx \frac{a/(a+b)}{c/(c+d)}. \quad \text{OR} \approx \frac{a/b}{c/d}. \tag{16}$$

Interval Estimation

We describe how to obtain confidence intervals of RR and of OR. Put

$$w = \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}, \tag{17}$$

$$w' = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}. \tag{18}$$

Then, the confidence intervals of RR and OR with confidence level $1-\alpha$ are given by

$$\frac{a/(a+b)}{c/(c+d)} e^{-z(1-\alpha/2)w} < \text{RR} < \frac{a/(a+b)}{c/(c+d)} e^{z(1-\alpha/2)w}, \tag{19}$$

$$\frac{(a/b)}{(c/d)} e^{-z(1-\alpha/2)w'} < \text{OR} < \frac{(a/b)}{(c/d)} e^{z(1-\alpha/2)w'}, \tag{20}$$

respectively, where $z(1-\alpha/2)$ denotes the $(1-\alpha/2)\times 100$th percentile of the standard normal distribution and $e = 2.71828\cdots$ is a mathematical constant.

Table 1 A contingency table: a sample consisting of $a+b+c+d$ elements is divided into four categories according to whether or not each element has the properties X and Y, respectively.

	with Y	*without* Y	Total
with X	a	b	$a+b$
without X	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

Statistical Test

Let us consider statistical tests of the null hypothesis H_0: $p_1 = p_2$ by the sample data shown in **Table 1**.

The simplest procedure to test H_0 is a chi-square test (or Fisher's exact test) for a contingency table. More sophisticated tests rely on the formulae (19) and (20). For a test using RR (or OR) with significance level α, we accept H_0 if and only if the value RR = 1 (or OR = 1) satisfies (19) (or (20), resp.).

Remarks. 1) Interval estimations of relative risk and of odds ratio are useful for assessing *causal relations* between exposures and outcomes in cohort studies and case-control studies. To confirm *statistical associations* (instead of causal relations) in cross-sectional studies, the chi-square test is usually used.

2) In a case-control study, we cannot estimate relative risk by (16). In fact, the ratios $a : b : (a + b)$ and $c : d : (c + d)$ are determined by the study plan (instead of the populations), hence they do not reflect objective properties of the populations. In this case, the odds ratio should be estimated, because the ratios $a : c$ and $b : d$ have objective meanings and odds ratio approximates relative risk if p_1 and p_2 are small, as is seen from (13) and (14).

3) *Matched-pair data* in a case-control study may have some dependence within the pair. We therefore cannot regard the data as being independently chosen from two populations M_1 and M_2. In this case, McNemar's test can be used to test the null hypothesis $p_1 = p_2$.

5. Multiple Comparisons

In this section, we deal with the problem of comparing means of three or more populations. It is essential to understand the reason why pairwise comparisons may cause a trouble.

Successive Applications of Statistical Tests

Let T_1 and T_2 be two tests of a null hypothesis H_0. We compose them into a test T as follows: 1) perform T_1 with significance level α; 2) perform T_2 with significance level α; and 3) reject H_0 if at least one of these tests rejects H_0. The key point is that the *overall significance level* α' of the test T is not α. In fact, if H_0 is true, H_0 may be rejected with probability α in each of T_1 and T_2, hence the probability α' with which the test T may reject H_0 is given by

$$\alpha' = 1-(1-\alpha)^2 = 2\alpha-\alpha^2 (>\alpha), \tag{21}$$

where we have assumed that consequences of T_1 and T_2 are independent. Although it is possible to set α so that α' may be a desired value, e.g., 0.05, the problem becomes serious, when we successively perform many tests, because α should be set extremely small.

For the same reason, pairwise comparisons may cause an undesired change of significance level, when we compare three or more populations. In what follows, we discuss a solution to this problem.

Comparison of Means of Three or More Populations

Suppose that we are comparing means of r populations $M_1, M_2, \cdots, M_r$. Denote the mean and variance of M_i by μ_i and σ_i^2, respectively, for $i = 1, 2, \cdots, r$, and assume

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2. \tag{22}$$

To test the null hypothesis H_0: $\mu_1 = \mu_2 = \cdots = \mu_r$, we randomly choose samples from the populations (**Table 2**). Let n_i (or $\bar{x}_i$) be the size (or the sample mean, resp.) of the sample $x_{i1}, x_{i2}, \cdots$ chosen from M_i for $i = 1, 2, \cdots, r$, and denote the total size by $N = n_1 + n_2 + \cdots + n_r$. Then, the sample mean over all groups is given by

$$\bar{x} = \frac{1}{N}\sum_i \sum_j x_{ij} = \frac{1}{N}\sum n_i \bar{x}_i. \tag{23}$$

We introduce two kinds of *sums of squares*:

Table 2 A set of samples for a one-way ANOVA: a sample chosen from the population M_i consists of n_i elements and has sample mean $\bar{x}_i$. The total size of the samples is N, and the sample mean over all groups is $\bar{x}$.

