

## **Testing and P-Values: Statistical Evidence**

<Material provided by>  
AMED "International Journals Project"

Unauthorized reproduction prohibited.

Reproduction Prohibited

## **Contents**

Introduction

What Is Statistical Evidence?

Why Reject Hypotheses?

One-Sided Test or Two-Sided Test?

Two-Sided Test

One-Sided Test

Interpretation of Study Results Using P-Values

Limitations of P-Values

P-Values and Confidence Intervals

Interpretation of Research Results Using Confidence Intervals

When Proving Superiority

When Proving Non-Inferiority or Equivalence

---

## Introduction

In clinical research, it is not realistic to collect data from the entire population. Instead, data is collected from a group of randomly selected subjects, which well-represents the population. The group of randomly selected subjects is called a “sample” and data collected from it is called “sample data.” “Statistical inference” means to infer what can be said about the population based on the sample data. There are two methods to make statistical inference: statistical estimation (see the modules “Proper Data Description”) and statistical hypothesis testing.

In this module, you will learn about the latter method, statistical hypothesis testing. For example, let’s say you want to prove that Treatment A, which was developed and has been practiced in Japan, is more effective than the conventional treatment, Treatment B. In this case, it is almost impossible for you to collect data from all the patients who have ever received Treatment A. Therefore, you judge whether your hypothesis, “Treatment A is more effective than Treatment B,” is true for the population based on the sample data. This process is called “hypothesis testing.” The judgment made based on hypothesis testing is closer to the truth when the quantity of sample data is larger.

Hypothesis testing has two stages:

- (1) Establishing a hypothesis; and
- (2) Rejecting the hypothesis.

Whether or not to reject the hypothesis is decided using a p-value. In this module, you will also learn about p-values in addition to hypothesis testing.

### Learning Objectives

Your goals in this module are to be able to:

- Describe the process of hypothesis testing.
- Explain the meaning of p-values.
- Interpret the results of hypothesis testing.

---

## What Is Statistical Evidence?

---

### Why Reject Hypotheses?

For example, there are two ways to prove the hypothesis that water boils at 100 C°.

- (1) **Support** the hypothesis that “water **boils** at 100 C°.”
- (2) **Reject** the hypothesis that “water **does not boil** at 100 C°.”

When obtaining scientific evidence, we apply the second approach above, which uses a **double negative** of **rejecting** the hypothesis that is a **negative** statement. This is identical to say, “I don’t think I don’t like apples,” instead of simply saying, “I like apples.” A hypothesis that supports the intended conclusion, (1) above, is called an “**alternative hypothesis**” and a hypothesis that rejects the intended conclusion, (2) above, is called a “**null hypothesis**,” because this type of hypothesis is intended to be nullified later. Surely, we would very much like to use the alternative hypothesis (1) above and say, “water boils at 100 C°” or “I like apples.” However, in order to acquire scientific evidence, we need to hypothesize “it is not possible that water does not boil at 100 C°” or “I don’t think I don’t like apples” and reject this null hypothesis.

Why should we use such devious and indirect approach?

For example, what should we do if we would like to prove the hypothesis that water boils at 100 C°? Let’s say you first observe water come to a boil at 100 C° in Tokyo. Next, you see water come to a boil at 100 C° in Osaka. Then, you see water come to a boil at 100 C° in Hokkaido, New York, and everywhere you go. Are these facts enough to provide scientific evidence? No, the data is still far from enough. Now, you go to Alaska, the U.K., and 10,000 places around the world. In every single place, water boils at 100 C°. Are the data enough to prove your hypothesis? You choose Mt. Fuji as the 10,001st place, where you see water boil at 88 C°. Now, the hypothesis that water boils at 100 C° must to be rejected, because of this last test result. **After rejecting the original hypothesis**, you make a new hypothesis that “water does not always boil at 100 C°; rather, the boiling point depends on the altitude” and repeat the testing process to prove this new hypothesis. Scientific evidence solidifies gradually through this process.

As you can see from this example, no amount of data is enough to support a hypothesis, but **a single fact is enough to reject it**. It is for this reason that **a double negative of “rejecting null hypotheses”** is used to obtain scientific evidence.

For example, let’s say we would like to prove a clinical hypothesis that Drug A is more effective than Drug B in reducing the blood pressure of adult Japanese males. We first administered Drug A and B to 50 patients each, measured their blood pressures and compared the mean of their blood pressures (the average blood pressure) between the group that received Drug A (Group A) and the group that received Drug B (Group B). The main data to be evaluated in this example is the average blood pressure after drug administration. Such results of primary interest are called “**outcomes**.” In this example, we generate the following null hypothesis.

**Null hypothesis:** The difference in the average blood pressure between Group A and B is zero.

What criteria can we use to reject this null hypothesis? First, let’s assume that this null hypothesis is correct regarding the population of adult Japanese males. We also suppose the collected data show that the number of subjects who experienced a blood pressure reduction is clearly larger in Group A than in Group B. We could make the following two judgments based on this test result.

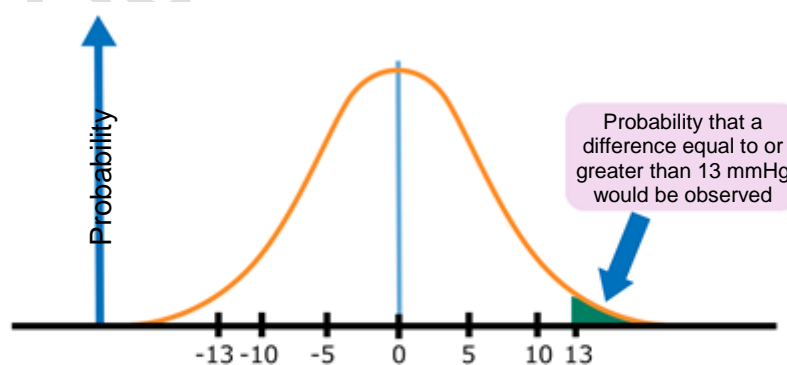
(1) The null hypothesis is correct, i.e. Drug A is no more effective than Drug B and the above

test results were just a coincidence.

- (2) The test results were not a coincidence. The null hypothesis is wrong, i.e. Drug A is more effective than Drug B.

We use p-values to judge whether to reject a null hypothesis. A p-value is **short for a realized value of probability**, that is, **a realized value of the probability that either the phenomenon observed in sample data or a phenomenon even more divergent from the hypothesis occurs when the null hypothesis is true for the population**. In the above example, a p-value is a realized value of the probability that a difference equal to or greater than the difference observed by chance occurs when Drug A is no more effective than Drug B. In other words, a small p-value suggests that the null hypothesis is incorrect. Then, how small does a p-value have to be to reject a null hypothesis? It goes without saying that the reference value for this judgement should not be determined arbitrarily; otherwise, null hypotheses can be rejected conveniently. Therefore, a reference value for this judgement must be chosen beforehand. This reference value is called “significance level” and has conventionally been 5%. If the p-value is smaller than the significance level, we conclude as the (2) above states. In this case, we report that “the difference was statistically significant (at a significance level of 5%).” On the other hand, the p-value can be viewed as “**the realized value of the probability** that Drug A is wrongly judged” as more effective than Drug B from the sample data “when Drug A is no more effective than Drug B.” It is obvious that the smaller the realized value of the probability of misjudgment is, the better. If we decide to “reject the null hypothesis” when the p-value is smaller than 5%, the above-mentioned conventional reference value, we can limit the probability of wrongly rejecting the null hypothesis to 5% at maximum. If we can limit the probability of wrongly rejecting the null hypothesis to a small value (5% in this case), it is fair to adopt the judgment (2) above.

Now, let’s see how to calculate the p-value in the above example. Here, we assume that the average blood pressure of Group A and that of Group B are 100 mmHg and 113 mmHg, respectively (thus the difference is 13 mmHg), and the standard error of the average blood pressure is 5 mmHg for the two groups.



Difference in the average blood pressure between the two groups based on the sample data randomly collected from the population

Let's say that an unspecified number of researchers collected data separately and compared the average blood pressure of Group A against that of Group B in order to test the null hypothesis. Each researcher obtained data on the difference in the average blood pressure between the two groups. The first researcher found that the average blood pressure of Group A was 10 mmHg higher than that of Group B, whereas the second researcher observed that the average blood pressure of Group B was 5 mmHg higher than that of Group A. In this manner, the value of the difference may vary from one researcher to another. Now, let's think about the distribution of the values of the difference in the average blood pressure. An unlimited number of the values are possible. If the null hypothesis that "Drug A is no more effective than Drug B" is correct, the most frequent value of the difference in the average blood pressure is zero, and the frequency decreases as the value shifts away from zero.