Population	Sample	Sample size	Sample mean
M_1	$x_{11}, x_{12}, \cdots\cdots$	n_1	$\bar{x}_1$
M_2	$x_{21}, x_{22}, \cdots\cdots$	n_2	$\bar{x}_2$
⋮	⋮	⋮	⋮
M_r	$x_{r1}, x_{r2}, \cdots\cdots$	n_r	$\bar{x}_r$
Total		N	$\bar{x}$

within-Groups Sum of Squares:

$$SS_{\text{within}}=\sum_i\sum_j (x_{ij}-\bar{x}_i)^2, \quad (24)$$

Between-Groups Sum of Squares:

$$SS_{\text{between}}=\sum_i n_i\,(\bar{x}_i-\bar{x})^2, \quad (25)$$

and put

$$F=\frac{\dfrac{1}{r-1}SS_{\text{between}}}{\dfrac{1}{N-r}SS_{\text{within}}}. \quad (26)$$

Since the quantity SS_{within} measures the variances of data within groups, SS_{within} is insensitive to the population means $\mu_1, \mu_2, \cdots, \mu_r$, whereas SS_{between} that measures the variance of data between groups is sensitive to population means and is likely to be large unless

$$\mu_1=\mu_2=\cdots=\mu_r \quad (27)$$

holds. Then, we expect that F may tend to be large unless (27) holds.

Assume that $M_1, M_2, \cdots, M_r$ are normally distributed populations with (22) and (27). Under these assumptions, F obeys an F distribution with $r-1$ and $N-r$ degrees of freedom. Therefore, in the situation where (22) has been confirmed, the null hypothesis H_0: $\mu_1=\mu_2=\cdots=\mu_r$ is tested with significance level α by the following rule:

Reject H_0 if and only if $F>F^{r-1}_{N-r}(1-\alpha)$,

where $F^{r-1}_{N-r}(1-\alpha)$ denotes the $(1-\alpha)\times 100$th percentile of the F distribution with $r-1$ and $N-r$ degrees of freedom. Note that this is a one-sided test. If the null hypothesis H_0 is false, the statistic F is likely to be large, hence we should reject H_0 if and only if F is large.

The above procedure is called a **one-way analysis of variance** (abbreviated as a one-way **ANOVA**). In a situation where the population M is decomposed into subpopulations $M_1, M_2, \cdots, M_r$ according to some factor, e.g., age of subjects, one-way ANOVA works as a test of whether or not the factor has an effect on outcomes.

Remarks. 1) As long as (22) is satisfied, the one-way ANOVA is also available in the situation where the distributions of the populations may slightly deviate from a normal distribution. In this sense, ANOVA is robust. However, if the distributions of populations are far from being normal, we should use a nonparametric alternative called the *Kruskal-Wallis* test.

2) To confirm the assumption (22) is a subtle problem. For this purpose, Bartlett's test is known. As an alternative of the one-way ANOVA which is available without the assumption (22), an approximate test using weighting is known. However, from a practical viewpoint, it may be advisable to plan a study so that sample sizes $n_1, n_2, \cdots, n_r$ are equal or close to equal, because the assumption (22) is not relevant in this situation.

3) Even if we succeed in rejecting the null hypothesis H_0: $\mu_1=\mu_2=\cdots=\mu_r$, we cannot clarify which population mean deviates. This problem needs a further analysis, which will be discussed in the next subsection.

4) Using the one-way ANOVA, we can study an effect of a factor on outcomes. To assess effects of two factors, we should use a *two-way ANOVA*.

Contrasts

To compare three means μ_1, μ_2, and μ_3 in detail, we may want to perform multiple comparisons, i.e., inferences of three differences $\mu_1-\mu_2$, $\mu_1-\mu_3$, and $\mu_2-\mu_3$. It will be convenient to introduce a linear combination $c_1\mu_1+c_2\mu_2+c_3\mu_3$ as a generalization of the differences, where c_1, c_2, and c_3 are constants satisfying $c_1+c_2+c_3=0$.

A linear combination

$$\lambda=c_1\mu_1+c_2\mu_2+\cdots+c_r\mu_r \quad (28)$$

of r means $\mu_1, \mu_2, \cdots, \mu_r$ is called a **contrast**, if

$$c_1+c_2+\cdots+c_r=0 \tag{29}$$

holds. In what follows, we discuss how to infer the value of λ from the sample shown in **Table 2**.