The horizontal axis in the above figure shows the difference between the two groups in the average blood pressure calculated from the sample data randomly collected from the population. The vertical axis represents the probability (strictly speaking "probability density" in statistics terminology) that a certain value of the difference in the average blood pressure is observed between the two groups **when the null hypothesis is correct**. Let's assume that the distribution of the values forms a symmetrical shape centered on the value, zero. The red curve indicates the distribution of the values and the area under the red curve represents the probability. For example, the area under the curve is 0.5, or 50% of probability, when the values on the horizontal axis are equal to or greater than zero. This means that the probability that the difference in the average blood pressure between the two groups is zero or greater is 50% when the null hypothesis is correct for the population. The area of the blue part in the figure represents the probability that the difference in the average blood pressure between the two groups happens to be 13 mmHg or greater by coincidence in the sample data when the null hypothesis is correct. Now, let's recall the definition of a p-value. A p-value is defined as **a realized value of the probability that either the phenomenon observed in sample data, which represents a part of the population, or a phenomenon even more divergent from the hypothesis occurs when the null hypothesis is correct for the population**. In the above example, the "phenomenon observed in the sample data" is "the difference in the average blood pressure calculated from the sample data" and is 13 mmHg. Therefore, the area of the blue part in the above figure represents *the* p-value. Let's assume that the area of this part is 3%. In this example, the p-value is **"the realized value of the probability that the average of the observed outcome (the difference in the average blood pressure) is 13 mmHg or greater when the null hypothesis is correct"** and is 3%. In other words, even if Drug A is truly no more effective than Drug B, there is 3% of probability that such a difference as is seen here is observed by chance. Since this 3% of probability is smaller than 5%, we can say that this probability is too small to be a coincidence, meaning it is not a coincidence. Thus, we judge that the null hypothesis is wrong, reject it, and gain evidence for the clinical hypothesis mentioned earlier, "Drug A is more effective than Drug B in reducing the blood pressure."

---

## One-Sided Test or Two-Sided Test?

---

### Two-Sided Test

In the above example, we established a null hypothesis that the difference in the average blood pressure between Groups A and B is zero and considered hypothesis testing, supposing only that Drug A is more effective than Drug B in reducing blood pressure. However, contrary to this supposition, there is the possibility that Drug B reduces blood pressure more than Drug A does. Now, consider the average blood pressure of Group A as  $M_a$  and that of Group B as  $M_b$ . In hypothesis testing that considers the two situations, one in which Drug A is more effective than Drug B and the other where Drug B is more effective than Drug A, the null hypothesis and the alternative hypothesis are as follows:

Null hypothesis:  $M_a = M_b$

Alternative hypothesis:  $M_a \neq M_b$

This is called a “two-sided test.”

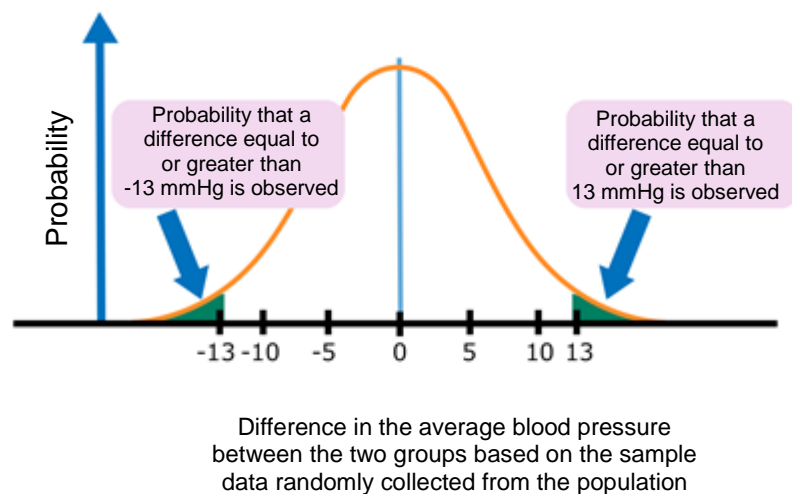
In a two-sided test, the alternative hypothesis, “ $M_a \neq M_b$  (the average blood pressure is not the same between Group A and B),” includes two opposing hypotheses: one is  $M_a > M_b$  (the average blood pressure is lower in Group B than in Group A) and the other is  $M_a < M_b$  (the average blood pressure is lower in Group A than in Group B). This alternative hypothesis can be expressed as follows:

The alternative hypothesis for the null hypothesis, “ $M_a = M_b$ ,” is either “ $M_a > M_b$ ” or “ $M_a < M_b$ .”

Therefore, the null hypothesis is rejected not only when  $M_a < M_b$  is supported, but also when  $M_a > M_b$  is supported. For this reason, commonly used p-values are calculated for two-sided test that considers the above two opposing hypotheses. This type of p-value is sometimes called p-value for two-sided hypothesis (testing) or “two-sided p-value,” suggesting its nature. In the above example, the p-value is calculated assuming that the average blood pressure can be 13 mmHg higher in Group A than in Group B with an equal probability even if the average blood pressure is in fact 13 mmHg lower in Group A than in Group B in the sample data. **Therefore, the p-value for two-sided test is 6%, twice the p-value of 3% for one-sided hypothesis (testing) that only considers the one-sided hypothesis, “ $M_a < M_b$ .”**

If we set the significance level at 5% and consider only the one-sided hypothesis, “ $M_a < M_b$ ,” we can say that there is a statistically significant difference because the one-sided p-value (3%) is smaller than the significance level. However, if we set the significance level at 5% and consider the two-sided hypothesis, we cannot say that there is a statistically significant difference because the two-sided p-value (6%) is greater than the significance level. Since a one-sided p-value is always smaller than a two-sided p-value, a statistically significant difference can easily be obtained if we set the significance level at 5% all the time without describing the sidedness of

hypothesis. In general, two-sided p-values are used in hypothesis testing except for special cases. If you use one-sided test, you must describe the reasons. Occasionally, there are studies in which the sample size was determined using one-sided hypothesis testing to reduce the sample size when designing the study protocol, but two-sided hypothesis testing was chosen for actual analysis, finding no statistically significant results. This is a grave fault caused by the inconsistency between the study design and analysis regarding the sidedness of hypothesis (whether one-sided or two-sided). In order to avoid this type of mistake, you must keep in mind to maintain consistency between the study design and actual analysis.




---

### One-Sided Test

A one-sided test may be acceptable when the direction of the change in the outcome of interest is known. An example of such a case is a study in which the test drug is known for sure to reduce the blood pressure so that no study would ever show that the average blood pressure is higher in Group A than in Group B, including measurement errors, even if one hundred researchers conduct one hundred separate studies. You can imagine that this type of study is unrealistic because measurement errors and mistakes are common in research.

Another example in which one-sided test is acceptable is “non-inferiority studies where only one direction of the effect needs to be considered.” A non-inferiority study aims to show that “Drug A is no inferior to Drug B.” This type of study is used when you want to assert the effectiveness of Drug A if the effectiveness of Drug A is either comparable to or slightly but not considerably less (no inferior) than that of conventional Drug B, in other words, if Drug A is found no inferior to conventional Drug B, after accounting for its safety and convenience. When you want to prove non-inferiority, test hypothesis needs to consider only one side (concerning whether inferior or not and disregarding whether superior or not) because it does not matter how superior the test drug is to the conventional drug if it is superior.

Here,  $M_a$  represents the average blood pressure of Group A and  $M_b$  that of Group B. “ $M_a - M_b$ ” represents the difference in the average blood pressure between Group A and B. When this



subtraction is negative, that is, when Drug A keeps blood pressure lower than Drug B does, our interpretation is that Drug A is superior. However, this subtraction can also be positive because Drug A may be inferior. The range of the difference that leads us to conclude non-inferiority of Drug A is called “non-inferiority margin.” Non-inferiority margin is 10 mmHg in a study where we conclude non-inferiority of Drug A when the average blood pressure in Group A is higher than that in Group B by 10 mmHg or less.

In this non-inferiority study examining whether Drug A is no inferior to Drug B, hypotheses can be expressed as follows using the non-inferiority margin:

Null hypothesis:  $M_a - M_b > 10$

Alternative hypothesis:  $M_a - M_b \leq 10$

The null hypothesis, which is to be rejected, is that “the average blood pressure is more than 10 mmHg higher in Group A than in Group B (Drug A is inferior to Drug B).” The alternative hypothesis is that “the average blood pressure in Group A is no higher than that in Group B by 10 mmHg (Drug A is no inferior to Drug B).”

---

### Interpretation of Study Results Using P-Values

A common mistake in interpreting p-values is to support the null hypothesis when the p-value is 0.05 (5%) or greater and the null hypothesis is not rejected, and to conclude that “the average blood pressure is the same when Drug A or B is used” or “Drug A is ineffective.” **You should never conclude “the same” solely based on the p-values.** Some clinical trials, such as one used in developing generic drugs, look for evidence that “this generic drug is biologically equivalent to the original drug” by comparing drug concentrations in the blood and such. This type of trial that aims to **prove equivalence, not differences, is called “equivalence trial.”**

Making p-values larger is very easy. You **reduce the number of study subjects, p-values become larger as much as you want**, because p-values depend greatly on the number of study subject.

It is wrong to suggest equivalence when the null hypothesis is not rejected (the p-value is 0.05 or greater) and no significant difference is found. However, this erroneous approach was used until the 1980s. Some studies that used this approach were published in high-quality widely-read journals, including New England Journal of Medicine, and many drugs reached the market with equivalence as their catchphrase. Unbelievable to today’s eyes. Such studies were also common in Japan until the former Ministry of Health and Welfare issued the previous version of Statistics Guidelines (1992). P-values are affected not only by whether a difference truly exists, but also by the number of study subjects. When the p-value is 0.05 or greater and no significant difference is observed, it only means that the null hypothesis, “the average blood pressure is the same when Drug A or B is used” or “Drug A is ineffective,” is not rejected. **You cannot tell why the null hypothesis is not rejected; is it because the drug is indeed ineffective? or is it simply because the data (the number of study subjects) is insufficient?** Therefore, do not conclude “effective” or

“ineffective” when you find or do not find a significant difference. Instead, use proper descriptions as follows:

### Interpretation of Study Results Using P-Values (Example)

**Null hypothesis: The difference in the average blood pressure between Group A and B is zero.**

“No statistically significant difference was observed” when two-sided test with a significance level of 5% was applied, because the p-value was equal to or greater than 0.05. In other words, results do not suggest difference between Drug A and B in their effectiveness.