Because μ_i has an estimator $\bar{x}_i$, the value of λ is estimated by the statistic

$$L=c_1\bar{x}_1+c_2\bar{x}_2+\cdots+c_r\bar{x}_r. \tag{30}$$

Under the assumption (22), L has the squared standard error $SE(L)^2$ given by

$$SE(L)^2=\frac{1}{N-r}\left(\frac{c_1^2}{n_1}+\frac{c_2^2}{n_2}+\cdots+\frac{c_r^2}{n_r}\right)SS_{\text{within}}, \tag{31}$$

where SS_{within} is defined by (24).

Assume that population M_1, M_2, ⋯, M_r are normally distributed and (22) holds. Then, the statistic

$$t=\frac{L-\lambda}{\mathrm{SE}(L)} \tag{32}$$

obeys a t distribution with $N-r$ degrees of freedom. A t test by means of (32) being possible, its successive applications may cause a problem, in particular if we explore various contrasts one after another. The key point of the solution is to take into account *all* contrasts λ with (29) instead of looking only at the ones that have direct relevance to our interest. Let $F_{N-r}^{r-1}(1-\alpha)$ be the $(1-\alpha)\times 100$th percentile of an F distribution with $r-1$ and $N-r$ degrees of freedom, and put $\rho=(r-1)\,F_{N-r}^{r-1}(1-\alpha)$. Then, the following event occurs with probability $1-\alpha$:

the inequality $t^2<\rho$ holds for *any* contrast λ,

where t is defined by (32). Note that the inequality $t^2<\rho$ yields the following interval estimation of the corresponding contrast λ:

$$L-\sqrt{\rho}\cdot SE(L)<\lambda<L+\sqrt{\rho}\cdot SE(L). \tag{33}$$

The confidence level is not reduced (from $1-\alpha$), no matter how many contrasts we consider, because $t^2<\rho$ holds for *any* contrast (with probability $1-\alpha$). We refer to this method as **Scheffé's procedure**.

As an example, we compare three means μ_1, μ_2, and μ_3. Consider three contrasts and their estimators

$$\lambda_1=\mu_1-\mu_2,\ \lambda_2=\mu_1-\mu_3,\ \lambda_3=\mu_2-\mu_3, \tag{34}$$

$$L_1=\bar{x}_1-\bar{x}_2,\ L_2=\bar{x}_1-\bar{x}_3,\ L_3=\bar{x}_2-\bar{x}_3. \tag{35}$$

The corresponding t statistics are

$$t_i=\frac{L_i-\lambda_i}{\mathrm{SE}(L_i)},\ i=1,\ 2,\ 3, \tag{36}$$

and the following event occurs with probability *larger than* $1-\alpha$:

the inequalities $t_i^2<\rho$ *simultaneously* hold for $i=1, 2, 3$.

Because we consider only three contrasts (34) instead of all the contrasts, the probability is larger than $1-\alpha$. The inequality (33) yields simultaneous confidence intervals

$$L_i-\sqrt{\rho}\cdot\mathrm{SE}(L_i)<\lambda_i<L_i+\sqrt{\rho}\cdot\mathrm{SE}(L_i),\quad i=1,\ 2,\ 3 \tag{37}$$

for λ_1, λ_2, and for λ_3 with confidence level larger than $1-\alpha$. By means of (37), we can determine which contrasts differ from 0. Suppose we find that $\lambda_1>0$, $\lambda_2>0$, and $\lambda_3=0$. Then, our conclusion is $\mu_1>\mu_2=\mu_3$.

Remarks. 1) We can apply Scheffé's procedure to the contrast $\lambda=2\mu_1-\mu_2-\mu_3$ for comparison of μ_1 and the average of μ_2 and μ_3. More complicated comparisons are possible using contrasts.

2) The actual confidence level of the confidence intervals (37) is *larger* than $1-\alpha$. This means that Scheffé's procedure yields conservative results.

3) If the sizes $n_1, n_2, \cdots, n_r$ of samples are the same, the *Tukey test* is available for multiple *pairwise* comparisons.

References

1. Watanabe H: Applications of Statistics to Medical Science, I. Journal of Nippon Medical School 2011; 78: 274–279.
2. Rosner B: Fundamentals of Biostatistics, 7th ed. 2011; Brooks/Cole, Boston.
3. Armitage P, Berry G, Matthews JNR: Statistical Methods in Medical Research, 4th ed. 2002; Blackwell Science, Massachusetts.

(Received, October 24, 2011)
(Accepted, December 7, 2011)