When the p-value is smaller than 0.05

“A statistically significant difference was observed” when two-sided significance level of 5% was applied. In other words, results suggest a difference between Drug A and B in their effectiveness.

Thus, when the null hypothesis is not rejected, you are advised to use the expression that results do not conclude or suggest that “Drug A is effective,” rather than describing that “Drug A is ineffective,” admitting that the null hypothesis is correct.

---

### Limitations of P-Values

Calculations of p-values are strongly influenced by the number of study subjects as well as by “the differences in the outcomes (effects) between groups you intend to compare.” If the p-value is 0.05 or greater and you judge solely based on the p-value, you cannot tell whether it is because there truly are no clinical effects or it is because the number of study subjects is insufficient although there truly are clinical effects. Likewise, if the p-value is smaller than 0.05, you have no way to tell whether it is because there truly are clinical effects or because the number of study subjects is so large that statistical significance is found although clinical effects cannot be expected too much (due, for instance, to a huge variability of the effects).

Interpreting research results solely based on p-values is very risky because p-values can be quite small even for clinically meaningless differences (in the effects of drugs, for example) in studies that use data from tens of thousands of study subjects. Used in order to avoid this risk along with p-values is confidence intervals.

---

### P-Values and Confidence Intervals

Confidence intervals can show whether there is a statistically significant difference and whether the p-value is smaller than 5%.

#### Example 1

Six of 10 study subjects who received the new drug were cured, while two of 10 study subjects who received the conventional drug were cured. The cure rates for the new and conventional

drugs were 60% and 20%, respectively. The inter-group ratio of the cure rate for the new drug to that for the conventional drug is  $0.6/0.2 = 3$ . This means that the new drug can cure the disease three times better than the conventional drug. The 95% confidence interval for this inter-group ratio of cure rate is “0.786 - 11.445.”

**Example 2**

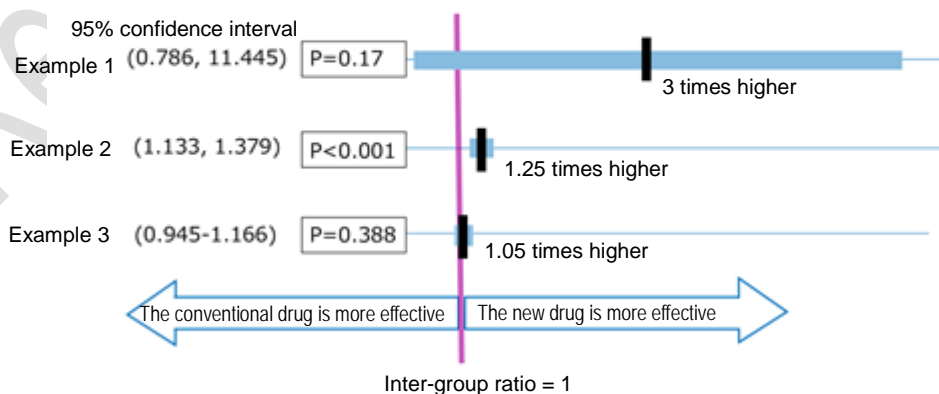
Five hundred of 1,000 study subjects who received the new drug were cured, while 400 of 1,000 study subjects who received the conventional drug were cured. The inter-group ratio of the cure rate for the new drug to that for the conventional drug is  $0.5/0.4 = 1.25$ . This means that the new drug can cure the disease 1.25 times better than the conventional drug. The 95% confidence interval for this inter-group ratio of cure rate is “1.133 - 1.379.”

**Example 3**

Four hundred and twenty of 1,000 study subjects who received the new drug were cured, while 400 of 1,000 study subjects who received the conventional drug were cured. The inter-group ratio of the cure rate for the new drug to that for the conventional drug is  $0.42/0.4 = 1.05$ . This means that the new drug can cure the disease 1.05 times better than the conventional drug. The 95% confidence interval for this inter-group ratio of cure rate is “0.945 - 1.166.”

In Example 1, the 95% confidence interval includes “(an inter-group ratio of) 1,” which means that there is no difference between the two groups. This suggests that the p-value is 5% or greater, confirming no significant difference even though calculations show that the new drug is found three times more effective than the conventional drug. In Example 2, the value, “1,” is not included in the 95% confidence interval. This suggests that the p-value is smaller than 5% and a statistically significant difference is found even though calculations show that the new drug is only 1.25 times as effective as the conventional drug. This is because the 95% confidence interval in Example 2 is narrower than that in Example 1 and does not include “1” due to the greater number of study subjects even though the inter-group ratio is a small value of 1.25.

Inter-group ratio of the cure rate with the new drug to that with the conventional drug and its confidence interval



The above figure shows the 95% confidence intervals of Example 1 through 3. Neither Example 1 or 3 finds statistically significant difference because the p-value is greater than 5%. However, you can see that the reasons behind are different. In Example 1, no statistically significant difference is observed because the confidence interval is wide and includes the value, “1,” meaning no difference, due to the small number of study subject, 10 in each group, although the cure rate is 3 times higher for the new drug than that for the conventional drug. In Example 3, no statistically significant difference is observed because the difference in the cure rate is small, with 42% and 40% for the new and conventional drugs, respectively, and the confidence interval includes “1.”

---

### Interpretation of Study Results Using Confidence Intervals

As shown in the above examples, use of confidence intervals allows hypothesis testing without p-values. Confidence intervals also reveal what is unapparent from p-values. Therefore, you are advised to describe inter-group differences in the effects and confidence intervals in addition to p-values when reporting the results of analysis.

The following three types of finding are possible in clinical trials, and the test results of each can be interpreted using confidence intervals as described in the later sections.

- (1) The effectiveness of the new drug is higher than that of the conventional drug (**superiority** trial).
- (2) The effectiveness of the new drug is not inferior to that of the conventional drug (**non-inferiority** trial).
- (3) The effectiveness of the new drug is equivalent to that of the conventional drug (**equivalence** trial).

---

### When Proving Superiority

In superiority trials, (1) above, when the 95% confidence interval does not include the value that means no difference, the p-value is less than 0.05 and, thus, you can conclude that there is a statistically significant difference. Roughly speaking, the dissimilarity between the groups can be expressed using either a **difference** or **ratio**. Let's say the average blood pressure of 50 subjects who received Drug A (a new drug) is 100 mmHg and that of a group of subjects who received Drug B is 113 mmHg. The difference in the average blood pressure is  $113 - 100 = 13$  mmHg whereas the ratio of the average blood pressure is  $113 \div 100 = 1.13$ . When the difference is used, the difference of 0 mmHg in the average blood pressure indicates no inter-group difference in the effectiveness and, thus, superiority is proven if the value, 0, is not included in the confidence interval. When the ratio is used, a ratio of 1 indicates no inter-group difference in the effectiveness and, thus, superiority is proven if the value, 1, is not included in the confidence interval.

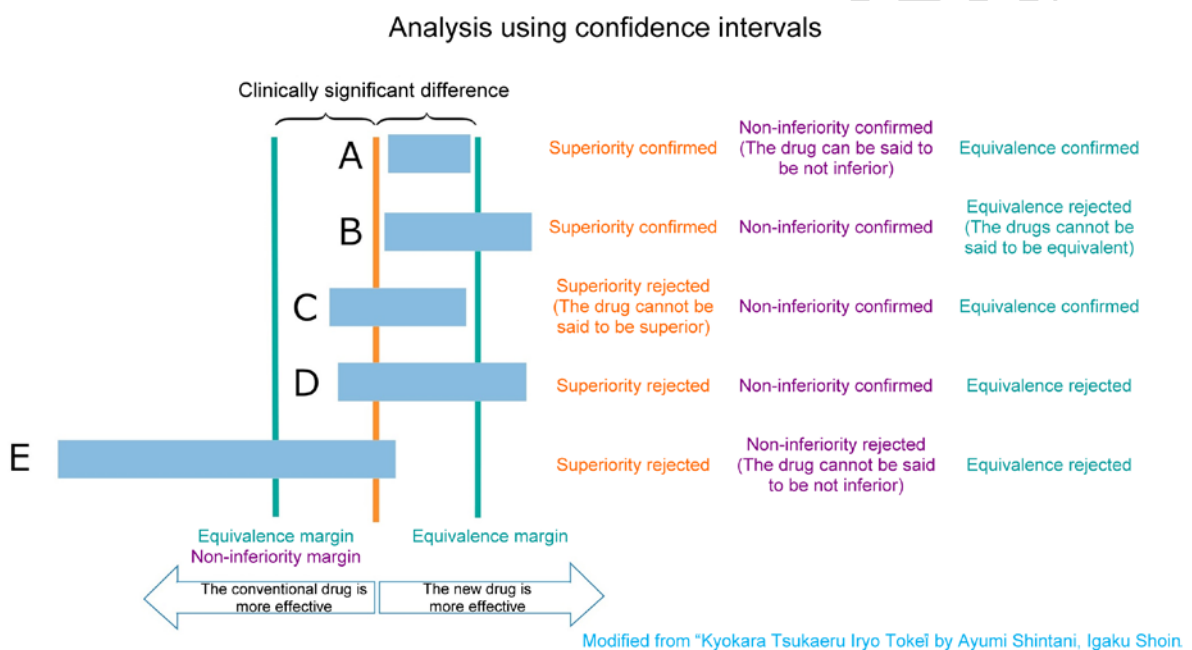
---

### When Proving Non-Inferiority or Equivalence

Confidence intervals also make it easier to understand in non-inferiority trials, (2) above, and

equivalence trials, (3) above.

In a study to prove equivalence, such as “the average blood pressure is the same when receiving the new drug or conventional drug,” you set a margin ( $\Delta$ ) that indicates equivalence before starting the study by saying, for example, that we conclude equivalence when the inter-group difference in the average blood pressure between the new and conventional drugs does not exceeds 5 mmHg. When the confidence interval of the inter-group difference in the average blood pressure falls within this margin, equivalence is statistically proven. Likewise, in a study to prove non-inferiority, such as “the new drug is no inferior to the conventional drug,” non-inferiority is statistically proven when neither the upper nor the lower limit of the confidence interval (the lower limit in the figure below) exceeds the pre-determined margin.



Now, let's look at the examples mentioned above one more time. This time, the equivalence and non-inferiority margins are set at  $\pm 6\%$  and  $-6\%$ , respectively, before starting each study.

**Example 1**

Six of 10 study subjects who received the new drug were cured, while two of 10 study subjects who received the conventional drug were cured. The cure rates for the new and conventional drugs were 60% and 20%, respectively. The ratio of the cure rate for the new drug to that for the conventional drug is  $0.6/0.2 = 3$ . This means that the new drug can cure the disease three times better than the conventional drug. The 95% confidence interval for this inter-group ratio of cure rate is “0.786 - 11.445.”

- The equivalence and non-inferiority margins are set at  $\pm 6\%$  and  $-6\%$ , respectively. In Example 1, the confidence interval is “0.786-11.445,” which includes the value, 1, that means there is no difference. Therefore, superiority is not proven. Moreover, the lower

limit of the confidence interval is 0.786. When we regard the effectiveness of the conventional drug as 1, this lower limit of the confidence interval indicates that the new drug is 21.4% less effective than the conventional drug. In addition, this lower limit exceeds both the equivalence and non-inferiority margins that allow effectiveness 6% less than that of the conventional drug. Thus, neither equivalence nor non-inferiority is proven (Pattern E in the above figure).

### Example 2

Five hundred of 1,000 study subjects who received the new drug were cured, while 400 of 1,000 study subjects who received the conventional drug were cured. The inter-group ratio of the cure rate for the new drug to that for the conventional drug is  $0.5/0.4 = 1.25$ . This means that the new drug can cure the disease 1.25 times better than the conventional drug. The 95% confidence interval for this inter-group ratio of cure rate is "1.133 - 1.379."

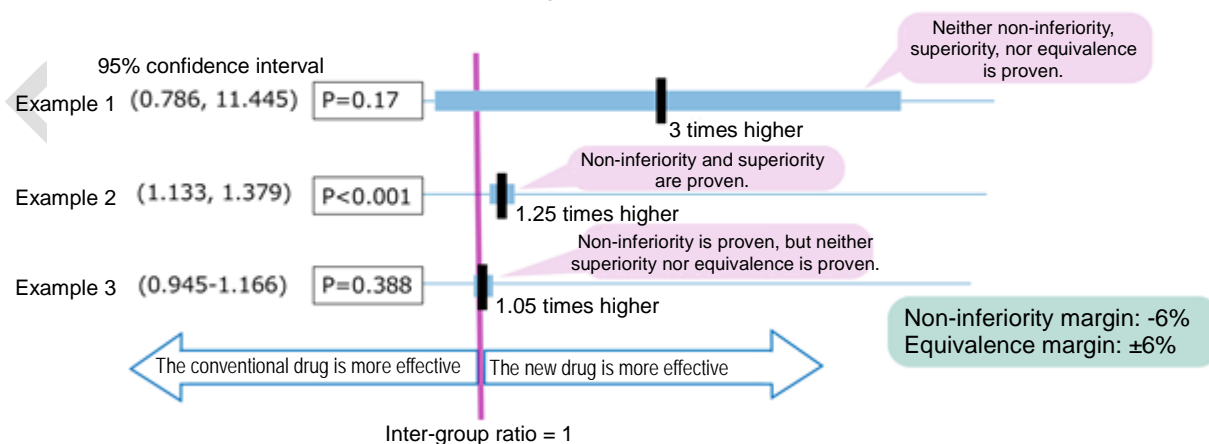
- ➔ Since the confidence interval does not include the value, 1, superiority is proven. Moreover, non-inferiority is also proven because superiority is proven (Pattern A in the above figure).

### Example 3

Four hundred and twenty of 1,000 study subjects who received the new drug were cured, while 400 of 1,000 study subjects who received the conventional drug were cured. The inter-group ratio of the cure rate for the new drug to that for the conventional drug is  $0.42/0.4 = 1.05$ . This means that the new drug can cure the disease 1.05 times better than the conventional drug. The 95% confidence interval for this inter-group ratio of cure rate is "0.945 - 1.166."

- ➔ Since the lower limit of the confidence interval is 0.945, non-inferiority is proven. However, neither superiority nor equivalence is proven (Pattern D in the above figure).

Inter-group ratio of the cure rate with the new drug to that with the conventional drug and its confidence interval



---

This module was prepared by “Ethics Education Program on the Research Reliability Standards of International Medical Journals” (or “AMED International Journals Project”) with funding support from “Research and Development of Material/Program for Research Integrity Education” of the Japan Agency for Medical Research and Development. Please visit [the APRIN website](#) for the names of the experts who prepared and/or reviewed this module.

Reproduction Prohibited

---

International journals have the following checkpoints in relation to the content of this module (partially summarized or supplemented to help understanding the content).

- [1] Nature (<http://image.sciencenet.cn/olddata/kexue.com.cn/upload/blog/file/2010/12/2010128212513557501.pdf>; visited on 2019.03.18)
- [2] New England Journal of Medicine (<http://www.nejm.org/page/author-center/manuscript-submission#electronic>; visited on 2019.03.18)
- [3] Science (<http://www.sciencemag.org/authors/science-editorial-policies>; visited on 2019.03.18)
- [4] The EMBO Journal (<http://emboj.embopress.org/authorguide#statisticalanalysis>; visited on 2019.03.18)
- [5] JAMA (<http://jamanetwork.com/journals/jama/pages/instructions-for-authors>; visited on 2019.03.18)

#### [1] Nature

- Alpha level is given for all statistical tests (e.g. 5%)
- Tests are clearly identified as one or two-tailed
- Actual *P* values are given for primary analyses

#### [2] New England Journal of Medicine

- Except when one-sided tests are required by study design, such as in noninferiority trials, all reported *P* values should be two-sided. In general, *P* values larger than 0.01 should be reported to two decimal places, and those between 0.01 and 0.001 to three decimal places; *P* values smaller than 0.001 should be reported as  $P < 0.001$ . Notable exceptions to this policy include *P* values arising from the application of stopping rules to the analysis of clinical trials and from genetic-screening studies.
- For tables comparing treatment groups at baseline in a randomized trial (usually the first table in the manuscript), significant differences between or among groups (i.e.,  $P < 0.05$ ) should be identified in a table footnote and the *P* value should be provided in the format specified above.

#### [3] Science

- The testing level (alpha) and whether one-sided or two-sided testing was used should be reported for each statistical test; typically, two-sided testing is appropriate, but if one-sided testing is used, its use should be justified.
- Authors should present results in complete and transparent fashion so that stated conclusions are backed by appropriate statistical evaluation and limitations of the study are frankly discussed.
- Results of each statistical test should be reported in full with the value of the test statistic and *p*-value, and not simply reported as significant or non-significant; more than two significant digits on *p*-values are usually not needed except in situations of extreme multiple testing such as in genetic association studies where stringent corrections for multiple testing might be used.

#### [4] The EMBO Journal

- The description of all reported data that includes statistical testing must state the name of the statistical test used to generate error bars and *P* values, the number (*n*) of independent experiments underlying each data point, and the actual *P* value for each test (not merely 'significant' or ' $P < 0.05$ ').

#### [5] JAMA

- Avoid solely reporting the results of statistical hypothesis testing, such as *P* values, which fail to convey important quantitative information.
- For most studies, *P* values should follow the reporting of comparisons of absolute numbers or rates and measures of uncertainty (e.g. 0.8%, 95% CI -0.2% to 1.8%;  $P = .13$ ).
- If *P* values are reported, follow standard conventions for decimal places: for *P* values less



than .001, report as “ $P < .001$ ”; for  $P$  values between .001 and .01, report the value to the nearest thousandth; for  $P$  values greater than or equal to .10, report the value to the nearest hundredth; and for  $P$  values greater than .99, report as “ $P > .99$ .” For studies with exponentially small  $P$  values (e.g. genetic association studies),  $P$  values may be reported with exponents (e.g.  $P = 1 \times 10^{-5}$ ). In general, there is no need to present the values of test statistics (e.g.  $F$  statistics or  $\chi^2$  results) and degrees of freedom when reporting results.

- For randomized trials using parallel-group design, there is no validity in conducting hypothesis tests regarding the distribution of baseline covariates between groups; by definition, these differences are due to chance. Because of this, tables of baseline participant characteristics should not include  $P$  values or statements of statistical comparisons among randomized groups. Instead, report clinically meaningful imbalances between groups, along with potential adjustments for those imbalances in multivariable models.